# Reviews

**Karin Aijmer** and **Bengt Altenberg** (eds.). *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg 22–26 May 2002*. Amsterdam and New York: Rodopi, 2004. vii + 419 pp. ISBN: 90-420-1741-4. Reviewed by **Kay Wikberg**, University of Oslo.

The volume opens with three introductory papers on 'The role of corpora in linguistic research'. Spoken language has always been a major concern of Michael Halliday's. In 'The spoken language corpus: A foundation for grammatical theory' Halliday examines the role of corpora in the study of spoken language. With reference to recent research he comments on a number of special features of the spoken language, such as non-standard patterns, the grammar of appraisal and grammatical intricacy. He discusses a number of problems with a spoken corpus such as transcription and the grammaticalization of the spoken mode. Halliday would like to see more research on the grammar of speech to form a basis for grammatical theory.

John Sinclair addresses the question of the role of intuition and annotation in corpus linguistics. As far as intuition is concerned, linguists cannot do without it in their role as analysers and interpreters of corpus evidence. By contrast, if by intuition you refer to the native speaker's competence, Sinclair is sceptical and goes in for his well-known principle of 'trusting the text'. In his treatment of annotation he distinguishes between 'mark-up' and 'annotation'. Mark-up refers to the process of recording things like bold face, italic, large fonts and layout generally whereas "[a]nnotation … uses the same conventions as mark-up but has no limits on the kind of information that is provided" (p. 48). According to Sinclair, annotations still have a place in particular applications but are not to be recommended in generic corpora since "they impose one particular model of language on the corpus" (p. 54). Therefore, if annotations are provided, they should be optional.

The third introductory paper is written by Geoffrey Leech. In 'Recent grammatical change in English: Data, description, theory' he does two things. First of all, he sets data-driven corpus linguistics against a generative framework and argues that the former is more than just 'mere data collection' or 'mere description'. Second, he investigates the development of modal auxiliaries and the phenomenon of 'colloquialization' in Present-day English, using LOB, FLOB, Brown and Frown plus some mini-corpora of spoken English based on the Survey of English Usage and ICE-GB (80,000 words each). It turns out that modals are generally most frequent in LOB and least frequent in Frown. While the corpus data show a sharp decline of frequency in the use of *must, may, shall* and *ought* in both British and American English, the use of semi-modals like *be going to, need to* and *have to* has been growing. As far as the process of 'colloquialization' is concerned, Leech registers a number of changes when comparing the LOB and FLOB corpora. On the one hand, there is a rise in the use of the present progressive, on the other, there is a general decline in the use of passives. Questions are more common in FLOB, which may at least in part be due to a higher percentage of quoted text in FLOB. Finally, there are changes in relative clauses, such as more zero relatives with a stranded preposition (*someone I spoke to*). Leech takes care to point out that his findings are provisional, that there is no straightforward explanation of increase or decrease of frequency except possibly the influence of mass media, and finally that only a usage-based model of language forms "a natural bridge between the study of naturally-occurring data and the cognitive and social workings of language" (p. 78).

One of the assumptions underlying cognitive grammar is that it is usage-based. In his analysis of *give* Joybrato Mukherjee shows that a corpus approach can very well supplement a usage-based cognitive grammar. He provides a detailed analysis of 1,064 instances of *give* in ICE-GB with frequencies of the various patterns. Mukherjee emphasizes that cognitive grammar can profit from corpus data by obtaining information on frequencies, lexicogrammatical patterns, and both recurrent usage and linguistic creativity. A corpus approach can also help to provide cognitive grammar with "context-dependent principles of pattern selection (such as lexico-grammatical co-selection, pragmatic principles and activation statuses of discourse entities)" (p. 97).

Two papers deal with semantic categories of verbs. Caroline David looks at 'Putting 'putting verbs' to the test of corpora' and Åke Viberg examines 'Physical contact verbs in English and Swedish'. Using the BNC, LOB, Flob, Brown and Frown, and semantic information from dictionaries, Caroline David shows that *put, set, lay* and *place* should not be classified as belonging to the same category of 'PUT-verbs' since *lay, set* and *place* all convey the more specific feature

of [MANNER OF MOVING] and therefore are hyponyms of the more general verb *put.* Although the author makes it clear from the start that she is dealing with 'putting verbs', she could have been more explicit about the polysemy of *put* since several of the verb-preposition combinations she lists (p. 103) represent communication verbs (*put down, put across, put forward*) and some convey mental processes. The second part of Caroline David's paper deals with the well-known SPRAY/LOAD-verbs.

Åke Viberg is one who has investigated the use of *put* in English and its Swedish equivalents (Viberg 1998). Viberg's physical contact verbs include the English *strike, hit* and *beat* and their Swedish equivalents as found in the English-Swedish Parallel Corpus. This is an interesting paper both theoretically and methodologically. It draws on two theoretical frameworks: conceptual representation (cognitive semantics and prototype theory) and the use of the local or topical context for word sense identification. The prototypical contact verb in Swedish is *slå,* which "is grounded more firmly in sensorimotor experience of limb movement than *strike, hit* and *beat*" (p. 349), and its major meanings are represented in a lucid contrastive display (p. 341). As regards word sense identification, a number of disambiguation cues are found in both languages. The semantic class of Subject turns out to be most useful with *strike* and *hit,* whereas the semantic class of Object is more helpful for *slå.* In many cases, though, a span of ± 2 is not enough, which means that topical and pragmatic information has to be utilized. Towards the end of the paper Viberg touches on the possible universality of semantic patterning.

Metaphor is analysed by Jonathan Charteris-Black in a paper based on the Inaugural speeches of American Presidents and party political manifestos of two British political parties in the period 1974–1997. He relies on conceptual metaphor theory and pays attention to the communicative purpose of the texts. For identification he makes use of I.A. Richards's notion of semantic tension between the source and target domains. The two subcorpora share the most common lexical fields expressed through metaphor: conflict, journeys and buildings. By contrast, the British corpus favours plant metaphors whereas fire, light and the physical environment are common source domains in the American corpus. The author has recently expanded this kind of research considerably to include other genres (Charteris-Black 2004).

Not surprisingly, the Web is attracting more and more attention among corpus linguists. Three papers in this volume deal in different ways with the Web. Thus Peter Tan, Vincent Ooi and Andy Chiang examine the phenomenon of spokenness as realized by the use of appraisal, more particularly amplification items (*lah, ever, a lot, really, too, very*) and mitigators (*somewhat, a bit, only,*

*just*), as occurring in a personal advertisement subcorpus from the Web and spoken vs. written portions of the Singapore component of ICE. The authors describe the background factors very carefully but the findings are still relatively modest. The pragmatic particle *lah*, the most common of the investigated items in the spoken corpus, hardly occurred at all in the advertisements, possibly due to its local flavour. By comparison, *just* was a high-frequency item in both spoken data and personal advertisements.

Many of us have already used WebCorp when even large corpora have turned out to be inadequate. Antoinette Renouf, Andrew Kehoe and David Mezquiriz, staff members attached to the WebCorp project, describe the WebCorp and discuss a number of issues connected with its use as a corpus in 'The accidental corpus: Some issues in extracting linguistic information from the Web'. For example, the inexperienced user may not know that the linguistic sample is not constant, and that you can only describe one linguistic phenomenon at a time in the same body of data. The paper describes a number of solutions to specific search problems such as pattern matching and search refinement. Thus it is possible to go in for domain-specific search (e.g. *.ac.uk* OR *.edu*) or to refine a search by specifying language (e.g. *.no* for Norway). This area obviously has tremendous potential for further development.

Natalie Kübler describes an experiment carried out to use WebCorp and finite corpora in the classroom for building specialized dictionaries. The finite corpus in this case was the parallel English-French HOWTO corpus, based on the user manual files of the Linux operating system, but other small corpora dealing with the domain of computing were also available. The students used a Terminology Extractor tool to extract terms in the two languages. Information from the dynamic WebCorp turned out to be very useful in the way it supplemented and updated the other time-bound sources.

Two papers treat different aspects of noun phrases. Peter Willemse studies 'Esphoric reference and pseudo-definiteness' in a collection of 200 existential sentences extracted from *The Bank of English*. A typical example of a pseudo-definite NP is Willemse's:

(1) Tomorrow afternoon, there will be the usual Christmas concert. (p. 118)

The NP is formally definite but actually introduces a new entity into the discourse. 'Esphoric' reference occurs when there is a 'forward' phoric relationship within the same NP and "is associated with presuming rather than presenting reference" (p. 117) as in this instance of listing, here abbreviated:

(10) … there was the stump of a cigar bearing the marks of a man's teeth; and there was a clump of fluffy dust freshly gathered from some floor (p. 124)

Willemse deals with three basic categories: type/subtype, possessives (as in (10)), and general/specific constructions, including appositives. The basis for Willemse's classification is the construction types of the pseudo-definite NPs and the semantic relation between NP1 and NP2. He explains the use of the definite article in an example like (10) as a 'bridging relationship' between NP1 (*the stump*) and NP2 (*a cigar…*).

Generations of students have had problems analysing another kind of NP1 *of* NP2-construction as in *loads of* (NP1) *furniture and other stuff* (NP2) and *a pile of* (NP1) *other books* (NP2). Liselotte Brems gets to grips with such measure noun (MN) constructions in her paper and considers the MNs in a process of ongoing delexicalization and grammaticalization. Whereas *loads* in its syntactic and semantic behaviour is not far from the quantifiers *a lot of* and *lots of*, *a pile of* and *piles of* mostly function as head nouns. Thus *loads* has lost some of its primary meaning of 'heavy things' and increased its quantifier status, which is a kind of grammaticalization.

Michael Hoey argues rightly that "[c]orpus linguistics has not attended much to text-linguistic issues" (p. 171). To remedy this situation Hoey zooms in on the word and assigns textual function to an infinite number of lexical items which either have a positive or negative preference for use in textual organisation. Hoey calls this function 'textual colligation'. 'Colligation' used to stand for a syntagmatic relation, but Hoey extends the meaning to include positioning within any chunk of text, such as in cohesive chains, themes, and paragraphs, even in semantic relations. Hoey makes a number of claims which together represent textual colligation. One such claim is

> Claim 2: Every lexical item (or combination of lexical items) may have a positive or negative preference for occurring as part of Theme in a Theme-Rheme relation. (p. 176)

Systemic Functional Grammar has taught us that certain (types of) words or expressions are typically thematic (e.g. words appearing in Subject position, Adjuncts like that is, anyway, in my opinion). Hoey is aware that an enormous lot of work remains to substantiate his claims. Since 2002 he has developed his idea of lexical priming into a more comprehensive theory, which is certainly welcome in a field in need of fresh theoretical ideas (Hoey 2005).

In the same section on 'Discourse and Pragmatics' Hilde Hasselgård looks at 'Adverbials in *IT*-cleft constructions' on the basis of material from ICE-GB. Her examination of their information structure and discourse functions shows "that *IT*-clefts with adverbials tend to be informative-presupposition clefts, while other *IT*-clefts are more likely to be stressed-focus clefts. The typical information structure of the *IT*-clefts with adverbials involves a clefted constituent carrying given information and a cleft clause carrying new information or, alternatively, one in which both parts of the cleft construction are new" (p. 208). She gives examples of *IT*-clefts having contrastive, topic-launching, transitional and summative functions and wisely points out that "assigning information values and discourse functions is no exact science" (p. 209). As regards frequency in relation to genre, Hasselgård finds that *IT*-clefts with adverbials are most common in scripted speech (lectures, broadcast narration, official speeches). She assumes that reasons for that "may have to do with the possibility of assigning unambiguous thematic prominence to the clefted constituent and the possibility of presenting new information in the cleft clause without asserting it" (p. 208).

Another scholar dealing with spoken English in ICE-GB is Bernard De Clerck, who investigates the pragmatic functions of *let's* utterances. Unlike *let us* (as in *Let us have a drink*), which can express non-inclusive permission ('allow us to'), *let's* (as in *Let's have a drink*) is used to express a hearer-inclusive proposal for joint action. *Let* in *let's* is considered to have undergone semantic bleaching, a process which seems to have progressed further in AmE than in BrE. De Clerck finds that *let's* utterances with joint agency are most common in face-to-face conversation, spontaneous commentaries, business transactions and broadcast discussions. The most common function of conversational *let's* utterances in the investigated material was to influence the conversational flow of the interaction (*e.g. let's backtrack for a second*).

Thomas Kohnen is interested in diachronic pragmatics in a paper called 'Methodological problems in corpus-based historical pragmatics. The case of English directives'. He takes up things like how to get at the range of sentence (utterance) forms expressing directive speech acts, difficulties of interpretation, the relation between attested corpus data and actual usage at a given point of time, and the lack of sufficient data. He calls his approach 'structured eclecticism', by which he refers to a procedure involving "a deliberate selection of typical patterns which we trace by way of a representative analysis throughout the history of English" (p. 238). His main corpus is the *Helsinki Corpus* supplemented with dictionary data. In his diachrony of directives he shows that there is a development towards "less explicit, less direct and less face-threatening" man-

ifestations (p. 246). He attributes the increased use of indirect speech acts to considerations of politeness. An important source of supplementary data not mentioned by Kohnen is drama texts. There is a large number of plays available in electronic form which could supplement the fragments in the *Helsinki Corpus* and which would certainly contain numerous directives.

The history of high-frequency words can be fascinating, at least as described by Göran Kjellmer, who has written on '*Yourself*: A general-purpose emphatic-reflexive'. He draws on the *OED* for historical data but uses CobuildDirect and the BNC to illustrate present-day variation in the use of *yourself*. "The number indeterminacy of *you* spills over to *yourself* by analogy" (p. 270), which allows *yourself* to have plural reference side by side with the singular (*Treat yourself to a Maltese odyssey*). The next step is when *yourself* clearly refers to the second person plural (*Do you all consider yourself to be Botards?*), which occurs quite frequently in the corpora. Kjellmer also mentions a colloquial use (*We have to think yourself!*), referring to the first person. Finally, there is a development where *yourself* is used in a generic sense 'one' as in *one is to do it yourself*.

By using the Polytechnic of Wales Corpus of Children's English (POW) Clive Souter shows that quantity is not everything. POW is a corpus of just 61,000 words (3,730 types), collected in the late 1970's. Souter uses it to show changes in vocabulary growth between the ages of 6, 8, 10 and 12. There were in all 96 informants, subdivided according to age, sex and socio-economic class. The data were collected in a recorded play session when the children were engaged in a *Lego* building task, and each child was also interviewed about favourite games and TV programmes. Souter found a marked difference between boys and girls in their use of words. Another finding was that the children's speech contained numerous grammatical structures (ellipsis, interruptions) not found in written corpora. Some Welsh dialect features were also found, such as a high frequency of tag questions and the dialectal locative adverbs *by-here* and *by-there* for *here* and *there*. In spite of the limitations of this corpus, the findings are interesting from both a language learning and a language teaching perspective.

By now a large number of studies of learner language have been carried out using corpora in the International Corpus of Learner English (ICLE) project. Roumiana Blagoeva's contribution to this volume is entitled 'Demonstrative reference as a cohesive device in advanced learner writing: A corpus-based study'. The adjective *Bulgarian* should have been added to the title. This study focuses on how Bulgarian learners use the English demonstratives and the reasons why differences occur. Blagoeva uses four corpora: *Corpus 1* is the learner corpus of argumentative essays written by Bulgarian university students; *Corpus 2* is the

British LOCNESS; *Corpus 3* is a subcorpus of non-fiction texts (Applied Science, Social Science, World Affairs) from the BNC; and *Corpus 4* consists of Bulgarian texts equivalent to those in *Corpus 3*. The size of each corpus is about 200,000 words. A major finding is that Bulgarian learners show a wider variety of patterns compared with the LOCNESS and the BNC material. The author discusses possible reasons for this, such as differences in the use of demonstratives to refer to extended text in the two languages. Whereas English *this* and *that* are both used anaphorically to refer to the content of text segments, Bulgarian has only one form for this function (*tova* [singular, neuter, near]). There is a clear overuse of *that* in *Corpus 1* compared with *Corpus 2* and *Corpus 3*. A discussion of the comparability of texts would have been in order since it appears that the frequencies in *Corpus 3* (BNC) differ radically from those in the other corpora.

As far as methodology is concerned, Helge Dyvik's paper on 'Translations as semantic mirrors: From parallel corpus to Wordnet' stands out in the way it combines data from a parallel corpus (ENPC) with lexical semantics and computational linguistics. To give a fair representation of Dyvik's interesting paper we would really have to use some of his illustrations. Here we will have to focus on a few points. Dyvik uses WordNet (e.g. Fellbaum, 1998) rather than semantic theory as the underlying model for formulating semantic concepts like 'synonymy' and 'hyponymy'. Provided you look up a word in either English or Norwegian, the corpus gives a set of translated sentences from which one can derive the translations and the senses of each word. The next stage involves grouping senses in semantic fields. "On the basis of the structure of a semantic field a set of features is assigned to each individual sense in it, coding its relations to other senses in the field… Based on intersections and inclusions among the feature sets a semilattice is calculated with the senses as nodes" (p. 311). This lattice is used to obtain information about semantic relations among the nodes. The final output of the programme will be thesaurus-like entries for words. Considering the limited size of the ENPC corpus, relatively high-frequency items have to be selected.

Anna-Lena Fredriksson deals with 'Exploring theme contrastively: The choice of model'. This is a thorough paper, in which she grapples with the problem of applying the concept of 'theme' as used in Systemic Functional Grammar to other languages, not least Swedish. Since the definition and interpretation of 'theme' in English grammar has given rise to a fair amount of research, it is natural that more difficulties arise in typologically different languages, for instance when word order is governed by the V2 constraint or when there are other system differences. Fredriksson takes a global view of theme and thinks that it is

necessary to go beyond the clause or sentence level when assigning the theme-rheme transition point. Within the sentence she includes all preverbal elements, but in Swedish, in addition, it seems necessary to include the postverbal subject as experiential theme.

Elena Tognini Bonelli and Elena Manca demonstrate what happens when lexical equivalence does not exist and when functional translation equivalence has to be sought instead. Their investigation concerns a comparison of how the field of 'Farmhouse holidays' (UK) and the corresponding 'Agriturismo' (Italy) are represented as discourse in the two languages. Corpus data was collected from web pages, the Italian 'Agriturist' corpus (115,000 words) and the English 'Farmhols' corpus (203,000 words). A surprising early finding was that there were only three instances of It. *benvenuto* corresponding to Eng. *welcome* (n=324). This resulted in search based on important collocates of *welcome* (*children*, *pets*, *guests*) and their translational equivalents. Finally, the collocational range of these Italian equivalents (*bambini*, *animali*, *ospiti*) was investigated. The authors emphasize that "any translating activity should start by considering very carefully the context in which a certain word or expression is embedded and the one into which it is going to be transferred" (p. 384).

Spectacular advances in corpus studies are not made in the interval between two annual ICAME conferences. Still, the present volume is proof of progress in several respects. There is a lot of descriptive work going on using existent corpora and applying various theoretical models at the same time as new corpora are being created for specific purposes. In both cases the methodology required is a very important issue.

## *References*

Charteris-Black, Jonathan. 2004. *Corpus approaches to critical metaphor analysis.* Basingstoke: Palgrave Macmillan.

Fellbaum, Christiane (ed.). 1998. *WordNet. An electronic lexical database.* Cambridge, Mass.: The MIT Press.

Hoey, Michael. 2005. *Lexical priming. A new theory of words and language.* London and New York: Routledge.

Viberg, Åke. 1998. Contrasts in polysemy and differentiation. Running and putting in English and Swedish. In S. Johansson and S. Oksefjell (eds.). *Corpora and cross-linguistic research. Theory, method, and case studies*, 342–376. Amsterdam and Atlanta: Rodopi.

**Paul Baker.** *Using corpora in discourse analysis*. London and New York: Continuum, 2006. 240 pp. ISBN 0-826-47725-9. Reviewed by **Karin Aijmer**, University of Göteborg.

Until recently discourse analysts have not been greatly interested in using corpora. The author of *Using corpora in discourse analysis* reports that he found "interest, disinterest and hostility" towards using corpora in discourse analysis in about equal amounts when he has been going to international conferences (p. 6). The present book was written to show that corpus-based techniques have an important role to play in addition to and as a complement to more traditional methods in discourse analysis.

The book has eight chapters. The focus is on different sorts of analytical techniques that can be used in corpus-based (critical) discourse analysis. Each chapter includes a number of case studies mainly using some kind of specialized corpus.

Chapter 1 sets the scene for the rest of the book and provides an introduction of what the author means by corpus linguistics and by discourse. The term discourse is especially problematic since it is used in different ways in social and linguistic research. The author chooses to focus on Foucault's definition of discourse as "practices which systematically form the objects of which they speak" (p. 4). 'Discourses' represent the different ways of constructing or viewing the world. One way that they can be constructed is via language, and language contexts are therefore a rich source for uncovering different types of discourses. Much of the work which has been carried out in discourse analysis is qualitative rather than quantitative and is therefore subjective. The availability of corpora has given us the opportunity to study discourse in a more objective way. Corpora are described as 'helping to restrict bias' by making it less easy for the researcher to be subjective in selecting data. Corpora focus on what is representative (because repeated) rather than on single cases, which makes the analysis more objective. Moreover by studying how meanings vary or change when we compare different corpora we can show how discourse positions in a society are in a flux and that what is acceptable today may be unacceptable tomorrow.

The time is ripe to include corpora in discourse analysis. This is also in line with a shift in post-structural thinking towards deconstructing binary arguments and either-or dichotomies and placing more focus on both-and (referred to as triangulation). The availability of corpus methods should be seen as a complement to other methods. For example, it only takes a few minutes to check whether a particular collocation is highly 'loaded' and can be used to support an analysis of what Baker refers to as hegemonic discourse.

Much of the material in Chapter 2 is likely to be familiar to corpus linguists. The main part of the chapter deals with different types of corpora (general, specialized), traditional corpus building and with issues such as size, sampling, representativeness and more practical problems, such as downloading internet texts, scanning, keying in the text, copyright. Different types of annotation schemes are introduced. The novice to corpus-linguistic studies can also get information about corpora in existence which are of interest to discourse analysts.

Chapter 3 is concerned with frequency and dispersion analyses and how they could be important in corpus linguistics. This is illustrated by means of a corpus analysis of holiday leaflets. The analysis of the most frequent lemmas showed how holiday makers were constructed in the discourse as being interested in places to drink. The phrase *work 2 live [work to live]* was shown by means of dispersion analysis to occur primarily at the beginning of the leaflet.

In Chapter 4 the author discusses concordance analysis and how concordances can be of use in discourse analysis to show how the context contributes to particular discourses. The semantic patterning of lexical elements may also provide evidence for people's attitudes and evaluations. The author distinguishes between discourse prosody and semantic preference, the difference being that discourse prosody has to do with "the relationship of a word to speakers and hearers" and "is more concerned with attitudes" (p. 87).

Chapter 5 deals with collocation and how studying collocates can help us to see how the text is discursively constructed. This is illustrated with an analysis of the lexical items *bachelor* and *spinster* and their collocates, perhaps no longer the most frequent words to describe unmarried people. The chapter also describes a number of statistical tests such as mutual information and log-likelihood needed to identify the strength of the collocates.

Chapter 6 illustrates how it is possible to use WordSmith tools to compare two word-lists and to compile a so-called keyword list on the basis of the comparison. Such a list contains all the words which are more frequent than expected when one compares the word-lists. As the author shows on the basis of data from a parliamentary debate on fox-hunting, such a list may be useful because it gives an indication of saliency rather than just frequencies. When key words are analyzed, interesting patterns emerge which can act as signposts to the discourses used by the two sides in the debate. By analyzing clusters in which a particular key-word occurs and by plotting links between the collocates, we can explain why a particular pattern is either favoured or disfavoured in the pro-hunt or anti-hunt debate. This case study provides a good example of why both overuse and underuse of certain lexical items compared with other corpora may be interesting to analyze from a social point of view.

Chapter 7 discusses the phenomena nominalization, modality, attribution and metaphor and their role in describing the way we see the world. These phenomena are approached through a detailed analysis of the lexical item *alleged*. The word was chosen because it was regarded as a key concept in constructing discourses about rape. Nominalization offers ideological opportunities (p. 153) because of the reductions and other changes involved in the process. It is for instance shown that the nominalization *allegation* has a discourse prosody associated with denial unlike the non-nominalized form (adjectives, verbs, adverbs). Interestingly the different forms of *allege* are used with different modal forms. There is not a single interpretation for this fact, but it can be hypothesized that the function of denial requires stronger modal forms. By attribution is meant that an analysis is made of the presence and description of different actors mentioned in the text (accuser, accused, victim, hearer of accusation). Metaphors are a particularly interesting area to try to study with the use of corpora. From a discourse point of view they can reveal a lot about the discourse surrounding a particular topic. We are beginning to see several studies of metaphor using corpora. One possible approach (proposed by Sardinha) is to look for strong collocates which are dissimilar in meanings and investigate them in more detail. However, other researchers have been more pessimistic about using corpora to study metaphor, for example Wikberg (2004: 246), who writes that "the nature of metaphor is simply such that it invites much less to quantification than to qualitative analysis", which is "probably one reason why so little has been said about metaphor at corpus conferences so far". Chapter 8 contains the conclusion.

In a general introductory book on the use of corpora for discourse analysis, some readers might have liked more discussion of how corpora can help us to see how discourse is built up. Compare, for example, Sinclair and Coulthard (1975), who in their pioneering work on classroom discourse are above all interested in the linguistic organization of the discourse rather than in the unequal relationship between teacher and pupils. Paul Baker, however, focuses only on discourse as a social phenomenon a view of discourse which is often associated with Critical Discourse Analysis.

The main point of the book is to show that we can combine a fine-grained analysis of linguistic elements with a critical social interpretation of these features. The author shows successfully that both corpus linguists and discourse analysts have something to gain from this. This is also the way forwards advocated by John Sinclair (2001: 168), "the way in which massed corpus evidence can show the ideological trappings or a word or phrase is very good news for those students of discourse who are prepared to accept a moderate discipline of objectivity". Compare also Amy Tsui (2004: 29), who argues that "there is

scope for relating the discourse description to the larger social context", and "that the work of critical discourse analysis and descriptive discourse analysis should be seen as complementary rather than competitive". The best demonstration of this will no doubt be research along the lines suggested by Paul Baker.

The book can also be warmly recommended as a practical and pedagogical handbook for linguists using corpora to study discourse. It offers rewarding reading both for corpus linguists and for discourse analysts. For example, the reader gets a helpful introduction (including figures) to the different steps of using WordSmith Tools, and there is an excellent step-by-step guide to collocational analysis concluding Chapter 5.

### *References*

Sinclair, John. 2001. A tool for text explication. In K. Aijmer (ed.). *A wealth of English. Studies in honour of Göran Kjellmer*, 163–176. Göteborg: Acta Universitatis Gothoburgensis.

Sinclair, John and Malcolm Coulthard. 1975. *Towards an analysis of discourse. The English used by teachers and pupils*. Oxford: Oxford University Press.

Tsui, Amy B. M. 2004. What do linguistic descriptions have to say about discourse? In K. Aijmer (ed.). *Dialogue analysis VIII: Understanding and misunderstanding in dialogue. Selected papers from the 8th IADA Conference, Göteborg 2001*, 25–46. Tübingen: Max Niemeyer Verlag.

Wikberg, Kay. 2004. English metaphors and their translation: The importance of context. In: K. Aijmer and A-B. Stenström (eds.). *Discourse patterns in spoken and written corpora*, 245–265. Amsterdam and Philadelphia: John Benjamins.

**Sabine Braun, Kurt Kohn** and **Joybrato Mukherjee** (eds.). *Corpus technology and language pedagogy.* Frankfurt am Main: Peter Lang Verlag, 2006. 214 pp. ISBN 3-631-54720-X. Reviewed by **Hilde Hasselgård**, University of Oslo.

This edited volume is one of many publications at the moment dedicated to the combination of corpus linguistics and language teaching. Most of its eleven papers were presented at the Language Technology Section of the LearnTec

Conference in Karlsruhe in 2005. What may distinguish this volume from many of its 'siblings' is the frank recognition that "there still remains a wide gap between the wide range of corpus-based activities that have been suggested by applied corpus linguistics and the relatively limited extent to which corpora are actually used in the ELT classroom" (p. 6). Furthermore, the enthusiasm of corpus linguists is not found among students; "on the one hand, students enjoy corpus work and have a positive attitude towards it. On the other hand, there are teaching methods they prefer and feel they profit from more" (p. 84). Several papers propose that language learners require different tools and methods from those used by linguistically minded corpus experts and that this may be part of the explanation for the relative scarcity of corpus use in classrooms.

The papers included in the volume are highly diverse; while some of them report on actual experiences with corpus use with students in both schools and universities, others present corpus materials and tools with a potential for being applied in teaching. One paper has an NLP perspective and thus affects language pedagogy only indirectly. In this review, the papers are not presented in the order in which they occur in the book; instead I have attempted to group them according to common topics.

The book opens with an overview chapter entitled 'Corpus linguistics and language pedagogy: The state of the art – and beyond' by Joybrato Mukherjee. While 'state of the art' is a daunting concept in such a fast-growing field,[1] the chapter nevertheless gives a picture – albeit not exhaustive – of current corpus work and current teaching practice, including the gap between the perceived potential of corpora in language teaching and the "relatively limited extent to which corpora are actually used in the ELT classroom" (p. 6). Concrete ideas are given for corpus-based or corpus-informed activities for classroom use on the basis of standard or 'DIY' corpora. Furthermore, there are suggestions for how learner corpora can be used in the classroom. Moving on from the general overview, Sandra Götz and Joybrato Mukherjee report on the student evaluation of various DDL (data-driven learning) activities they were involved in during the project phase of a linguistics seminar. The majority of students found the DDL activities useful, and most of them increased their interest in DDL activities during the project phase. In spite of this, not many felt that their own language had been improved. However, since many commented on structures and forms they had discovered during the project, the authors suggest that the newness of an inductive approach to language learning led the students to underestimate their learning outcome. An important finding of the project was that a thorough introduction to corpus-learning methods is of the essence: "the acquisition of some

kind of 'corpus literacy' […] seems to be *the* most central prerequisite for successful implementation of DDL activities in the English classroom" (p. 59).

Among the presentations of new corpus materials the use of speech corpora is a particularly interesting new development in the ELT context. Sabine Braun presents the ELISA corpus, a small corpus of interviews with native speakers from various countries and regions. The rich annotation and search facilities that go with ELISA enable learners to search not only for linguistic features, but also for texts on particular topics, thus enhancing the potential applicability of the material in the classroom where many activities are content-driven. The ELISA corpus is too small for investigating low-frequency features, but on the other hand it gives learners easy access to whole texts as well as video recordings of the interviews. Braun suggests that the corpus and software facilities will help learners "bridge the gap between the textual records in the corpus and the discourse situations they have to (re)construct in order to exploit the corpus materials efficiently" (p. 38).

Ulrike Gut reports on the LeaP (Learning Prosody in a Foreign Language) corpus, a corpus of both native speakers and learners of English and German. The corpus is text-to-tone aligned and can thus be used in the teaching of phonology and prosody. It is extensively annotated with orthographic, phonemic and prosodic transcription, POS-tagging, lemmatization, and semantic and anaphoric annotation. Gut gives concrete suggestions on how to use the corpus in teaching (pp. 74–76) and reports on how the corpus was used and evaluated by students at the University of Freiburg. Corpus work as a teaching method was rated higher than student presentations and reading, but lower than lectures and discussion. The majority said they had learned a lot about accents, but few thought they had improved their own accent through working with the corpus.

Christiane Brand and Susanne Kämmerer report on the compilation of the German component of the LINDSEI corpus (the Louvain International Database of Spoken English Interlangauge). Their detailed account of the compilation procedure is no doubt extremely useful for prospective compilers of speech corpora. Particularly the section on problems encountered during the process (p. 133) and how the guidelines for transcribers had to be adjusted may save hours of work for similar project teams.

Several papers emphasize the different perspectives on corpus use of researchers and language learners. Josef Schmied discusses problems of e-learning on the basis of the Chemnitz Internet Grammar (CING). An important insight is that students find inductive learning difficult. A better route for many students is 'rule-testing'. Furthermore, the complexity of the hypertext grammar coupled with a corpus along with the complexity of the grammar to be explored

is bewildering to many learners. Simplicity of the interface and easy navigation are thus crucial. By taking into account the difference between the 'tutor perspective' and the 'learner perspective' the new version of CING is hoped to improve the language learning outcome while also accommodating different needs and learning styles.

Yvonne Breyer's paper points to some of the problems of using standard corpora and corpus tools in the classroom (e.g. curriculum-based needs, vocabulary difficulties, corpora that are too big for the learner to handle). It is argued that "the prerequisites and requirements of research versus classroom application differ markedly", which calls for specialized software tools. *MyConcordancer* is such a "tailor-made approach to classroom concordancing" (p. 157) whose use is simple and intuitive. Some novel features are a field in each concordance line where users can insert their own comments/analysis and sort the concordance accordingly, a dictionary facility that suggests alternatives if a query word is spelled wrongly, and the possibility to save the whole 'workspace' so that unfinished queries and projects can be taken up in the next classroom session where they had to be left off.

Nadia Nesselhauf gives an introduction to the ICLE (International Corpus of Learner English) CD-ROM and discusses the role of learner corpora in L2 research. Although many of the ICLE subcorpora have existed for some time, and have been explored particularly in their country of origin, the CD-ROM is relatively recent (2002). Many of the possibilities built into the corpus and software design have so far been little exploited. Thus more use could be made of recorded variables such as language background (for comparative purposes), number of years of English instructions, and stays in English-speaking countries to investigate how these affect language proficiency.

The remaining papers are less directly concerned with language pedagogy. Chris Tribble presents a case study of how the Keyword and Wordlist functions of WordSmith Tools can be applied to discover explicit and implicit features of content in a corpus of newspaper text (*Guardian Weekly*), thus using corpus software for cultural rather than linguistic analysis. Its application in the language classroom is not stated explicitly, but the case study will be able to serve as a model for similar projects by students.

The paper by Sebastian Hoffmann and Stefan Evert describes 'the marriage of two corpus tools', the BNC*web* and the corpus query processor (CQP). The purpose of this marriage is an improved search tool that can cater for the needs of both specialist and novice corpus users. The interface remains as simple as before (judging from the accompanying screenshots) while there are possibilities of much more advanced searches, in combining lexical and word class

searches and also using regular expressions. The authors also envisage an offspring of this marriage: C*web*, which is expected to be much like the BNC*web* (CQP version) in functionality but independent of the BNC. However, this tool has not yet been developed.

Christoph Müller and Michael Strube describe the annotation tool MMAX2 and how it can be used at various stages in the annotation process. The annotation is added in layers according to the principle of multi-level annotation so that phenomena on different levels can be related without interfering with each other. MMAX2 is also a tool for querying the linguistic annotation. Some of the possibilities are demonstrated towards the end of the paper, along with possibilities of transforming the corpus to other formats and accessing the annotation by means of programming language.

The title of the volume is *Corpus technology and language pedagogy.* As the above summary will have indicated, the corpus technology component is present in all the papers (though with varying degrees of explicitness and technicality), while the language pedagogy component is absent from some of the papers and only implicitly present in others.

The papers following the introductory overview chapter are organized under two headings: 'New Methods' and 'New Resources and New Tools'. Although methods and resources are obviously related in language teaching, it is debatable how suitable the section headings are to the papers that appear in them. For example, Gut and Braun present new corpus resources, but their papers are found in the first section. In contrast, Nesselhauf's paper deals extensively with research method, but is placed in the second section. Tribble's paper is appropriately placed in the methods section, but here the question will be how *new* the method is, as most of the procedures demonstrated are not new to the latest version of WordSmith Tools and have thus been around for some time. In sum, the organization of the book could have been better, making its structure more transparent and consistent with the contents of the papers for the benefit of readers. The importance of this is related to the diversity of topics and perspectives; it is probably unlikely that a reader will find all of the papers useful or fascinating, while it is fairly certain that most readers will find a number of papers to interest them.

Regarding the formal features of the book, it should be mentioned that some of the papers would have benefited from an additional round of proofreading and/or language checking by a native speaker of English.

The great strength of this book is its closeness to teaching practice. The focus on the different needs of learners and corpus experts is particularly useful, as this is probably a key factor for the success or failure of corpus-based lan-

guage teaching methods in schools and undergraduate courses. Most of the papers report on the teaching of English to German-speaking students, but the tools and methods should be easily transferable to other situations. Furthermore, tools and methods are evaluated on the basis of teaching. This is a useful step forward in the promotion of corpus-informed language teaching. The evaluations by students are likewise enlightening, albeit not too encouraging. Teachers wanting to integrate corpus methods in their courses will thus find useful advice in this book on corpus tools and materials as well as concrete tasks and class projects. The awareness of the pitfalls of corpus techniques thrown uncritically (but enthusiastically) at students and of the need for tailor-made software and alternative approaches to the corpus material gives hope that more insights gained from corpus linguistics and the obvious advantages of autonomous, problem-based learning can finally find their way to students through improved tools and teaching methods.

### *Note*

1. An indication of the growth of the field can be gleaned from a Google-search for 'corpus' and 'language teaching', which gave 214,000 results.

**Joybrato Mukherjee.** *English ditransitive verbs. Aspects of theory, description and a usage-based model.* Amsterdam and New York: Rodopi, 2005. 295 pp. ISBN 90-420-1934-4. Reviewed by **Jan Aarts**, University of Nijmegen.

There was a time – many years ago – when it was not unusual to hear conference presentations along the following lines: 'I have counted the occurrences of phenomenon x in genres A and B and found that x is significantly more frequent in B than in A. Isn't that interesting'? And, to tell the truth, very often such findings *were* interesting, if only because they proved a useful tool in varieties differentiation. But they also tended to provoke a 'so what?' reaction, accompanied by the nagging thought that frequency data might be interpreted more meaningfully than as independent, merely numerical facts. This gradually led to the conviction that differences in frequency, in most cases, can and should also be interpreted as pointers to elements in the text that can be seen as providing an explanation of why in a particular context one linguistic element is more often

selected than another. In his ambitious and interesting book on English ditransitive verbs, Joybrato Mukherjee (henceforth JM) gives a central role to this use of frequency data; he establishes a systematic link between frequency and function in what he calls a from-corpus-to-cognition approach to ditransitivity.

In the first chapter an overview is given of the various ways in which ditransitivity has been approached in a great number of linguistic traditions and 'schools': descriptive grammar, generative grammar, valency theory, functional grammar, corpus-based grammar, corpus-driven lexicogrammar, construction grammar and cognitive grammar. A critical discussion of these various approaches lays the groundwork for an eclectic theory of ditransitivity placed in the tradition of (English) descriptive reference grammars. It is corpus-based and follows Langacker's version of cognitive grammar, because this is based on "lexicogrammatical entities comparable to patterns in corpus-driven lexicogrammar" (p. 260). Towards the end of the chapter a working definition of ditransitive verbs is given. A verb is regarded as ditransitive if it requires a subject (S), an indirect object ($O_i$) and a direct object ($O_d$). All three arguments should be realizable as NPs. The pattern S:NP – DV – $O_i$:NP – $O_d$:NP is called the basic form of complementation. If a verb occurs in this basic form in actual language use (i.e. in a corpus) "it is also considered a ditransitive verb in all other forms of complementation" (p. 65). A ditransitive verb has "an underlying proposition that represents the situation type of TRANSFER with three semantic roles involved: the ditransitive verb denotes an action in which the *acting entity* transfers a *transferred entity* to the *affected entity*" (ibid).

In Chapter 2 an account is given of the collection of the data. The primary source is ICE-GB, for the obvious reason that it contains rather exhaustive tagging and parsing information, including seven types of verb complementation. In spite of this rich annotation – or perhaps because of it – there are several discrepancies between the definition of ditransitivity in Chapter 1 and the tagging and parsing found in ICE. One reason for this is that JM's definition of ditransitivity is basically semantic in character; another is that a verb that is just once attested in the basic type of ditransitivity is regarded as ditransitive in all other forms of complementation. The ICE parser, on the other hand, is based on syntactic, and hence predominantly formal, criteria, while at the same time the tagging and parsing of each verb in its specific context reflects its syntactic behaviour *in that context*; in other words, if a verb is accompanied by only one complement, it cannot be ditransitive. A third reason why there is no perfect fit between the ICE analysis and that applied in this book are some straightforward differences in the analysis of specific structures. For example, JM follows the Survey grammars in regarding a constituent as indirect object also if it is

removed from its position immediately after the verb and its semantic role of 'affected entity' therefore needs to be marked by the preposition *to*, since this meaning is no longer signalled by its postverbal position. In such cases the grammar underlying the ICE parser considers the *to*-phrase to be an adjunct and hence the verb as monotransitive. JM concludes his discussion of such discrepancies by saying that "there is a danger, therefore, that already available corpora with their syntactic annotation predetermine the linguistic theory of and research into syntax" (p. 79). I would say that this is a fact rather than a danger, and that it can only be circumvented if one wants to make the effort of re-interpreting the analysis found in the corpus.

From the corpus seventy verbs were collected that occurred at least once in the basic form of ditransitive complementation (S:NP – DV – $O_i$:NP – $O_d$:NP). Together, these verbs were found 1,741 times in an explicit ditransitive pattern. By comparing the overall frequency of each of the seventy verbs with the number of times it occurs with an explicit ditransitive syntax, three groups of verbs were created reflecting a gliding scale of ditransitive typicality. The first group, that of 'typical ditransitive verbs', which are used very frequently in general as well as with an explicit ditransitive complementation, comprises GIVE (with 562 'explicit' occurrences) and TELL (491). Next comes a group called 'habitual ditransitive verbs' consisting of ASK (91), OFFER (54), SEND (79) and SHOW (84). The third group is called 'peripheral ditransitive verbs' whose explicit ditransitive use varies between one and 34 occurrences. This group includes verbs such as AFFORD (4), DRAW (1), SAVE (4) and THROW (2). For an examination of the syntactic behaviour of such low-frequency items the BNC was used.

The rest of Chapter 2 is devoted to a discussion of the nature of linguistic description on the basis of corpus data, and a comparison of the corpus-based and the corpus-driven approaches ("a strictly corpus-driven approach [is] not a viable option": p. 261) and is concluded with a quotation from Schmid (2000: 39) which expresses the guiding principle of the 'from-corpus-to-cognition' approach exemplified in the book under review: "Frequency in text instantiates entrenchment in the cognitive system".

Chapter 3 forms the core of the book. Here JM gives a highly detailed analysis of the verbs constituting the groups of typical and habitual ditransitives, resulting in an integrated quantitative and qualitative description of each verb. First of all, the various syntactic patterns in which each verb occurs in the corpus are described. Five major patterns are distinguished. The first of these is the pattern of ditransitivity in which both the affected entity and the transferred entity are realized as noun phrases. In the second pattern the affected entity is

not realized by a NP, but by a *to*-phrase in final position. The third pattern is a monotransitive one, with only the direct object realized, and the affected entity 'understood'. In pattern four both the affected entity and the transferred entity are understood, so that the pattern is intransitive, syntactically speaking. A fifth pattern realizes only the indirect object ('John told me'). Not all verbs occur in all of these patterns; GIVE, for example, does not occur in the fifth pattern, TELL occurs in all five. Within almost all patterns several subpatterns are distinguished. These are variations on the five major patterns in terms of the realization of syntactic functions (e.g. clause instead of NP), lexical variation (e.g. *for* instead of *to* in the prepositional phrase denoting the affected entity) or changes in word order caused by general syntactic operations (passivization, relativization). The total number of patterns thus increases considerably; GIVE has a total number of 24 patterns, the total number for TELL is 23. For each verb the absolute number of times it occurs in a given pattern is presented, as well as the percentages of each pattern in relation to the total number of occurrences. In the further functional analysis of the verbs only the more frequent patterns are examined. In the case of GIVE, for example, this means that eight patterns, together covering 91.2 per cent of all the occurrences are taken in consideration. The assumption is that such a percentage will reflect the 'routinised core area' of language use. It is JM's conviction "that a model of lexicogrammatical routines in language use should not attempt to explain each and every occurrence in performance, but abstract away from the performance data a model that is able to account for some 90% of all cases" (p. 100).

After this inventory of patterns and their quantification, the next question to be answered is "which principles and factors cause language users to choose a specific pattern" (p. 101). For an answer to this question it is first determined which of all the patterns associated with a given verb is the 'default pattern'. This decision can be based on a number of factors. A very important one is, of course, the frequency of a pattern. An obvious candidate in the case of GIVE is the pattern with two objects, both realized by NPs, which makes up 38 per cent of all the occurrences. A second factor taken into account is the 'structural simplicity' of the pattern. But apart from quantitative and structural considerations other factors are also taken into account. In the case of TELL, for example, a third factor overrides the structural simplicity criterion. Here, the pattern that has a *that*-clause as the direct object, is selected as the default pattern, not only because it is by far the most frequent, but also because TELL, as a verb of verbal communication, selects most naturally a *that*-clause as the means to convey the content of the message which is the regular transferred entity.

For each of the typical and the habitual verbs a description is given of the 'principles of pattern selection', that is, of the contextual conditions that favour the choice of one particular pattern over the other patterns "in a more or less significant proportion of all relevant instances" (p. 262). These conditions vary from verb to verb, but they are, on the whole, of the following kinds: the presence or absence of the need to specify the semantic roles involved; pragmatic factors, such as the principle of end-weight; lexical preferences, when the presence of a given lexical item in one of the semantic roles tends to trigger a particular pattern – for instance, the use of the noun *example* realizing the role of transferred entity with the verb GIVE, favours a pattern without an indirect object. The identification of the principles of pattern selection is a question of interpretation of course (although to some extent supported by quantitative considerations), but on the whole the arguments adduced for the principles are quite convincing.

After the discussion of patterns and pattern selection principles for typical and habitual ditransitive verbs, attention is given to the group of peripheral verbs. Since peripheral verbs are not or very rarely attested in explicit ditransitive patterns, they are not treated separately, but as a group. The discussion is put within the perspective of a process of grammaticalization, which starts with the question which verbs can emerge in a ditransitive pattern in the first place. The meaning of such potentially ditransitive verbs can be seen as being "in line with" (p. 205) or having the potential to be extended (metaphorically and/or by analogy) to the ditransitive situation schema. The verbs, however, are not yet attested in the basic ditransitive pattern. Once a verb has emerged but still occurs infrequently with the basic form of ditransitive complementation, it is regarded as a 'grammatically institutionalised ditransitive verb'. If after that the frequency of the verb's explicit ditransitive use passes a critical threshold, it becomes 'a conventionalised ditransitive verb'. The process of grammaticalization can thus be seen as a reflection of the classification into peripheral, habitual and typical ditransitive verbs. In contrast to the meticulous and factual analysis of the typical and habitual verbs in the earlier sections of this chapter, the discussion of the peripheral verbs is a bit speculative and intuitive. Sometimes this leads to vagueness and confusion. In Figure 3–14 (p. 205), for example, which visualizes the successive stages of institutionalization and conventionalization, the formal criterion for a potentially ditransitive verb stipulates that "the verb is *not attested with the basic form of ditransitive complementation*" [italics JA]. For a grammatically institutionalized verb it is said that it is "infrequently attested" in this way. For a conventionalized verb the formal criterion is that it "is ± frequently attested with the basic form of ditransitive complementation".

At the same time and in the same column of the figure, a conventionalized verb is semantically characterized as a verb which "is, by default, associated with the ditransitive situation schema (*even if it is not used in the basic form of ditransitive complementation*)" [italics JA]. Clearly, there is a contradiction between the two italicized passages: a verb cannot be a 'potentially ditransitive verb' and a 'conventionalized ditransitive verb' at the same time. In the discussion of the figure this contradiction is not solved, so that the reader begins to suspect his understanding of the chapter is insufficient.

All this is not to say that the model that is sketched of the grammaticalization of potentially ditransitive verbs is not an attractive one, based as it is on frequency and the analogous power of the lexicon. If one subscribes to the viewpoint that frequency is a measure of the entrenchment of words, phrases and lexico-grammatical structures in the cognitive system, it follows that it should play a prominent role in the description of grammaticalization processes. The other valuable notion is that of the potentiality of verbs to extend the number of semantic roles with which they are associated from two to three, enabling basically monotransitive verbs to be used ditransitively. The notion is intuitively appealing, and might be extended from single lexical items to classes of lexical items. For example, verbs of creation and transformation (cf. Levin 1993) like *bake a cake, chop some wood,* are basically mono-transitive, but can easily be extended to admit in their situation schema an affected entity for whose benefit the act of creation or transformation is performed: *I'll bake you a cake, chop you some wood.*

In Chapter 4 JM sketches a model of "speakers' linguistic knowledge about ditransitive verbs" (p. 221) using the corpus findings discussed in the two previous chapters; it represents "a cognitive-linguistic abstraction of language use" (ibid.) and is lexicogrammatical in character in that it ignores the traditional boundary between lexis and syntax. Four principles are formulated to which such a model should adhere (and which were applied in the analysis of the verbs in the third chapter): 1. it should be based on the analysis of large amounts of representative corpus data; 2. it should include frequency data which are combined with and supported by qualitative considerations; 3. it must be expressed in terms of lexicogrammatical patterns as basic units; 4. it must "distinguish between the routinised core area and the creativity-guided periphery of language use" (p. 263). The first principle is a bit problematic, especially in relation to the fourth. It would be difficult to maintain that there are corpora at the moment that are sufficiently large and representative *and* have the level of annotative sophistication that is required to automatically search for complex syntactic and lexical patterns. Consequently, as the author admits, while it is possible to give an

account of the core, the insufficient availability of data makes it impossible to also give an account of the periphery, although of course it *is* possible to make the distinction between core and periphery on the basis of the frequency data. Models are therefore only presented for the typical ditransitives GIVE and TELL. The author sketches out a network-like model of ditransitivity incorporating the lexical set of ditransitive verbs, a set of principles of pattern selection and a set of ditransitive patterns. Seen in the light of what has been presented in the preceding chapters, the model is quite convincing in filling out the notion of a from-corpus-to-cognition model.

Chapter 5 provides a summary of the preceding chapters and ends with a few suggestions for further research.

To conclude this review, I want to bring up two general points that do not quite fit in the above chapter-by-chapter account. The first concerns the place of syntactic functions in the proposed model. Syntax plays its most prominent role in the verb patterns. The patterns are presented in the form of a string which reflects the word order of the pattern and contains information about the pattern's syntactic functions and the syntactic categories by which these are realized (see the example of the basic pattern of GIVE above). The definition of ditransitive verbs, on the other hand, is basically semantic, conceived in terms of semantic roles – quite rightly, I think. But the relation between the semantic and the syntactic level is not straight-forward. There is a one-to-one mapping between semantic roles and syntactic functions and categories only in the case of the basic ditransitive pattern; with the pattern variants, uncertainty creeps in, due to the different ways in which syntactic functions may be interpreted. I have already given the example of the *to*-phrase in sentences with GIVE, which is interpreted as an indirect object by JM, while the ICE-parser labels it an adjunct. The former interpretation seems to me entirely semantically motivated, while the latter is not only motivated by the variant word order, but also by the (semantic) fact that loss of the postverbal position also entails that the NP in question is no longer marked as the affected entity, a role that must now be explicitly expressed by the preposition *to*. Such differences in the interpretation of function labels result in uncertainty about what the function labels in the patterns 'mean' and on the basis of what criteria they are attributed to constituents. Moreover, their only usefulness in the patterns seems to be that they can be used as cover terms for more than one linguistic category. That being so, I think the patterns would gain much in clarity if they were not expressed in terms of syntactic functions; instead, they could be expressed directly in terms of semantic roles and the syntactic categories associated with them. Another point is the delimitation of the class of ditransitive verbs. JM sets the membership of this

class on a firm footing by applying the formal criterion that a verb should occur at least once in the basic pattern of ditransitivity. This can have some freakish results, however. It makes, for example, the verb *chop* a full-fledged member of the class (see above), while *address*, as in *she addressed her remarks to the children* (p. 11) is only a potentially ditransitive verb, in spite of the fact that *address* is much more strongly associated with the situation type of TRANSFER than *chop*. It might be better to apply the principle of 'frequency in use is entrenchment in the system' also to the question of class inclusion, and make primary membership of the class of ditransitives dependent on frequency of occurrence. Bearing in mind that many verbs can occur in a number of complementation classes, it would seem natural to assign a verb for its primary membership to the complementation class in which it occurs most frequently. This would doubtlessly put a verb like *chop* in the class of monotransitives.

Something should be said about the more formal features of the book. It is carefully edited: there are almost no misprints. It has some reader-unfriendly features though. One of these is the far too great number of footnotes, which, in spite of their relevancy, break up the line of argumentation in the main text. More trivial – but just as irritating – is the way patterns and subpatterns are referred to. This is done by using Roman numerals and lower-case letters; but even with this knowledge it remains difficult not to read combinations like IP or If as alphabetical sequences instead of alphanumeric ones.

The book is a convincing exercise in corpus-linguistic description and methodology. It is an important step forward in the development of corpus linguistics as a linguistic discipline in its own right, thanks to its careful and detailed description of the corpus data and the integration of quantitative and qualitative analysis. No one who is interested in corpus-linguistic methodology or, more specifically, in ditransitive verbs, can afford to ignore it.

## *References*

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago and London: The University of Chicago Press.

Schmid, Hans-Jörg. 2000. *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: Mouton de Gruyter.

**Junsaku Nakamura**, **Nagayuki Inoue** and **Tomoji Tabata** (eds.). *English corpora under Japanese eyes* (Language and Computers: Studies in Practical Linguistics 51). Amsterdam and New York: Rodopi, 2004. xi + 249 pp. ISBN: 90-420-1882-8. Reviewed by **Shunji Yamazaki**, Daito Bunka University.

This book commemorates the tenth anniversary of the Japan Association for English Corpus Studies (JAECS), a unique national association of corpus linguistics, with a membership over 350. The association has been continuing its dynamic academic activities, holding biannual conferences, and publishing an annual journal, *English Corpus Studies*. The current book is a sequel to the first description of corpus-based activity on English in Japan (*English corpus linguistics in Japan*, 2002, Rodopi, edited by Toshio Saito, Junsaku Nakamura, and Shunji Yamazaki), which included eight papers on contemporary English analysis, nine papers on diachronic corpus analysis, and two papers on English language teaching as well as one explanation of software for analyzing corpora. Most of the contributors were members of JAECS.

The latest book covers a wide range of corpus-based or corpus-driven research studies including synchronic or diachronic studies of English and applications in literary analysis or in English language teaching, and also clearly shows some current trends of corpus linguistic studies in Japan. The foreword and preface succinctly outline the history of JAECS and the ways in which the book has been published. There are five sections in the book: the first includes Stig Johansson's paper, 'Overview of Corpus-based Studies', followed by 'Corpus-based Studies in Contemporary English' in section 2. Section 3 deals with 'Historical and Diachronic Studies of English', one of the most popular corpus research areas in Japan as demonstrated by the length of the section. There are comparatively few papers in the last two sections on 'Corpus-based Studies in English Literature' and 'Corpus and English Language Teaching' respectively, albeit these two research areas are well-researched in Japan.

In his keynote article, Johansson clearly summarizes "an overview of development in the use of corpora in English language research" from corpora before computers to the vast collections of corpora available now, focusing particularly on ICAME which has become a world-wide community of scholars of corpus linguistics. Though a considerable amount has been achieved, he stresses that there is a great potential for further research in future in the areas of language variation, such as lexis, grammar, and contrastive linguistics, and concludes that we are still at the beginning of the era of corpus linguistics.

There has been a rapid growth of interest in contrastive linguistics using parallel corpora in Sweden, Norway, and some other European countries. The book contains two noteworthy contrastive papers utilizing two languages: English and French by Uchida and Yanagi, and English and Japanese by Shimizu and Murata. The former deals with 'copula and infinitive' constructions (English *be to* and French *être à*) in English and French, and finds that the French form is less frequent than the English construction despite the fundamental similarities in form and function. The analyses of examples reveal French preferences for lexical expressions of modality, non-passive voice, and a human argument in its subject position. By contrast, the latter paper sheds light on patterns with transitive verbs and reflexives in English and Japanese (*Vt reflexive (prep)* patterns) by using the BNC and an English-Japanese dictionary, and confirms that their earlier findings of the majority of English reflexives do not correspond to Japanese reflexives, but to several types of Japanese expressions. As a rule-based grammar cannot predict the idiosyncrasies of the relation between a word in one language and its translation in another language, they strongly suggest that a bilingual pattern grammar (BPG) should be constructed in order to map a pattern in one language into another pattern in another language.

Kimura used the BNC and *OED2* for the comparison of a relatively new loanword *tycoon* and a long existing word *magnate* in order to demonstrate the similarities and the differences in the use of synonymous words in English. *OED2* was used in order to reveal the origin and development of these words. Different semantic categories of use are identified for *magnate*, 'a man of wealth', and *tycoon*, initially meaning 'shogun' but more recently 'a dominant person in business'. At the same time, *magnate* appears in the 'art' domain and more in books, whereas *tycoon* has more negative connotations and is used in the 'leisure' domain. The findings of different origin and development of these sample words strongly suggest the need for further research using other pairs such as *tidal wave* and *tsunami*, or *reckless* and *kamikaze* in order to illustrate the process affecting the use of a relatively new loanword alongside a long existing word.

After an earlier study Milsark (1974) claimed that definite NPs (*the* + NP) do not occur as notional subjects in existential *there* constructions (i.e. the 'Definiteness Restriction'), several attempts have been made to classify exceptions to its claim (Lakoff 1987; Lumsden 1988). The paper by Nishibu presents quantitative research undertaken in order to describe the characteristics of '*the* + NP' existential constructions by using the BNC. He provides some interesting information: 1) approximately 80–90 per cent of notional subjects in the construc-

tions are indefinite nouns and about five per cent '*the* + NPs', and the Definiteness Restriction is more binding in the written register; 2) cataphoric use of the definite article (66.3%) is most common in the samples of '*the* + NP' subjects in the written register, and an abstract noun headed by a cataphoric *the* is most typical '*the* + NP' subject; 3) the listing function produces '*the* + NP' subjects at high frequency in both registers. Nishibu has theoretically challenged the previous common understanding of notional subjects in existential *there* constructions, but it might be necessary to verify his findings by using different regional corpora such as American, Australian, or New Zealand corpora.

Although a previous study (Wright and Hope 1996) claims that lexis has been considered less important than syntax in stylistic studies, Takami suggests that lexis can be recognized as the more important, powerful element in stylistic studies, as exemplified by Leech (1974), and Takami clearly explains it by demonstrating a method to identify "the words which show significantly different frequencies between two types of text groups with the emphasis on corpus-driven evidence using the log-likelihood ratio" (p. 131) and presenting a set of words which make a significant contribution to stylistic differences by using three broadsheet (*Independent, Guardian, Times*) and two tabloid (*Today, Sun-now*) British newspapers in the Bank of English. Takami's method has been effectively applied to detect lexical differences of adjectives between two types of newspapers, and has interestingly elucidated the different use of adjectives in the British newspapers; broadsheets display "a special preference for words of possibilities and rationality as well as technical terms of society and culture" (p. 131), whereas tabloids emphasize "people's nature, characteristics or emotions" (p. 131). The paper clearly shows the effectiveness of lexis in stylistic studies, summarizing very succinctly the analysis of two types of adjectives: tabloid adjectives are more about people, whereas broadsheet adjectives rather describe society and culture.

Despite the fact that there are only three papers in this book on historical and diachronic studies of English, it is undoubtedly appropriate to say that there has been tremendous interest in this area of corpus studies in Japan, as the first corpus-based research book (Saito et al. 2002) shows. The paper produced by Nakao, Jimura, and Matsuo is an interim report of a project for a computer-assisted comprehensive textual collation between the Hengwrt Manuscript and the Ellesmere Manuscript of *The Canterbury Tales*. They toiled away at comparing the two manuscripts line by line, word by word, and made a comprehensive collation between these two manuscripts by using the machine-readable text compiled by Stubbs (2000). For all that, the paper is an interim report; they logically outline their project, objective, database, methodology, and some future

research directions. The completion of the collation of the two manuscripts can be expected to reveal the similarities and differences between the edited texts and the manuscripts they used.

Interesting research combining two separate linguistic perspectives, generative grammar and historical linguistics, is reported on in Ohkado's paper, entitled 'On Verb Movement in Old English Subordinate Clauses', which clearly illustrates that there are considerable amounts of corpus research done in this field and in literature in Japan. Ohkado's main aims are 1) to investigate "whether or not an independent leftward verb movement operation should be assumed in Old English subordinate clauses" (p. 151), with special emphasis on the position of objects in relation to finite or nonfinite verbs in subordinate clauses; i.e. (S)VO/(S)OV patterns in subordinate clauses with finite main verbs and nonfinite clauses, and 2) to investigate whether or not the higher frequencies of SVO patterns in subordinate clauses with finite main verbs observed in some of the texts can be accounted for in terms of the notion of embedded main clauses. The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English has been used in this paper, and the findings show a statistically significant correlation between the frequencies of (S)VO/(S)OV patterns in subordinate clauses and nonfinite clauses, illustrating, in opposition to previous studies (Pintzuk 1991, 1993, 1999), that an independent leftward verb movement operation is not motivated, and "higher frequencies of SVO patterns in subordinate clauses with finite main verbs than in nonfinite clauses observed in some of the texts can be accounted for in terms of the notion of embedded main clauses" (p. 163).

Tsukamoto, whose well-known free analysis software (*KWIC Concordance for Windows*) was developed in Japan, illustrates an effective application of a statistical technique to text analysis, employing 'multiple regression analysis' with seven variables (non-argument NPs, WH-words, free relatives, untensed auxiliary verbs, floated quantifiers, negation, and degree complement subordinate clauses) in order to estimate the date of texts, using only internal information by utilizing the Penn-Helsinki-Parsed Corpus of Middle English Phase 1. His equation technique has proved effective with two thirds of texts within ± 50 years deviation, and 45 texts among 93 texts are shown to be dated correctly to within ± 30 years, and 21 texts to within ± 15 years. His equation is also verified by being applied to independent texts of Middle English, and the three variables (WH-words, untensed auxiliary verbs, and negation) can be seen as showing prominent developments in the history of English.

Not only synchronic and diachronic studies of English language, but also corpus-based studies in literature and in language teaching are represented in the

contributions by Ishikawa, Kaneko, and Chujo. Ishikawa shows how much the eleven colour terms (*black, blue, brown, green, grey, orange, pink, purple, red, white,* and *yellow*) in the novels of D. H. Lawrence deviate quantitatively from the norms elsewhere in English and how importantly the four key colour terms (*black, white, blue,* and *grey*) function in his novels. The analyses of three separate corpora, the Lawrence Novels Corpus (LNC), the BNC, and the English Novels Corpus (ENC), show that the eleven colour terms are used much more in the LNC than in other novels, and that those four key colour terms play important roles in Lawrence's novels, with especially black and white being extraordinarily frequently used in his later works.

The next two papers are concerned specifically with the Japanese English language teaching and learning situations and the findings undoubtedly give worthwhile suggestions. As one of the members of the team compiling the Japanese portion of the Louvain International Database of Spoken English Interlanguage (LINDSEI), Kaneko has been carrying out several studies on Japanese learners of English, and raises two questions in this paper: 1) to what extent do Japanese-speaking learners of English use the past tense forms correctly? 2) does the lexical meaning of verbs affect the learners' use of the past tense verbs? Her findings using the 53 LINDSEI interview sub-corpus files (38,767 tokens and 2,702 types), which were transcribed manually and tagged for four types of verbs (regular verbs, irregular verbs, *be* forms, and auxiliary verbs), show that the accuracy rates for those four verbs range from 50 to 66 per cent, with the irregular verbs being used most correctly (65.7%), and *be* forms, least correctly (50.4%). Japanese learners of English tend to mark the past tense more easily in sentences with 'event' and 'activity' verbs than in sentences with 'state' verbs (which is compatible with the finding of Andersen 1991).

Chujo's thorough study reports on a means to compare the vocabulary levels of Japanese textbooks, college qualification tests, and proficiency tests in order to determine what the levels of English used in those materials are, and how many more vocabulary items are required for students to understand 95 per cent of the materials as Nation (2001) has suggested is the language knowledge threshold. She has created a lemmatised and ranked high frequency word list (BNC HFWL) from the BNC, and used Japanese junior and senior high school (JSH) texts, college qualification tests, English proficiency tests such as Eiken (an English test carried out nationally in Japan), TOEIC, TOEFL, and some college textbooks. Her findings clearly suggest that the Daigaku Center Nyushi (DCN, a unified university entrance examination) and the JSH textbook vocabulary have similar levels, whereas many college entrance examinations are higher than the JSH levels, which indicates that "the college qualification tests need

more careful consideration of JSH textbook vocabulary" (p. 245). Additionally, the Eiken second grade test and TOEIC vocabulary are similar to JSH vocabulary, though TOEFL and some ESP textbooks require a greater knowledge of vocabulary. The paper gives useful suggestions and implications in terms of English learning and teaching in Japan, and shows that software development can have a huge beneficial impact on textbook selection and test design in Japan.

Although corpus-based research in JAECS has continued to be vigorous as the book shows, there is room for further development in corpus-based studies in Japan. Firstly, many more papers could be written in English in the journal *English Corpus Studies*. Secondly, more corpus linguists could participate in and give papers at the ICAME or similar international conferences. At the same time, leading scholars from overseas could be invited to lecture in Japan. These efforts will ultimately lead to more collaboration and integration with the research of international corpus linguists, which, in turn, could greatly benefit all corpus linguists working in Japan.

## *References*

Andersen, Roger W. 1991. Development sequences: The emergence of aspect marking in Second Language Acquisition. In T. Huebner and C. Ferguson (eds.). *Second Language Acquisition and linguistic theories*, 305–324. Amsterdam: John Benjamins.

Lakoff, George. 1987. *Women, fire and dangerous things*. Chicago: University of Chicago Press.

Leech, Geoffrey. 1974. *Semantics*. Harmondsworth: Penguin.

Lumsden, Michael. 1988. *Existential sentences: Their structure and meaning*. London: Croom Helm.

Milsark, Gary L. 1974. *Existential sentences in English*. Ph. D. Dissertation, MIT.

Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Pintzuk, Susan. 1999. *Phrase structures in competition: Variation and change in Old English word order.* New York: Garland.

Pintzuk, Susan. 1993. Verb seconding in Old English: Verb movement to Infl'. *The Linguistics Review* 10: 5–35.

Saito, Toshio, Junsaku Nakamura and Shunji Yamazaki (eds.). 2002. *English corpus linguistics in Japan*. Amsterdam and New York: Rodopi.

Stubbs, Estelle (ed.). 2000. *The Hengwrt Chaucer digital facsimile*. Leicester: Scholarly Digital Editions.

Wright, Laura and Jonathan Hope. 1996. *Stylistics: A practical coursebook*. London: Routledge.

**Antoinette Renouf** and **Andrew Kehoe** (eds.). *The changing face of corpus linguistics* (Language and Computers 55). Amsterdam and New York: Rodopi, 2006. 408 pp. ISBN 90-420-1738-4. Reviewed by **Claudia Claridge**, University of Kiel.

The title of the present volume is intended to highlight developments within the field of corpus linguistics, in particular the increasing breadth as well as depth of compilation and research activities, the ongoing readjustment of the concept of 'corpus', the integration of several strands of synchronic and diachronic research, and, not least, the growing amount of methodological and theoretical self-reflection. Those aspects were amply exemplified at the 2003 ICAME conference, as the twenty-two papers and conference-concluding panel discussion collected in this volume bear witness.

Ten papers discuss new collections of data or new ways of presenting, retrieving and analysing data, underlining both the fundamental importance of data management for the field and the fact that in spite of all the progress made in this area there is still much left to do. The contributions by Stefan Dollinger and Clemens Fritz introduce two new corpora, the *Corpus of Early Ontario English* (CONTE, 1776–1899, 225,000 words) and the *Corpus of Oz Early English* (COOEE, 1788–1900, 2m. words). These fill a substantial data gap for two important varieties of English. Both compilers had to answer the crucial question of which texts were to count as Canadian or Australian. Dollinger's solution is a geographical one: texts written in Ontario are included, with the added condition that the writers must have been resident there for a considerable time. Fritz proceeds in a similar way, but also admits texts written in surrounding areas – a potentially problematic approach, as the latter also include New Zealand. Ian Lancashire's paper introduces the *Lexicons of Early Modern English* (LEME, an expansion of EMEDD, the Early Modern English Dictionar-

ies Database). LEME is a database of monolingual glossaries, bilingual dictionaries, lexical encyclopedias and linguistic treatises written between 1477 and 1702 together with an internet-based search interface. It is the intention of Manfred Markus that Hedgehogs, a collection of *Historical English Dictionaries, Grammars and Educational Handbooks of German Schools* (1700–1850), will become a corpus at some future date, but some of the material is currently accessible simply as OCR image scans on the web. Both projects add a valuable metalinguistic perspective to the corpus-linguistic exploration of (Early) Modern English, in the latter case also opening up a potential link with present-day EFL research. The paper by Antonio Miranda García, Javier Calle Martín, David Moreno Olalla and Gustavo Muñoz González illustrates the usefulness of the Old English Concordancer (OEC) and also the Morphological Analyzer of Old English Texts by applying them to *Apollonius of Tyre*. The OEC enables, among other things, the analysis of the distribution of morphological classes and inflections, which is of especial importance for Old English.

While the above papers dealt with originally non-electronic texts, the following contributions emphasize the growing importance of web texts as a source of data. Both Barry Morley and Andrew Kehoe present new features of WebCorp, which make a wider range of research questions possible and offer output possibilities similar to non-web tools. Of special importance are the restriction of searches to web domains in order to create web sub-corpora, the use of date heuristics to enable short-term diachronic research, and the enhanced possibilities for research into phraseological combinations. The WebPhraseCount tool, described in the contribution by Josef Schmied, is geared towards a purely quantitative approach by producing statistical output on the absolute and relative frequencies of search terms/phrases in particular countries or regions, delimited with the help of web domains. A drawback of this tool so far is that the raw data (in KWIC or other format) is not part of the output, so that the user cannot check how much potentially irrelevant data may have entered into the statistics. Cedrick Fairon and John Singler present GlossaNet, a system that allows the use of user-determined specialized newspaper corpora for monolingual or comparative multilingual studies and, as the system makes use of dynamic corpora, for documenting developments over time. The system works in combination with a parser, so that it allows the user to process more sophisticated queries using morphological constraints. While all of the above tools seem highly useful, there remains the question of their availability. In four cases, no mention is made of when and where the corpus or tool described will become publicly accessible (the Ontario and Australian Corpora, the Old English Concordancer and Morphologizer, WebPhraseCount). And despite ongoing corpus-

compilation activity, there are of course still gaps, as is evident in the papers by Marianne Hundt and Ute Römer, who, while also using existing corpora, had to resort to self-provided text collections for dealing with relatively rare features (mediopassive) and with specific research fields (EFL).

The concept of 'corpus', or of what constitutes suitable material for a corpus-linguistic study, which underpins the papers of this collection is evidently subject to a wide range of interpretations, reflecting the different questions and needs with which corpus linguists approach their data. Corpus material can be very small in size and specialised (Römer's 100,000-word EFL corpus); fairly small and specialised (Meurman-Solin: *Corpus of Scottish Correspondence*); fairly small and general-purpose (the Brown family as used by Leech and Smith); big, general and representative (the BNC, used by Deutschmann); big and unpredictable in its composition (internet data). It may be spoken (Stenström: COLT) or written, annotated (Ozón: ICE-GB); contain metadata (Lancashire: dictionaries); and even be non-computerised (Hundt's mail order catalogues). On the one hand, this may reduce the comparability and replicability of the results obtained; on the other hand, it will lead to a more comprehensive description and analysis of English. Indeed, Christian Mair argues for a 'vast and expanding corpus-linguistic working environment' in which the researcher will choose, from among a multitude of corpus(-like) sources, those most relevant to the question at hand. He argues both for small corpora, which need to be improved, in particular with respect to annotation and speech representation, and for "big and messy" corpora like the web – and especially for integrating the two approaches.

Methodology also plays a more explicit role in some other papers and in the panel discusssion. Bas Aarts, actually not unlike Mair, takes an 'instrumentalist' approach, advocating the use of whatever kind of data seems useful and a wide understanding of 'corpus'. The discussion generally reveals an openness for using various kinds of data. Elena Tognini Bonelli argues strongly for observational adequacy and the corpus-driven approach; the latter is applied in the contributions by Römer and Mahlberg, while the other papers can rather be characterised as corpus-based. However, as the discussion shows, these two approaches might not be irreconcilable after all, but simply represent a slightly different focus. The other major topic of the panel concerns the connection between corpus linguistics and the writing of reference grammars. Joybrato Mukherjee argues both in the discussion and in his paper for a truly corpus-based reference grammar of English, one which is fully transparent in its use and analysis of data, balanced between comprehensive and genre-specific description, and open to constant modification. In contrast, Bas Aarts and Mair

both have reservations about the need for or even possibility of a large-scale corpus-linguistic approach to grammar writing. Geoffrey Leech sketches out a gradient of grammars from the 'most theoretical' (not corpus-related) to the 'most observational' (corpus-driven) types, apparently accepting the need for all of these grammar types. One interesting point arising out of the discussion is the necessity of a work that treats the interface between lexicon and grammar in a more satisfactory way – an aspect also taken up by Michaela Mahlberg's paper in this collection.

The breadth of thematic coverage provided by the papers is impressive, ranging from Old English to 21st-century English, from native varieties to ESL and EFL varieties, and from orthography to discourse characteristics. Charting the evolution of the predictive function of *shall* and *will* from Middle English onwards with the help of the *Helsinki Corpus*, Maurizio Gotti finds confirmation of previously proposed grammaticalisation scales and places the shift from *shall* to *will* in EModE. Anneli Meurman-Solin and Päivi Pahta's thought-provoking paper identifies the discoursal functions of the connectives *seeing* and *considering* in the *Corpus of Scottish Correspondence* and the *Corpus of Early English Medical Writing* as cohesive, focusing, narrative and argumentative, with the latter being the most frequent. Caren auf dem Keller's contribution adds to our knowledge about the evolution of text types by chronologically tracing three basic models of book advertisements in the ZEN corpus. Marianne Hundt finds a significant increase of mediopassive constructions in advertising language throughout the 20th century, thereby contradicting Leech's (1966) claims that verbs are inconspicuous and that there was no important change after the 1920s in this text type. She shows how mediopassives are in fact a logical element within the typological development of English. Unfortunately, the diagrams in this very interesting contribution are not as clear as one might wish. Comparing various grammatical features in 20th-century British and American English with respect to change induced by Americanisation and/or colloquialisation, Geoffrey Leech and Nicholas Smith find some proof for these tendencies, but also contradictory evidence. They rightly point out that such general explanatory labels need to be used with care. These last two papers, as well as the contributions by Mair, Kehoe and Fairon and Singler, attest to the ever increasing attention paid to change in progress in Present-day English within corpus-linguistic studies.

The remaining papers present synchronic analyses. Mats Deutschmann's investigation of explicit apologies in the BNC successfully couples corpus-linguistic methodology with pragmatics and sociolinguistics. He attributes the higher apology rates of younger and middle-class speakers to a high involve-

ment style, a form of social conditioning and the negative politeness culture of dominating social groups. Göran Kjellmer's succinct study of resolving ambiguity induced by the polysemy of *recent* points to five possible strategies used by addressees, involving semantic, grammatical, pragmatic and contextual factors. Ute Römer takes an applied perspective, as her aim is for corpus linguistics to contribute to a greater degree of authenticity of the English used in teaching contexts. Based on the example of *looking*, she shows that German textbooks of English inconsistently either under- or overrepresent the amount of contracted progressive forms as well as the instances of *looking* followed by *at* and *for*, and completely neglect the functions of expressing repeated actions and general validity. Gabriel Ozón uses two parameters to approach the variation between the double object construction and the NP-PP complementation with ditransitive verbs, namely medium (written vs. spoken English) and information structure. While medium does not influence the choice of construction at all, the 'given before new' information hypothesis is not fully supported by the data and needs to be supplemented by other factors, such as the concept of focal information. The last paper in the synchronic section continues the emphasis on spoken English noticeable in Deutschmann's, Römer's and Ozón's papers. Anna-Brita Stenström's contrastive approach to the Spanish pragmatic marker *pues* and its closest English equivalents shares with Römer's paper a potentially applied perspective, though this is not made explicit here. Stenström finds that *pues* shares four functions with English *cos* and eight with *well*, but also corresponds sometimes to *therefore, okay, yeah* and even zero. Michaela Mahlberg's paper revolves around the high-frequency noun *time*, whose 'investing time' pattern she investigates with the help of Hunston and Francis' Pattern Grammar approach. In the process she shows how the pattern approach might be improved by taking lexical items as reference points in order to add detail to the description and to enable a better grouping of patterns.

It is impossible to do justice to individual papers within such a large and varied collection in the space of a brief review. I have therefore tried to highlight the developments in methodology and their reflection within the field as perhaps the most important aspect of this collection. Beyond this, the volume attests to the richness of corpus linguistics, and its ability to incorporate and benefit from research questions from diverse fields.

## *Reference*

Leech, Geoffrey N. 1966. *English in advertising: A linguistic study of advertising in Great Britain*. London: Longman.