Reviews

Karin Aijmer. *English discourse particles: Evidence from a corpus*. Studies in Corpus Linguistics 10. Amsterdam and Philadelphia: John Benjamins, 2002. xv + 298 pp. ISBN 90-272-2280-0 (Eur.)/ 1-58811-284-5 (US). Reviewed by **Laurel J. Brinton**, University of British Columbia.

This attractive and highly readable book contains a finely-nuanced and richly documented study of a set of discourse particles (DPs) in Modern English, based on data from the London-Lund Corpus of spoken English, with some comparison to the Lancaster-Oslo/Bergen Corpus of written English and the COLT Corpus of London teenager speech, where relevant. Covered in detail are the particles *now, oh/ah, just, sort of, actually,* and tags such as *and that sort of thing*, chosen because of their frequency in the corpus. The book has its genesis in studies of individual DPs that Karin Aijmer has published over the past fifteen or more years, though substantially revised and updated. It is a pleasure to have this work brought together in a unified text.

Aijmer begins by defining DPs as grammaticalized (or partially grammaticalized) elements in which pragmatic (textual and phatic) functions override 'literal' (lexical or referential) meaning. They may be oriented either backwards or forwards in the discourse. Formally, DPs are characterized by syntactic position (in the 'pre-front field', as insertion, or as tail), prosodic features (often separate tone units), textual distribution (in dialogic, interactive texts), and clustering tendencies (with other DPs).

Aijmer's overall approach is strictly corpus-based and 'bottom-up' (beginning with the linguistic description of individual particles). Eschewing any one technique of analysis, such as speech act theory or relevance theory, Aijmer takes a broadly functionalist perspective and utilizes a variety of discourse-analytic techniques. She attempts to identify the 'core meaning' of each particle, relating its different functions to this core or prototype in a polysemous way (what she calls a 'modified minimalist description' (p. 21)). Multiple functions can often be explained by reference to linguistic factors such as collocation,

prosody, and text type distribution. Such linguistic clues are also used to distinguish the DP use from the adverbial or interjectional use of each particle. While she admires Schiffrin's (1987) integrated approach, which explains the behavior of DPs on five different levels, she finds it sufficient to restrict her analysis to two macrolevels – textual and interpersonal. On the textual level, DPs may function on either the local or the global coherence level, what Aijmer calls a "qualifier" or a "frame", respectively. On the interactive level, DPs may be expressions of evidentiality, may function as hedges or as boosters, may relate to politeness, or may be used for floorholding.

Important to Aijmer's conception of DPs are their indexical quality and their grammatical status. The indexicality of DPs is their link "to attitudes, to participants and to text" (p. 39). Like other indexical elements, DPs require a fair amount of inferencing in order to be decoded. Indexicality plays a role in the ongoing process of grammaticalization (or pragmaticalization) of DP's. The multifunctionality of DPs follows from their indexical properties, their grammaticalization, and their emergence as fully formed DPs. (Grammaticalization is defined according to the work of Paul Hopper and Elizabeth Traugott.) Pragmatic functions are derived from propositional meaning via certain paths of grammaticalization and on the basis of pragmatic principles (such as inferencing).

Following an introductory chapter in which the theoretical background and analytic framework of the study are set out, Chapter 2 begins the analysis of individual DPs by focusing on the 'topic-changer' *now.* Aijmer argues that the core function of *now* in denoting a boundary is a direct outcome of its temporal meaning 'at the present moment'. On the textual level, *now* has numerous foregrounding, boundary-marking functions: to shift topic, to frame discourse units, to mark off turns, to delimit sub-topics, to denote steps in an argument or moves in a narrative, or to draw attention to elements in a list. But it may also serve in the background to elaborate a sub-topic or provide explanation or clarification. On the interpersonal level, *now* is a marker of subjective modality. It may introduce meta-comments (*now let me see*) or prefaces, or may be used to heighten the effect of reported or one's own speech. It may function alternatively as a speaker-oriented stance marker expressing evaluation (*now that's dreadful*) or introducing a disclaimer or opinion (*now I think*), or as a hearer-oriented stance marker of impatience, resistance, or intensity (*now come on, now wait, now look*).

The interjections *oh* and *ah* (Chapter 3) are the most multifunctional of the DPs discussed. *Oh* is often used in contexts in which the core meaning of 'surprise' is backgrounded: to arrive at a realization (*oh I see*), to express clarifica-

tion after correction, to denote emphasis or intensification, or to register objection or reaction (*oh but*). It has special uses after statements and in elicitational contexts. It may function as a 'topicalizer' or 'newsmark' to promote topic development (*oh are you?*), as a backchannel device to register reception and recognition, as a sign of assessment (*oh that's good*), or as a signal of endorsement (*oh yes, oh no*). When embedded in a turn, *oh* may also demarcate the transition to a clearer formulation or to an aside. An interesting use occurs when *oh* precedes direct quotation and marks the change to a different deictic center of talk. In comparison to *oh, ah* is more formal, does not occur in lexicalized combinations, does not serve as an intensifier, and always contains a component of pleasure. Both forms, but especially *oh*, have a variety of politeness functions in thanking, inviting, apologizing, and expressing appreciation.

The shortest discussion (Chapter 4) treats the interpersonal particle *just*, which Aijmer sees as having a double function as a weakening (downtoning) and strengthening (intensifying) particle. With expressions of extreme or excess, *just* may denote the speaker's emotional bond with the hearer and serve the purpose of positive politeness, while in collocation with markers of tentativeness such as *I think*, it can soften the force of a face-threatening act and serve the purpose of negative politeness. In persuasive discourse, *just* may serve a rhetorical purpose in emphasizing the illocutionary force of an utterance.

In Chapter 5, Aijmer argues that the 'adjuster' *sort of* has two central functions. As an evidential, it may adapt a lexical item to a new instance, mark an expression as a type of metaphor, indicate a numerical approximation, signal lexical imprecision or a lexical gap, and introduce a self-repair. As an affective (interpersonal) marker, it serves as a downtoner (or compromiser), it hedges strong opinions (hence positive politeness), it establishes common ground, especially in collocation with *you know*, and it reduces imposition (hence negative politeness). These functions relate to the 'core meaning' which is metalinguistic and procedural: "to signal that the hearer will be able [to] figure out the meaning of what is said even if it [is] only approximate" (p. 209). Aijmer notes that, unlike other DPs, evidential *sort of* may affect the truth value of an utterance.

Chapter 6 treats a variety of 'referent-final tags', such as *and so on, and things, and things like that, or something, or anything,* and *or so*, which constitute lexicalized phrases and must be treated non-compositionally. On the textual level, they serve as a signal to the hearer to interpret the preceding element in the discourse as an illustrative member of a more general set. On the interpersonal level, they may express tentativeness, intensification, or approximation. They frequently collocate with *you know/see* and other forms that help negotiate common ground. *And*-tags have a 'concretizing' function in expanding and

illustrating; by circumventing the need to give an exhaustive list, they avoid tedious description, speed up a narrative, or invoke a certain ambience. If they contain the universal quantifiers *all* or *everything*, they may serve an intensifying function. *Or*-tags express numerical approximation or tentativeness, and thus serve purposes of negative politeness.

The final discussion (Chapter 7) is of the DP *actually.* Because meaning and use are often unhelpful, Aijmer considers position (utterance- or clause-final, utterance-initial, and post-head) as the defining characteristic of the DP in contrast to the adverbial function of *actually.* The core meaning of the DP relates to the lexical meaning of *actually*: it expresses a discrepancy between reality and what appears to be the case. It has two major functions, contrastive ('but actually') and emphatic ('and actually'). In the former function, the speaker may distance himself from the factuality of an earlier utterance, express an opposition between different points of view, or attempt to change the hearer's perspective. In the latter function, the speaker may provide explanation or justification (*actually, I think/to tell you the honest truth*) or may suggest that information is unexpected. In final position, *actually* may be interpersonal and positively polite, serving to soften what has been said by foregrounding it as a subjective opinion.

The few criticisms that I have of this work do not reflect on its overall – and obvious – strengths. The discussion of grammaticalization remains rather underdeveloped, and alternates between a synchronic and a diachronic view of the process. Discussions of the grammaticalization of individual particles either make brief reference to the work of others or do little more than rehearse general principles of grammaticalization (e.g. subjectification, change in scope) without focusing on changes in the particle in question. Undoubtedly, I am revealing my own interests here in the diachronic development of discourse particles (see Brinton 1988). In the end, the discussion of grammaticalization seems rather tangential to the main line of synchronic analysis in this work. For me the concept of the "indexicality" of DPs , albeit Aijmer sees this as their "most important property" (p. 5), also remains somewhat nebulous, and its contribution to grammaticalization unclear. Finally, although Aijmer is critical of the overly abstract 'core' meanings provided by those taking the minimalist approach to semantics (e.g., the approach of Anna Wierzbicka), her postulated core meanings are often as equally broad, as in the case of *just* whose core meaning "functions as an instruction to the hearer to interpret the utterance as the expression of an attitude" (p. 158). And Aijmer is forced several times to admit that the core meaning of a particular particle is difficult or impossible to specify.

These minor points aside, Aijmer's book represents an important contribution to research in the area of discourse particles in English, and to research in

discourse particles generally. While studies of individual discourse markers abound, they are scattered throughout journals and collected works and are variable in their methodology, source of data, and reliability. Very few full-length studies of English discourse particles exist. The most recent, Blakemore (2002), takes a 'top-down' approach using relevance theory. Closest in approach to the current study is Lenk (1998), likewise based on data from the London-Lund Corpus (as well as the still unpublished Santa Barbara Corpus of Spoken American English). However, Lenk examines a rather different set of particles (*anyway/anyhow, however, still, incidentally, actually, what else*) and is primarily concerned with their global textual function as discourse-structuring devices. Some older book-length studies (Goldberg 1980; Schourup 1985; Erman 1987; Schiffrin 1987) examine rather different sets of particles and are based on more limited data.

References

Blakemore, Diane. 2002. *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge Studies in Linguistics 99. Cambridge: Cambridge University Press.

Brinton, Laurel. 1988. *Pragmatic markers in English: Grammaticalization and discourse functions.* Topics in English Linguistics 19. Berlin and New York: Mouton de Gruyter.

Erman, Britt. 1987. *Pragmatic expressions in English: A study of "you know", "you see", and "I mean" in face-to-face conversation.* Acta Universitatis Stockholmiensis, Stockholm Studies in English 69. Stockholm: Almqvist and Wiksell.

Goldberg, Julia. 1980. Discourse particles: An analysis of the role of *y'know*, I *mean*, *well*, and *actually* in conversation. Unpublished Ph.D. dissertation, Cambridge University.

Lenk, Uta. 1998. *Marking discourse coherence: Functions of discourse markers in spoken English*. Language in Performance 15. Tübingen: Gunter Narr Verlag.

Schiffrin, Deborah. 1987. *Discourse markers.* Studies in Interactional Linguistics 5. Cambridge: Cambridge University Press.

Schourup, Lawrence C. 1985. *Common discourse particles in English conversations: "like", "well", "y'know".* New York and London: Garland.

Karin Aijmer (ed.). *A wealth of English. Studies in honour of Göran Kjellmer.* Göteborg: Acta Universitatis Gothoburgensis**,** 2001. 319 pp. ISBN 91-7346-398-1. Reviewed by **Fanny Meunier**, Université Catholique de Louvain.

This edited volume consists in a collection of twenty-three essays dedicated to Professor Göran Kjellmer on the occasion of his seventieth birthday. The title of the book pays a real tribute to his successful academic life and his profound love of the English language. The papers included are offered both to Professor Kjellmer and the readers as a colourful bouquet.

Before embarking on the reading of the various contributions, I recommend a tour of Kjellmer's bibliography[1] which unmistakably reveals his sense of humour. 'Concerning thirst in battle and dog-riding', 'Why is Winnie the Pooh?', ' The cups that cheer but not inebriate', 'How to crash into a kangaroo', 'He is one of the few men in history who plays jazz on a violin' and 'Cowed by a cow or bullied by a bull?' are only a few of his scientific papers' headings. Undoubtedly inspired by Kjellmer's talent, some of the contributors to the volume also came up with intriguing titles such as: 'A funny thing happened to me on the way to Sidney' (Svartvik), 'Did rhoticity kill the Hillbilly Cat?' (Mobärg) or 'Pin, pin bring me luck, because I stop to pick you up' (Persson), thereby inviting the curious reader to further discovery.

A wealth of English contains six sections: grammar, semantics & word-formation, text & discourse, contrastive studies, ELT, and the music & magic in the English language. The grammar and semantics & word-formation headings total seven papers each, while the four additional sections include fewer papers, with the ELT component containing only one paper. The multiplicity of the topics covered, and hence the large scope of the book, may tend to give a rather patchy impression, which does not facilitate the reviewer's task. The present review does therefore not comment on all the articles and is based on a – necessarily subjective – selection. Despite the lack of focus on a central theme, the majority of papers have adopted a corpus linguistics perspective, thereby offering an interesting panorama of linguistic phenomena tackled with the help of corpora.

Tottie and Hoffmann's study illustrates the **grammaticalization** phenomenon through a careful analysis of 'participles that have passed into prepositions' (Fowler and Fowler 1936: 119). They focus on the *based on* collocation when used as a synonym for *because of*. Corpus analysis reveals that, while a substantial increase of the frequency of use of *based on* can be found both in British and American English corpora, the potential prepositional use has quadrupled in American English only. The authors also comment on other word combinations,

such as *due to*, which have followed the same pattern and are now considered as bona fide complex prepositions. In reference to the usually rapid acceptance of features of American English in other varieties of English, Tottie and Hoffman conclude by predicting the grammaticalization of *based on* as a complex preposition into British and other Englishes. The related petrification or **lexicalization** phenomenon is addressed in Seppänen's article. Starting from one of Kjellmer's papers on *as is*, he studies some sequences more literary in style: the *as/that was* combinations. With the help of numerous examples, he demonstrates how such sequences have undergone a lexicalization process leading towards idiomaticity. Implications for lexicography are also discussed, and the author shows how dictionaries have been somewhat slow in recognizing innovations in the language.

Other papers focus on **variation studies**. Olofsson's analysis of *one of whose* addresses comparisons between *American and British English* using a variety of corpora, namely the Brown, LOB, Frown, FLOB, BNC and Cobuild-Direct. While the *one of whose* structure seems to lack ample corpus evidence, Olofsson shows how English has "found a way to express the inexpressible" (p. 21). De Haan provides further insight into the syntactic make-up of *spoken and written English* and shows how the frequency of tags and tag sequences can help establish crude syntactic differences between speech and writing. He establishes the more 'clausal nature' of speech vs. a more 'nominal tendency' for written language. Mobärg compares *singing and speaking pronunciation* and focuses more specifically on the rhoticity[2] phenomenon. With the help of precise comparisons and rhoticity percentages, he demonstrates the links between linguistic behaviour (pronunciation) and choice of identity in Elvis Presley's career. Although his study cannot prove the conscious or unconscious use of changes in Presley's pronunciation, Mobärg demonstrates that the songs tested differed in a patterned fashion according to genre. Wichmann's study focuses exclusively on speech and shows how progress in *spoken corpus management* can be used to analyse spoken parentheticals. While the latter were usually considered as disfluency or performance phenomena not really worth analysing, Wichmann uncovers the actual pragmatic role of spoken parenthesis as real interaction management tools in discourse.

Some articles offer detailed distribution analyses of **syntactic elements**. Oostdijk provides a comprehensive syntactic description of the English adjective phrase (hereafter AJP), accounting for the most frequent syntactically simple phrases to the most syntactically complex constructions that occur less readily. The author draws a very detailed picture including a descriptive account of the various types of adjective pre- and post-modification. She also presents the distribution of AJP types and the distribution of AJPs over different func-

tions. She demonstrates that complex structures are rare in actual language use and that the more complex structures attested are relatively simple when compared to the vast array of theoretical potential for complexity. Kennedy, too, carries out a distribution analysis of two syntactic phenomena associated with the use of the passive in English. However, while Oostdijk's paper is limited to a descriptive account, Kennedy goes one step further in trying to seek whether the distribution can help account for the difficulties associated with the learning and actual use of the passive.

The usefulness of corpus linguistics is also revealed in the **contrastive** papers of the volume. Altenberg's article on the delexical English *make* and Swedish *göra* can be considered a methodological model for contrastive studies. Not only does it illustrate the various essential steps in contrastive analysis, but it also shows how contrastive studies can supplement interlanguage research by "revealing the degree of correspondence between languages (…), [giving us] a better chance to understand the problems facing the learners (…), [providing] a firmer empirical basis for interpreting their behaviour (…) [and helping] to form hypotheses about interlanguage which can be further checked against learner data" (p. 218). Johansson's article is in the same methodological vein and analyses the English verb *seem* and its correspondences in Norwegian. Beyond the interesting results directly linked to the topic, his study reveals that contrastive studies make it possible to map correspondences across languages but also help highlight the specific characteristics of each language. Aijmer's contrastive interlanguage analysis deals with *I think* as a marker of discourse style in argumentative student writing. She demonstrates that learners have problems mastering the genre conventions and tend to over-personalize academic essays. Her study also has pedagogical implications, and she shows the potential (both L1 and L2) teaching induced features of the over-use of *I think* by Swedish learners.

The topic of **language change and evolution** is also covered by a number of articles. Peters analyses the Latin legacy and its evolution in the language. She acknowledges the influence of Latin on the English language, both in lexical and grammatical terms. She also provides interesting data on, among other aspects, sociolinguistic variations in the maintenance of Latin elements (with women being less likely to anglicise Latin words) and regional differences between British and American English (with British English being more conservative). Stålhammar's contribution, 'Through the computer screen', shows how new technologies require new lexical terms. The author demonstrates that new terms have generally been borrowed from general purpose language or from more specialized fields, and that coinage constitutes the exception rather than the rule. He provides numerous examples of metaphorical computer terms

related to office environment (*document, file, desktop, library*) or to anthropomorphization (*memory, dumb, smart, mother board, daughter boards, hosts, master*, etc.) thereby demonstrating that there is "a life beyond electricity" (p. 121).

Although several authors have included pedagogical implications in their contributions, Svartvik's paper is the only one appearing in the **English Language Teaching** section. While the expressions 'English as a lingua franca' or 'English as an international language' are becoming increasingly fashionable, I dearly welcomed Svartvik's plea for at least some sort of native language model in ELT. His article goes along the same lines as Quirk's 1990 paper on 'Language Varieties and Standard Language' and comes in reaction to Modiano's article (2000) 'Rethinking ELT'. While the plea for a lingua franca, with less importance given to a native speaker model, is based on philosophical and humanistic statements which should be respected and valued, the results of such an approach might in the end turn up to be rather disappointing for the learners and, instead of empowering them, this approach might 'ghettorize' them, as Jones (1998) shows in her 'Not White, Just Right' essay.

Finally, some papers can easily be classified in an '**atypical**' section. Ohlander's contribution ('Onomastics, Grammar and Rock'n'roll') is probably the only linguistic article whose bibliography contains as many references to rock and music documents as to linguistics. However, the article does offer a solid linguistic and grammatical perspective on the names of rock bands (articles, sentences, verb phrases and noun phrases). The author demonstrates the variations in structures and the lack of grammatical systematicity which can be explained, in the case of rock bands, by a quest for ingenuity and originality. Persson's article on magic in the English language addresses the close relationship between magic and language through the presentation of charms and spells, divination, taboo expressions, naming and timing of utterances. Bergh's article also belongs to this atypical category, with a bibliography containing no less than nine (out of a total of eleven) references to Chemistry Journals. Bergh analyses the semantic categories, morphological patterns and word formation trends of the periodic table featuring elements' names, atomic numbers and symbols as recommended by the International Union of Pure and Applied Chemistry and shows that it contains features that are unusual in other word formation contexts. Some hot questions do however remain unanswered and I will end the review on this intriguing note (p. 152): "[is there] a chance that the *–ium* suffix will yield to the *–on* suffix when it comes to the definitive naming of the last element in the noble gas series, *ununoctium*?"

Notes

1. A select list of publications is included at the end of the volume.
2. People with a rhotic accent pronounce the 'r' wherever there is an <r> in the spelling of a word. Scottish English pronunciation is typically rhotic, while Black American pronunciation is considered as non-rhotic.

References

Fowler, Henry W. and Francis G. Fowler. 1936. *The King's English.* Oxford: Clarendon Press.

Jones, Rachel L. 1998. Not white, just right. In V.P. Clark, P.A. Eschholz and A.F. Rosa (eds.). *Language. Readings in language and culture.* Sixth edition, 489–491. New York: St. Martin's Press.

Modiano, Marko. 2000. Rethinking ELT. *English Today* 62: 28-37.

Quirk, Randolph. 1990. Language varieties and standard language. *English Today* 21: 3-10.

Bengt Altenberg and **Sylviane Granger** (eds.). *Lexis in contrast: Corpus-based approaches*. Studies in Corpus Linguistics 7. Amsterdam and Philadelphia: John Benjamins, 2002. 337 pp. ISBN 90-272-2277-0. Reviewed by **Magnus Levin**, Växjö University.

For many years, lexis was treated as a rather chaotic area of language where those unpredictable and idiosyncratic features that did not fit into syntax were put. Now, a major shift in priorities has occurred and lexis features high on the research agenda. In recent years lexical properties and their influence on syntactic phenomena have been receiving increasing attention. Instead of being considered as separate entities, lexis (which was earlier called 'vocabulary') and grammar are now seen as interdependent. The revival of contrastive linguistics (CL) is another change in the focus of linguistic theory that has been important to this collection of papers. Like the increasing interest in lexis, this trend has been greatly facilitated by the computer revolution, but CL has also benefited from increasing internationalization and the integration of Europe, and the current interest in real-life communication. The volume *Lexis in contrast: Corpus-based approaches* is a sign of these developments. The collection features thir-

teen papers divided into four sections and an introduction by the editors. The contributions include studies both of translation corpora and of L1 corpora. All the papers involve English as the object of study, while the contrasting languages mainly involve French, Swedish, Spanish, Chinese and Italian. Although practical applications of multilingual corpora form the main part of the volume, more theoretical issues are also discussed at some length.

Most papers in the book are thoroughly investigated pieces of work, while a few seem somewhat superficial, but that may be due to the necessary restrictions of space in a volume like this. Nevertheless, this book contains a wealth of ideas and approaches and is a valuable addition to both lexicology and CL.

Bengt Altenberg and Sylviane Granger's introduction provides a comprehensive overview of the renewed interest in lexis and CL, and the motivations for it. The editors present a number of applications of multilingual text corpora in contrastive lexical studies. They mention, for instance, that such corpora can provide insights into the languages that might have been overlooked in monolingual corpora, they provide material for the study of contextual influence and they are essential in investigations of multilingual lexicography and terminology. These areas are also in focus in many of the papers in this volume. Some more theoretical issues are also brought up in the introduction. It is argued that equivalence in translation is relative and a matter of judgement. This is an important point since one of the common features of the empirical studies in this book is the low translation equivalence across languages for pairs that at first glance would seem to be very close. In such cases a corpus can lend some intersubjectivity to the findings.

As for the future of lexical CL, the editors think that there are "exciting times ahead" (p. 39). The papers collected certainly seem to be precursors of such a future.

The first section of the volume is entitled Cross-Linguistic Equivalence and contains three papers. As indicated in the heading, the papers explore the complex problem of finding equivalents across languages. In the first of these Raphael Salkie investigates the issue of translation equivalence with the aid of two examples – the translation of the German word *kaum* into English, and the English word *contain* into French. Although rather limited in scope, the study clearly illustrates how parallel corpora can solve translation problems.

In the next paper, Elena Tognini Bonelli compares the English phrases *in the case of, in case of* and *in case* in an English newspaper corpus with their *prima facie* Italian equivalents *nel caso di, in caso di* and *se per caso*. The phrases in question produce quite similar patterns. For instance *in case of* and *in caso di* both have a strongly negative semantic preference, co-occurring with words like

death, *war*, *massive calamity* and *constipation*. The author proposes the term 'functionally complete units of meaning' to characterize the co-selectional patterns of words, because "words do not live in isolation but in strict semantic and functional relationship with other words" (p. 91). Words have different collocational (lexical) and colligational (grammatical) patterns, and through co-selection multi-word units are formed. This study demonstrates clearly that the fine-grained evidence obtained from the repeated patterns in corpora can provide essential information for translators. Although the fit between the languages was very good this time, it cannot be assumed in other cases, Tognini Bonelli suggests.

The section on cross-linguistic equivalence ends with Bengt Altenberg's paper, which is a thorough comparison of the Swedish equivalents of English causative *make* + Object + Infinitive. The conclusion is that the two languages have a similar range of options for translations: analytical constructions with *make* in English and *få* in Swedish, other causative verbs (*get*, *cause*, *komma*, *tvinga*), synthetic causative verbs and miscellaneous constructions. Altenberg connects these findings with results from learner corpora and argues that, since analytical constructions are cross-linguistically similar and unmarked in both languages, they are likely to be overused by learners. Altenberg's study shows convincingly how findings from parallel corpora can enhance our understanding of cross-linguistic phenomena.

The next section is devoted to Contrastive Lexical Semantics and contains three case studies. In his contribution, Åke Viberg continues his quest to map lexical differences between English and Swedish. In this article he compares the Swedish high frequency verb *få* with English *get*. These high frequency verbs have developed a number of new meaning extensions in different languages (some of which have become grammaticalized), and this can account for the different patterns in English and Swedish.

In a highly stimulating contribution, Lan Chun compares English metaphors with *up/down* with Chinese metaphors with the corresponding *shang/xia* within the framework of cognitive semantics. Four target domains are focused on, namely QUANTITY, SOCIAL HIERARCHY, TIME and STATES, with examples such as *The price of milk should be* down *next week*, *the* upper *strata of society, from 1918* up *to 1945* and *That was a low*-down *thing to do*. The comparison reveals remarkable similarities between the languages in the L1 corpora. Both pairs of words are frequently used in these domains, and what is oriented *up* is also oriented *shang* (with only one exception), and what is oriented *down* is consistently oriented *xia*. This indicates that "there may indeed exist a univer-

sal metaphorical system" (p. 173). This paper clearly demonstrates how CL can be used to test linguistic theories.

To conclude this section of the collection, Michel Paillard's paper produces some tentative findings which suggest that metonymy is more common in French, while hypallage (the reversal of the normal functions of elements in order to create a specific effect, as in *Melissa shook her doubtful curls*) is more readily used in English.

The section entitled Corpus-based Bilingual Lexicography offers a wide selection of methodologies. The papers are largely concerned with problems occurring in translations. A major focus is on how to use corpora to facilitate translation. In the first contribution, Wolfgang Teubert argues that bilingual databases will supplant traditional printed dictionaries because these databases can cope better with translation units in context. He exemplifies his point by comparing the words *work*, *travail* and *Arbeit* in different version of such diverse sources as Plato's *Republic* and EU documents. It is striking how rarely the standard translation equivalent occurs in actual translations. For instance, *Arbeit* is only rendered *travail* in three out of twenty instances, while the plural *travaux* is used eight times in the corpus of EU documents. Teubert suggests that recurrence should be used as a parameter for distinguishing good translation practice from bad, and that "actual translation practice offers a wider choice of options and a larger design space for translation than the traditional bilingual dictionary" (p. 212).

Victòria Alsina and Janet DeCesaris take an entirely different approach in their paper. Instead of looking at translations, which contain transfer from the source language, they compare the information provided in monolingual dictionaries and in English/Spanish and English/Catalan dictionaries for the polysemous adjectives *cold*, *high* and *odd* with native-speaker usage in the British National Corpus.

Sylviane Cardey and Peter Greenfield's concern is the construction of computerized set expression dictionaries. Set expressions, like *when pigs have wings* and *to breathe down someone's neck*, cause problems for natural language processing, and one of the aims of the authors is to build a system that can recognize these expressions. Not surprisingly, they conclude that "although machines are useful in advancing and verifying the work of the linguist, there remains much core work which only the linguist is competent to carry out (conception, understanding and organisation), and such work is also essentially manual in nature" (p. 246).

This section is concluded by Christine Chodkiewicz, Didier Bourigault and John Humbley, who work on the production of a glossary for specific purposes.

They explore English and French equivalents in legal texts and find that some terms, such as English *friendly settlement* and French *règlement amiable*, are equivalents in the corpus, whereas the term *proceedings* has no less than twelve equivalents in the French texts. Automatic processing of terms offers a large number of advantages: the total number of occurrences can be accessed, the context enables the translator to disambiguate meanings and to facilitate the harmonization of terms.

The last three papers in the volume are gathered under the heading Translation and Parallel Concordancing. To begin with, Olivier Kraif reflects on translation alignment and lexical correspondences. The contribution by François Maniez, somewhat oddly placed under the heading "Parallel Concordancing", looks at the problem of resolving potentially ambiguous items, like for instance the word *rate* and the phrase *based on*, in translations. He compares a corpus of medical texts with a newspaper corpus and finds that there are clear differences between the distributions of the various alternatives in different corpora. For example, *based on* is more frequently found as a complex preposition in medical writing than in news text, and such information would be a considerable help to translators.

In the final paper, Patrick Corness demonstrates how the software Multiconcord can be used to investigate translation variants. He exemplifies this by looking at some length at the phrasal verb *pick up* and its translations into Czech and Lithuanian.

The articles in *Lexis in contrast* provide ample evidence for the empirical and practical benefits of contrastive lexical studies. The volume highlights the potential of modern translation corpora and of comparisons of L1 corpora. Both theoretical linguists and translators will profit from reading this book. *Lexis in contrast* covers a wide range of topics and applications of a neglected area and is warmly recommended for anyone working in any of the fields covered in the volume.

Sylviane Granger, Estelle Dagneaux, and **Fanny Meunier** (eds.). *International Corpus of Learner English*. Version 1.1. Université catholique de Louvain: Centre for English Corpus Linguistics, 2002. Reviewed by **Erik Smitterberg**, Stockholm University.

The corpus-based study of learner English, from scientific and pedagogical perspectives, is an area of research that is attracting more and more scholarly interest, as evidenced by publications such as Granger, Hung, and Petch-Tyson (2002). By combining insights from Second Language Acquisition theory and English Language Teaching practice with a corpus linguistic methodology, researchers are able to describe interlanguage features and suggest implications for language teaching with greater confidence than has hitherto been possible.

Any area of corpus linguistics is necessarily dependent on available, reliable, and – preferably – comparable corpora that can serve as sources of data. Although a look at the corpora used by the scholars who contributed to Granger, Hung, and Petch-Tyson (2002) reveals that several learner corpora are currently being compiled in different parts of the world, few of these corpora appear to be publicly available as yet. In addition, some of the corpora chiefly contain specific types of learner English, such as ESP English, or English produced in an examination situation that may be more or less specific to the nation where the examination takes place. While all of these corpora appear to be reliable and valuable sources of data, there is still a need for learner corpora that are publicly available and comparable across several native languages. The publication of the International Corpus of Learner English (ICLE) is an important step forward in this regard.

The ICLE is stored on a CD-ROM, which contains a database of the corpus texts and the learner profiles. A license agreement and a handbook are also included. All page references in the present review are to the handbook, which has four sections: a description of the corpus, a user manual, a survey of the status of English in the countries of origin of the learners whose essays were sampled,[1] and a list of ICLE-based publications.

The ICLE contains about 2.5 million words of learner English; it consists of academic writing – mainly argumentative – produced by "university undergraduates in English (usually in their third or fourth year)" (p. 14). The corpus is divided into "eleven national subcorpora" (p. 27) of between 200,000 and 278,000 words each. Eleven native language backgrounds are represented, but there is no exact match between the backgrounds and the subcorpora: learners with a Swedish language background, for instance, are represented in both the

Finnish and Swedish subcorpora.[2] The term "national" is somewhat misleading regarding some subcorpora: for instance, the French subcorpus consists of essays written in Belgium (by native speakers of French), and the German subcorpus of essays written in Austria, Germany and Switzerland. This potential source of confusion is not serious, given the powerful selection tool that comes with the corpus texts (see below), but may still puzzle users, who will be faced with a list of countries to choose from that does not match the list of national subcorpora.

The learner profiles are stored in a database, and contain a great deal of information on each essay and essay-writer (see below). The profiles are linked to the texts by essay codes, which contain, among other things, a national code and an institution code (e.g. FIHE for Finnish, Helsinki University). The texts are in ASCII format, untagged, and contain no markup except for essay codes linking each text to its profile, and codes for deleted quotes, deleted bibliographical references, and illegible words. The text format is designed to work well with software tools for linguistic analysis such as WordSmith Tools.

After the corpus has been installed and the program started, the Query window, which "consists of two superimposed pages which can be moved to the foreground by clicking on the relevant header tab" (p. 54), appears on the screen. These two pages represent a major strength of the ICLE package: they contain about 20 variables (alphanumerical, numerical, alphabetical, or selection lists) according to which corpus users can select texts. The coverage is impressive: it is possible to select essays according to features of the essay (e.g. type, length, and production circumstances) as well as features of the learner (e.g. sex, country, native language, language at home, age, and years of English at school). The advantage of this coding scheme is that corpus users can design their own tailor-made subcorpora, which clearly helps to increase the validity and reliability of, for instance, comparisons across native languages. For example, Aijmer (2002: 73f.) emphasizes the importance of controlling for topic in research on modality in learner writing; the ICLE package enables users to select essays according to both type ("argumentative", "literary", or "other") and (words in) title. The only drawback in this respect is that some of the subcorpora that are selected by combining several variables will be quite small.[3] The handbook describes the selection process well, and help files are also available via the menu system of the program itself. However, some further information on how, exactly, each variable has been classified might be a useful addition to the handbook. Moreover, one variable I missed was whether each text had originally been submitted electronically or on paper (both methods were used), as this may affect the number and type of spelling errors. On the one hand, the use of spell

checkers may reduce the number of erroneous spellings; on the other hand, if spell checkers are not used, the keyboard also makes misspellings possible that would not be likely to occur in a handwritten essay (e.g. *bsd* for *bad* owing to the adjacency of the *s* and *a* keys, or *langauge* for *language* owing to fingers hitting keys in the wrong order).[4] However, the overall impression of the Query window is that of a very powerful tool indeed.

After carrying out the selection process in the Query window, the user is ready to click the "Search" button. This takes him/her to the Response window, where the search results are displayed in a grid, with the texts selected as rows and the variables as columns; there is also information on how many texts were selected. Among other things, the user can sort the essays according to their values on the variables (though only for one variable at a time), view, save, and print each text selected, and generate search reports that list the variables used and provide detailed profiles on each essay and essay-writer. However, the most important function may be the "Merge texts" option. This makes it possible for the user to conflate all texts selected into one single subcorpus, which "can then be printed or saved in an ASCII file for further processing or analysis" (p. 67). There are several reasons why this is a very useful feature. First, researchers can devote time to creating subcorpora that are comparable across several variables; they can then save these subcorpora as separate files and carry out several linguistic analyses on them without having to go through the selection process again. Secondly, the subcorpora selected can be processed further: for instance, part-of-speech tagging or error tagging could be supplied.[5] This feature is another major strength of the ICLE package.

The merged file can be saved on the researcher's hard disk and then analysed using text retrieval software tools. This process is very simple and straightforward. However, when I used WordSmith Tools to run a search for expressions of the future in Swedish essays, a potential problem appeared: a few expressions seemed to have the same context in the concordance.[6] Further investigation revealed that the two essays with the codes SWUG2028 and SWUG2040 were virtually identical; there were only a few differences regarding, for instance, word order (e.g. *rich, well-off people* vs. *well-off, rich people*), spelling (e.g. *mobil phones* vs. *mobilphones*), and punctuation. Several mistakes were also the same in the two essays (e.g. *looses out* for *loses out*), which suggests a common origin. This inclusion of virtually the same essay twice in the corpus need not be due to a mistake on the part of the compilers; instead, it may be the result of plagiarism, which is becoming a widespread problem in EFL composition courses.[7] Nonetheless, the discovery of two texts that are virtually identical in the corpus prompted me to look for further examples. Owing to time

limitations, I only carried out a few investigations in this regard, creating subcorpora and running searches to see whether the resulting concordance would reveal identical passages. The problem does not appear to be widespread, but an analysis of the occurrence of *might* in texts by German students revealed at least one other case of two essays that appeared identical.[8] On the one hand, these problems affect less than one per cent of all texts selected in the respective searches, and are thus unlikely to have any significant impact on quantitative results. On the other hand, there may be further identical texts that I have not discovered, as the search word(s) had to appear in the relevant essays for the inclusion of identical texts to be detected.

In sum, the publication of the ICLE is a milestone in the corpus-based study of learner English. The fact that researchers can easily create subcorpora of their own and the power of the software tool that allows them to do so are significant advantages. The long list of international collaborators makes it clear that a truly impressive coordinating effort must have been required to make all subcorpora comparable. It is to the editors' credit that they point out some limitations as regards the current version of the product, such as the lack of linguistic annotation and the fact that about 200,000 words per national subcorpus "precludes any investigation other than that of high frequency linguistic phenomena" (p. 38). The inclusion of a bibliography of ICLE-related publications, brief descriptions of learner corpus research methodology, and brief articles on the status of English in the countries of origin of the learners further adds to the usefulness of the publication. It is hoped that future versions of the ICLE will include tagged texts and further subcorpora (both of which the editors aim to do), as well as more details concerning the coding scheme for the ICLE database. Revisions of the database to ensure that identical essays do not occur in the material would also be welcome. Subsequent versions of the ICLE could thereby improve on the highly promising impression of version 1.1.

Notes

1. Austria and Switzerland, which account for a mere 70 and 60 essays respectively, are not included in the survey.
2. The national subcorpora – and native language backgrounds – present in version 1.1 of the ICLE are Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. Subsequent versions aim to include texts by Brazilian, Chinese, Japanese, Norwegian, Portuguese and South African learners also.

3. For instance, a search for argumentative essays written by male Spanish Spanish-speaking students who did not produce their essays in an examination situation yielded 15 essays of between 306 and 1,101 words in length.
4. One of the variables makes it possible for researchers to select only essays that were (or were not) produced with the use of reference tools, but as a reference tool may be both a dictionary for a handwritten essay and the spell checker of a word processor, this variable probably cannot be equated with that of whether the essays were submitted electronically or as handwritten documents.
5. However, the license agreement supplied in the handbook states that "[l]icensee shall not modify, decompile, disassemble, decrypt, extract or otherwise reverse the Product" (p. 49), and that users who wish to make other use of it are requested to contact the Licensor. In this respect, it is unclear to me what the legal status is as regards ASCII files that have been merged and saved separately: for instance, are licensees allowed to tag these merged files?
6. I am grateful to Petra Balog for originally drawing my attention to this issue.
7. However, most of the variables have the same values for both essays: for instance, they were written under exam conditions on the same day. This may suggest that the same essay was included twice, with different codes.
8. The two essays have the filenames GEAU3002 and GEAU3024 in the Response window. However, the essays linked to these codes appear to be identical, and both essays have the code <ICLE-GE-AUG-00024.3> in the Text window where the actual text file is presented, which may suggest an error in the coding scheme.

References

Aijmer, Karin. 2002. Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, and S. Petch-Tyson (eds.), 55–76.

Granger, Sylviane, Joseph Hung, and Stephanie Petch-Tyson (eds.). 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam and Philadelphia: John Benjamins.

Sylviane Granger, Joseph Hung, and **Stephanie Petch-Tyson** (eds.). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam and Philadelphia: John Benjamins, 2002. viii + 245 pp. ISBN 90 272 1702 5. Reviewed by **Erik Smitterberg,** Stockholm University.

In recent years, corpus linguistic methods have gained an increasingly central place in English language teaching (ELT). For instance, students use a range of materials that draw on corpus linguistics, from modern learners' dictionaries to concordances produced for the purpose of data-driven learning. The overwhelming majority of all English-language corpora used in this way consist of native-speaker English, but recent findings have shown that the compilation and analysis of learner corpora are also relevant from a pedagogical perspective (see e.g. Granger 1998), regarding fields such as materials design. There are also indications that students may benefit from analysing learner English as a complement to looking at native-speaker output. However, there is a need for studies of learner English that help to open up the field in this respect. *Computer learner corpora, second language acquisition and foreign language teaching* helps to satisfy this need. It is a collection of contributions intended both to help researchers assess the relevance of research on computer learner corpora for second language acquisition (SLA) theory, as well as ELT practice, and to "give practical insight to researchers who may be considering compiling a corpus of learner data or embarking on learner corpus research" (p. vii); this is a broad scope for a single volume, something which I shall return to towards the end of this review.

The book is divided into three sections. The first, "The role of computer learner corpora in SLA research and FLT", contains only one paper, by Sylviane Granger. She focuses on the contribution learner corpora can make to SLA and Foreign Language Teaching (FLT) research. As the title of her contribution indicates, Granger offers a brief but lucid and informative "bird's-eye view of learner corpus research". She outlines the field of corpus linguistics and the role of learner data in FLT and SLA research, and – importantly – offers a definition of learner corpora as well as commenting on aspects of that definition. Approaches such as Contrastive Interlanguage Analysis (CIA) and Error Analysis (EA) are discussed, as are the possibilities of software-aided analysis of computerized learner English. Granger is careful to point out potential pitfalls in this area, for example the fact that the accuracy rate of automatic taggers may decrease when they are applied to non-native English. Granger also discusses

practical applications of learner corpora, e.g. materials design, and future challenges, such as the need for corpus linguists and, among others, SLA specialists to co-operate (see also Hasselgård 1999: 152). Granger's contribution provides the non-specialist reader with a good deal of background information that is necessary in order to benefit fully from several of the subsequent, more specialized contributions. However, the wealth of abbreviations used is a drawback in this respect; even though they are usually explained, their frequency may discourage non-specialists.

The second part of the book, "Corpus-based approaches to interlanguage", comprises three contributions. In his analysis of Swedish students' overuse of causative *make* (e.g. *make someone happy*), Bengt Altenberg stresses that "reliable interpretations of interlanguage features require thorough knowledge of the three 'languages' involved: the learner's interlanguage, his/her mother tongue and the target language" (p. 38). Altenberg argues that the overuse is due to transfer from the students' first language (L1) rather than to overgeneralization of the main English target pattern, as French students display an underuse of causative *make*. Focusing on patterns where the complement is an adjective phrase, Altenberg uses the English–Swedish Parallel Corpus to compare the two most closely corresponding constructions in English and Swedish (i.e. English causative *make* and Swedish causative *göra*, as in *göra någon lycklig* 'make someone happy'), and to relate their use to that of other options.[1] By considering Swedish and English source texts as well as analysing translations bidirectionally, he demonstrates that causative *göra* appears to be more central in Swedish than causative *make* is in English. The results thus suggest that Swedish learners' overuse of the dominant target pattern (causative *make*) is due to transfer supported by cross-linguistic similarity where the similar pattern in the L1 (causative *göra*) is even more dominant.

Karin Aijmer also looks at advanced Swedish learners in her study of the expression of modality. Aijmer uses the Swedish component of the International Corpus of Learner English (ICLE) as her main primary material, and compares the Swedish texts with similar native English material as well as with the French and German components of the ICLE. In addition to modal auxiliaries, Aijmer examines modal adverbials (e.g. *perhaps*) and modal combinations (e.g. *would probably*). Aijmer's results reveal "a generalised overuse of all the formal categories of modality examined" (p. 72); she points out, however, that not all categories of modality were included in the study.[2] Aijmer sees several possible reasons for this overuse, including influence from spoken English, transfer from Swedish, and the topic of the essays. She also makes several suggestions for teaching, such as studying modal auxiliaries from a discourse perspective.

The third and last contribution to this section is Alex Housen's study of Dutch- and French-speaking learners' acquisition of parts of the English verbal system: the base form, third person singular present *-s*, the *-ing* form, and regular as well as irregular preterite/past participle forms. Housen looks at spoken language produced by learners between *c*. 9 and 17 years of age, who were divided into four proficiency groups based on lexical and grammatical criteria (a reference corpus of native English is also used).[3] One of the grammatical criteria "measures a speaker's morphological accuracy against the target norm" (p. 89). The inclusion of this measure may be problematic, as the proficiency groups are later used to characterize the learners with respect to a type of morphological accuracy, viz. their use of English verbal morphology; to the extent that the same forms were used in the criterion and in Housen's study, there is a risk of circular reasoning here. Considerable variation is found between the forms investigated as regards overuse, underuse, etc., and Housen shows that learners frequently acquire a form without yet being able to use it correctly. Housen also analyses parts of his data in order to test the Aspect Hypothesis, which predicts that learners will initially use a verb form predominantly with the verb type with which the function of the form is chiefly associated (e.g. *-ing* with activity verbs). Again, the results point to differences between the forms investigated. Housen offers several speculations, which would clearly be worth pursuing further, on why the forms investigated do not seem to be acquired in the same way: they include differences between temporal/aspectual and grammatical markers, different learning processes for regular and irregular morphology, and L1 transfer.

The third section of the volume is devoted to "[c]orpus-based approaches to foreign language pedagogy", and contains five contributions. The first and most general of these is by Fanny Meunier; it centres on the relevance of using corpora in EFL teaching with a focus on form. While she argues that findings based on native and learner corpora have not yet brought about "major changes in EFL curriculum design" (p. 124), she shows that reference tools such as dictionaries and grammars have benefited considerably from corpus-based research. In terms of teaching, the inclusion of authentic examples in textbooks and the use of data-driven learning with concordances are important developments; in this context, the use of learner English as a complement to native-speaker data is controversial, but appears to have several advantages. Meunier also lists ongoing and possible changes in grammar teaching from a short-term (e.g. data-driven learning), medium-term (e.g. using corpus linguistic methods and tools), and long-term perspective (e.g. a discourse-based rather than sentence-based view of grammar).

Angela Hasselgren addresses the issue of assessing learners' fluency. Fluency is difficult to define and describe; on the basis of previous research as well as her own investigation, Hasselgren demonstrates that so-called smallwords[4] may be an important indicator of learners' fluency. Like Housen, Hasselgren works with spoken material, consisting of the speech of 14- and 15-year-old Norwegian learners taking a spoken interaction test, as well as that of a British control group. Her investigation shows that the Norwegian students who were judged as relatively fluent in the test situation use fewer disruptive pauses, longer utterances, and more smallwords (in terms of both types and tokens) than the less fluent students; however, their output rarely approaches that of native speakers. Drawing on relevance theory, Hasselgren also discusses how smallwords contribute to fluency by, for instance, helping to indicate the state of success of communication; this discussion further strengthens the division between more and less fluent learners.

The contribution by Ulla Connor, Kristen Precht, and Thomas Upton illustrates a textlinguistic approach to learner English. They analyse the genre of (simulated) letters of job application written by non-native and native speakers of English: undergraduates from Belgium, Finland, and the U.S. Their analysis of the letters focuses on seven genre moves, described as "semantic/functional units of texts which can be identified first because of their communicative purposes, and second because of linguistic boundaries typical of the moves" (p. 180); for example, offering to provide more information is one such move. Overall, the results suggest "a cross-cultural consensus on the use of the majority of moves" (p. 185), but a few significant differences emerge: for instance, when arguing for the application, Belgian students tend to emphasize the benefit to the applicant more than Finnish and U.S. students, who mention benefits to the hiring company more often. The authors contend that genre-specific learner corpora will be useful for teachers, in that they make it easier to assess student needs, and that textlinguistic analyses of learner data are valuable; they also suggest that analyses such as theirs can help learners by clarifying genre characteristics, in terms of what moves are expected. Their study is interesting in that it opens up a textlinguistic perspective on learner English. However, I miss tables with raw frequencies that would enable readers to study the results in more detail, especially since the total number of letters (99) is fairly low, considering that the number of rhetorical moves, rather than the number of instances of, say, a grammatical feature, was analysed.

Quentin Grant Allan's contribution concerns the TLSC (TELEC Secondary Learner Corpus, where TELEC stands for Teachers of English Language Education Centre, Hong Kong). In 2002, the corpus, which is still under development,

contained 2.2 million words of student writing. Each text is coded for a number of parameters, which makes it possible to extract more homogeneous subcorpora (e.g. argumentative writing only). The main function of the TLSC is to form the basis for "systematic linguistic analysis of areas of English in which Hong Kong secondary students experience difficulty" (p. 200). The results are used in a hypertext database on grammar and usage aimed at teachers of English in Hong Kong: for instance, corpus extracts may illustrate incorrect or unidiomatic student output pertaining to an area of grammar, together with an explanation and correct versions. There are plans to improve and expand the corpus by, for instance, providing part-of-speech tagging, adding a spoken component, and creating a concordancer that would allow teachers to explore the corpus themselves.

Barbara Seidlhofer, finally, reports on an approach she dubs "learning-driven data", in which advanced learners analyse a corpus which they have themselves produced collectively during a course. The learners thus work with their own output, a practice Seidlhofer links to the Pushed Output Hypothesis, which states that "pushed output, i.e. sustained output that stretches the limits of learners' current linguistic capacity, can further their development significantly" (p. 218). The students' short written responses to the same article are conflated into corpora controlled for topic (the individual responses now being anonymous). The students then construct questions about the corpus texts, and discuss and try to answer many of these questions with the aid of corpus data. Seidlhofer argues that the students' motivation increased significantly as a result of their working with non-threatening texts that were already familiar to them.

Computer learner corpora, second language acquisition and foreign language teaching is a valuable and important publication. It contains several studies of great interest to corpus linguists in general, but also demonstrates the relevance of examining learner corpora both outside and in the classroom, regarding, for instance, curriculum development, materials design, and data-driven learning. Probably as a result of the broad scope of the volume, the contributions differ somewhat concerning matters such as how much detail they provide, and how much background knowledge they require, as regards, for example, terminology, corpus linguistic methods, and linguistics. However, Granger (p. 28) explicitly emphasizes the need for interdisciplinarity in research on learner corpora, and in order to bring several disciplines together some differences are probably unavoidable in this respect. The overviews that introduce the contributions help to familiarize the reader with the content of each contribution; the inclusion of a Name Index and a Subject Index is also an advantage. However, a list of abbreviations used in the volume as a whole, and perhaps a

list of explanations for specialist terms used, would have made the volume even more accessible to readers from different disciplines. The division of the contributions into sections appears logical for the most part, although the third section gives a more heterogeneous impression than the other two sections. Also, given that Aijmer's and Hasselgren's contributions both analyse learner English, compare it with native English, and suggest pedagogical implications, they might have been placed in the same section. The layout is inviting and the text usually runs smoothly, with only occasional infelicities regarding punctuation and spelling. This volume will be a definite asset to readers with an interest in learner corpus research, SLA theory and/or ELT practice.

Notes

1. The most frequent alternative is the use of a synthetic verb instead of causative *make*, e.g. *make something easier* ≈ *facilitate something*.
2. Aijmer also finds occasional underuse by Swedish students: for instance, they did not use root *may* in the texts examined.
3. While most learners were only interviewed once, a few were interviewed five times, at five-month intervals (p. 83). Consequently, a small number of students contributed considerably more material than the others.
4. "Smallwords" are defined as "small words and phrases, occurring with high frequency in the spoken language, that help to keep our speech flowing, yet do not contribute essentially to the message itself" (p. 150). A total of 19 smallwords (or smallword groups) were included in the study.

References

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

Hasselgård, Hilde. 1999. Review of: Granger (1998). *ICAME Journal* 23, 148–152.

Bernhard Kettemann and **Georg Marko** (eds.). *Teaching and learning by doing corpus analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July, 2000*. Language and Computers: Studies in Practical Linguistics 42. Amsterdam and New York: Rodopi, 2002. 390 pp. ISBN 90-420-1450-4. Reviewed by **Anne Curzan**, University of Michigan.

This volume of papers from TALC-2000 takes as its premise the value of corpora and data-driven learning to effective, student-centered teaching and learning in language classes, be they focused on second- or foreign-language learning, on linguistics, on translation studies, or on teaching for specific or for academic purposes. John Kirk, in his contribution to the volume, notes the larger shift from "teaching" to "learning" in pedagogical theory, and this volume exemplifies this pedagogical approach in practice, using corpora as the means for autonomous student-learning experiences. In the process, the volume both reaffirms the value of incorporating corpora into learning languages – or about languages – and highlights exciting innovations in available or developing corpus-based resources and pedagogical strategies. Many of the papers will be of interest to language and linguistics instructors designing student-centered, corpus-based linguistic investigation, as the authors share their best practices, cautions, successes, and failures in working with corpora. Other papers discuss issues at the heart of corpus design and will appeal to a perhaps even broader audience in corpus linguistics. The papers are both retrospective and forward-looking, as this community of scholars shares their experiences in order to further the development and exploitation of language corpora in teaching, learning, and research.

The volume's twenty-three generally very concise papers are divided into six sections: it begins with a broader section on "General Aspects of Corpus Linguistics" (four papers) and a short section on "Corpus-based Teaching Material" (two papers); at the heart of the volume are eight articles in "Data-driven learning"; the last three sections focus on more specialized corpora and language learning – "Learner Corpora" (three papers), "Corpus Analysis of ESP for Teaching Purposes" (three papers), and "Corpus Analysis and the Teaching of Translation" (three papers). The editors of the volume rightly note that these section headings capture only one way to categorize the papers and that contributions extend across category boundaries. In the brief introduction to the volume, Tony McEnery usefully clusters the articles within the framework of the developing focus of the TALC conferences, particularly on multilingual corpora

and on the increasing variety of corpus-based approaches and applications. In this review, I create a different set of connections, drawing out themes that run through the volume. In a relatively short review of a volume containing this many papers, I want to acknowledge upfront that I cannot do justice to the detailed content of many of the articles. In highlighting this set of issues and questions that run through the volume, I hope to demonstrate the richness of the contents from multiple perspectives.

Several papers describe new projects, often still in progress and/or representing early steps toward more ambitious designs. Federico Zanettin, in "CEXI: Designing an English-Italian Translational Corpus," introduces this bilingual, parallel, bidirectional, translation-driven corpus of English and Italian. Zanettin carefully leads readers through the selection process that resulted in the CEXI corpus of all books that have been translated between 1976-2000, have been published in Italy, the United Kingdom, or the United States, and are directed at adult readers – a process that raises questions of desired representativeness in these kinds of specialized corpora (an issue revisited in several other papers in the volume). Mike Scott provides a summary of the Guardian Keyword Database, to accompany the CD-ROM included with the volume, in "Picturing the Key Words of a very Large Corpus and their Lexical Upshots or Getting at the Guardian's View of the World." Keywords, their associates, and the calculation of "clumping" appear to be a feasible way to process very large databases and view structured hierarchies of "aboutness" or occasionally of stylistic features – although, as Scott notes, this work and its implications are currently exploratory. In "The Influence of External Factors on Learner Performance," Ylva Berglund and Oliver Mason describe their attempt to perform automatic stylistic analyses of texts using low-level features (e.g., sentence length, type/token ratios, average word length) as a way to identify how language-learner data differ from the production of native speakers. Their pilot study compares data from the Uppsala Student English corpus (USE) and from a subset of Frown; it would be very interesting to match these results with a comparison of data from USE to data from a corpus of native-speaker student essays, which may be significantly less polished than the material in Frown.

Agniezka Leko-Szymaska, in "How to Trace the Growth in Learners' Active Vocabulary?: A Corpus-based Study," exploits the PELCRA corpus of learner English, in comparison with the results of Vocabulary Level Tests which demonstrate receptive knowledge, to test the validity of two measures of lexical richness. Her useful conclusion that the Condensed Lexical Frequency Profile is the most meaningful measure of lexical richness is equally valuable for the methodology employed to test the measures and reach this conclusion.

Several papers raise fundamental questions about corpus design, including large representative corpora and smaller, more specialized ones. Lou Burnard's paper, "The BNC: Where did we Go Wrong?," offers a useful, concise retrospective on the development of the BNC, from acquiring permissions to sampling techniques, from annotation and encoding to distribution. Burnard's candor makes this paper an important read for prospective corpus designers. Near the end of the paper, Burnard discusses the repositioning of the BNC as a repository of language diversity (versus a "representative" corpus) and the insufficiency of the taxonomy of text types to exploit the BNC effectively in this regard. David Lee, four sections later, provides an almost direct response in "Genres, Registers, Text Types, Domains and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle."

Lee's contribution is in at least two ways anomalous in the collection: at forty-five pages, it is two to three times longer than any other paper; and the first half of the paper is an extended theoretical treatment of an ongoing terminological and conceptual issue in the field that is only indirectly related to the use of corpora in classrooms. That said, it is a very smart, interesting treatment of the distinction among the terms *register*, *genre*, and *text type* that provides a wide-ranging survey of the published material on the topic and stakes a well-justified position on how best to categorize texts in corpora, drawing on insights from prototype theory. In terms of corpus design, Lee argues that we are interested in genres, and he provides one of the clearest, most persuasive distinctions of *register* and *genre* that I have seen published; as he summarizes: "I contend that it is useful to see the two terms *genre* and *register* as really two different angles or points of view, with *register* being used when we are talking about lexicogrammatical and discourse-semantic patterns associated with situations (i.e. linguistic patterns), and *genre* being used when we are talking about memberships of culturally-recognisable categories" (p. 260). (It should be noted that scholars in genre theory would probably push his description of genre further in terms of genre's constitutive power of rhetorical situations.) Drawing particularly on work by Gerard Steen, Lee argues for genre categories at the basic level, where genres are maximally distinct; many existing corpora, he points out in a survey of ICE-GB, LOB, and the BNC, mix supergenres, genres, and subgenres in their "genre" classifications. After an extended critique of the BNC categories and titles, Lee offers the BNC Index (which works from three existing resources), "a comprehensive, user-friendly, 'one-stop' database of information in the BNC" (p. 274). Lee notes that some decisions were, of course, subjective, and some corpus users may disagree with his decisions, but the taxonomy and decisions are laid out plainly here. Importantly, the Index opens the possibility of creating

specialized sub-corpora for research or teaching/learning. And in the first paper of the volume, "The Learner as Corpus Designer," Guy Aston argues persuasively for the benefits of asking students to extract sub-corpora from larger ones, which can be specifically targeted and provide learners experience in corpus design and evaluation.

Laura Gavioli, in "Some Thoughts on the Problem of Representing ESP through Small Corpora," addresses the lack of an agreed-upon set of criteria for adequate representativeness of small corpora. Her description of students' work with small corpora of medical research articles highlights some of the benefits and pitfalls of these small ESP corpora. Claire Kennedy and Tiziana Miceli ("The *CWIC* Project: Developing and Using a Corpus for Intermediate Italian Students") argue for accessibility over representativeness in smaller corpora, given their work teaching Italian to intermediate students in Australia. The Contemporary Written Italian Corpus (CWIC) is made up of interactive, short (whole) texts that can serve as models of expert performances in the types of texts students must themselves produce. This way, students can find models for expressing particular rhetorical moves and answer their own questions along the lines, "Should I use X or Y here?" The rationale behind CWIC echoes Averil Coxhead's argument earlier in the volume: language teachers should teach materials which are directly relevant to the learners ("The Academic Word List: A Corpus-based Word List for Academic Purposes"). Coxhead is primarily focused on the implications of the Academic Word List for vocabulary learning and teaching, but the principle clearly applies to the design of many of the specialized corpora described in the volume.

John Flowerdew, in "Computer-assisted Analysis of Language Learner Diaries: A Qualitative Application of Word Frequency and Concordancing Software," describes perhaps the most specialized corpus in the volume: students' diaries reflecting on their learning a language as preparation for ESL teaching. Flowerdew exploits the corpus only in using keywords to locate stretches of text that capture the learners' preoccupations – it demonstrates the potential of corpus data as a qualitative research tool to assess the effectiveness of a program.

Several papers directly address the ways in which corpus-based learning empowers students. Tim Johns, in "Data-driven Learning: The Perceptual Challenge," describes learners as detectives, who when confronted with data must draw conclusions from clues. The practical examples here, from teaching collocations with prepositions to helping graduate students learn more nuanced collocational patterns, will be of interest to many ESL instructors. In "Exploring New Directions for Discovery Learning", Silvia Bernardini describes learners browsing with teachers as guides and also offers very useful specific examples, such

as teaching adverb and adjective collocations and helping students find patterns that vary by register (the benefit of incorporating multiple corpora). Natalie Kübler points out how much fundamental linguistics students are required to learn – and are motivated to learn – in working with corpora to learn authentic, specialized English. The kind of discovery learning and problem solving required by querying corpora highlight important issues in natural language processing for students and allow them to go beyond dictionaries to examine specialized meanings and syntactic environments. All of these articles describe a close, interactive relationship between instructors and students that defies any notion that computerized learning can (and should?) lead to distance learning – a concern that Christian Mair wisely raises at the end of his paper, given the widespread "technophilia" at universities and the circulation of ideas like "virtual universities."

Offering a different cautionary note about these kinds of corpus-based, discovery-learning experiences in the ESL classroom, David Wible, Feng-yi Chien, Chin-Hwa Kuo and C. C. Wang argue that unfiltered examples, which may surpass the lexical range of less advanced students, can actually be detrimental ("Toward Automating a Personalized Concordancer for Data-Driven Learning: A Lexical Difficulty Filter for Language Learners"). They describe a new tool, the Lexical Difficulty Filter (LDF), which filters examples based on the frequency of the words in the line and can be adjusted to different thresholds of lexical difficulty. The LDF, they assert, assures that concordancing tools are not restricted as "elite" tools and can simulate more specialized corpora by extracting examples from larger corpora.

John Kirk, in a paper focused on teaching linguistics as opposed to ESL ("Teaching Critical Skills in Corpus Linguistics Using the BNC"), stresses the importance of students learning critical skills (in a systematic manner, with systematic assessment) in addition to querying or concordancing skills. For example, through replicating the searches in published corpus studies, students learn to assess others' methodologies as a step toward designing their own studies.

In "Empowering Non-Native Speakers: The Hidden Surplus Value of Corpora in Continental English Departments", Christian Mair focuses specifically on empowering non-native speakers through the use of corpora. Corpora allow students to test the judgments of native speakers and the authoritative prescriptions of grammar books. As Mair points out, "using the appropriate corpora, any student can disprove statements made in the most authoritative reference grammar of English in less than half an hour" (p. 124).

Two papers in the volume demonstrate how corpus-based work can trouble traditional grammar categories. Gunter Lorenz, in "Language Corpora Rock the

Base: On Standard English Grammar, Perfective Aspect and Seemingly Adverse Corpus Evidence", calls for a distinction between perfective aspect and perfect forms, as part of a larger argument for replacing the teaching of "Good English" as the model for teaching ESL with the teaching of a "multi-layered, multi-variety standard of English" (p. 132). "Adverse" corpus findings are, in fact, a critical component of this constructivist approach to grammatical "rules." In "A Corpus-based Grammar for ELT", Dieter Mindt describes his corpus-based grammar of the English verb system (published in 2000), in which the categories are inductive. His five classes of verbs, which categorize *have to* and *like to* as catenative, offer a fascinating new way to think about the distinction between finite and non-finite verb phrases. The examples typically provide frequencies of different constructions and the entire approach of the grammatical descriptions targets English language learners.

Three other studies of verbs round out the volume. Paul Thompson focuses on core modal auxiliary verbs in selected agricultural theses in the Reading Academic Text corpus ("Modal Verbs in Academic Writing"). He concludes that EAP material tends to overemphasize the hedging role of modals; his study indicates that in various rhetorical sections within a thesis, modals serve other important functions, such as objective modality. Noëlle Serpollet tests to see if mandative *should* is decreasing through a comparison of LOB, FLOB and INTERSECT (a bilingual French-English translation corpus) and examines how mandative *should* is translated into French ("Mandative Constructions in English and their Equivalents in French – Applying a Bilingual Approach to the Theory and Practice of Translation"). Normalizing all the frequency counts could have strengthened the case made here, but it does demonstrate how bilingual corpora can aid translation studies and teaching. Claudia Claridge, in "Translating Phrasal Verbs," investigates the translation of selected English phrasal verbs into German using the Chemnitz English-German Translation Corpus. She outlines five different translation strategies evidenced in the corpus, noting the frequency and variation between translations with German particles and prefixes. Bilingual corpora clearly hold exciting possibilities for the teaching of translation and translation studies.

The world-wide web is undoubtedly, perhaps inevitably central to future developments in corpus linguistics, particularly the monitoring of ongoing language change. Antoinette Renouf tackles the question of how to study recent language change in "The Time Dimension in Modern English Corpus Linguistics" – an effective complement to the primarily synchronic focus of the rest of the volume. After describing the journalistic corpus compiled at Liverpool and the software they have developed there, Renouf outlines two systems for man-

aging the web and calls for the development of more resources and methodologies. Although we have yet to figure out how to manage "the diversity and unpredictability" of the web, to quote Burnard (p. 68), it has the potential to allow corpus linguists and their students to keep up with language change during the necessary gaps created by the time-consuming process of corpus compilation, annotation and encoding, and distribution.

The constraints of a volume with this many papers mean that authors often have to summarize and gesture towards the richness of their pedagogical approaches, of their corpora and tools, and of their studies; at the same time, this conciseness allows the volume to capture the wide range of work happening in corpus-based teaching and learning. There are minor glitches in the editing of the volume – for example, Mair's abstract seems to have been replaced by a duplicate of Bernardini's and Lee is left out of the list of contributors – but these do not distract from the content of the volume. I wished that some of the reproduced images could have been clearer, but they remain readable.

The volume as a whole highlights exciting developments in approaches to teaching and learning with corpora and in the development of resources and methodologies relevant to research as well as teaching. It stresses the importance of discovery learning – both in the classroom and in research. As one of Bernardini's students puts it, after working with corpora: "There is little certainty left: relying on intuitions, even regarding one's own native language, becomes more problematic …" (p. 179). The papers in this volume highlight the value of studying spoken and written language in use, captured in modern corpora, in terms of learning language, translating it, and studying it for linguistic description.

Charles F. Meyer. *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press, 2002. xvi + 168 pages. ISBN 0 521 80879 0 (hardback). ISBN 0 521 00490 X (paperback). Reviewed by **Claudia Claridge**, University of Kiel.

English Corpus Linguistics joins a number of other introductory corpus-linguistics books published in recent years. However, what distinguishes this publication from others available is that, instead of dealing with the field as a whole (e.g. McEnery and Wilson 1996/^{2}2001; Kennedy 1998) and/or pursuing a particular research agenda (e.g. Stubbs 1996; Biber et al. 1998), it can be described as

a kind of basic manual for corpus construction and analysis, with the emphasis on the former. Thus, it fills a gap in the existing literature.

The structure of the book falls into five sections. First, there is a preface presenting basic definitions and aims, followed by a first chapter linking corpus linguistics with linguistic theory and (practical) applications of corpus linguistic research. Then come three chapters (2-4) describing corpus construction from planning, via collection and computerization to corpus annotation, and one chapter (5) presenting a detailed case study of corpus analysis. Finally, a very brief sixth chapter both sums up and highlights possible future developments of the areas dealt with in the book. The whole is rounded off by two appendices listing available corpus resources and concordancing programmes.

In the preface, Meyer states his view of corpus linguistics as essentially a methodology, not a linguistic theory, and argues that, therefore, an increased awareness of methodological assumptions and procedures on the part of both corpus creators and users is vital for the progress of corpus linguistics (p. xiv). Corpus linguistics is indeed probably best viewed as a methodology; however, some further discussion of how the choice of a particular methodology correlates with broad, pre-existing theoretical assumptions about language and has potential theoretical repercussions or – to mention a clearly contrary view – can in fact be seen as a linguistic paradigm in its own right (cf. corpus-driven linguistics, Tognini-Bonelli 2001), would have provided a more balanced and informative approach. The preface defines a corpus as "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (p.xi). This definition at first seems overly brief and general, but the approach is narrowed down to the creation of "balanced corpora" and their use in "descriptive linguistic analysis" (p. xv), thus excluding most corpus research in computational linguistics/natural language processing, for example. This seems a wise restriction as the corpus-linguistic views and needs of the approaches just mentioned differ considerably and would have made the book unwieldy. The intended audience of the book seems to be the beginner in corpus linguistics: although Meyer does not explicitly state this (speaking only of "corpus linguists" as such, p. xiv), the structure and content, including numerous very basic aspects, as well as the study questions at the end of each chapter, imply this readership.

Chapter 1 discusses the relationship of corpus linguistics to generative linguistics and to functional theories of language, concluding – unsurprisingly – that it is the latter, not the former, that shows any interest in corpus linguistics. While Meyer gives examples to show that corpus linguistics can in fact contribute not insignificant insights to generative theory (p. 4f.), he thinks it unlikely

that generative linguists will ever develop much interest in using corpora. If this is so, it prompts the question why corpus linguists repeatedly feel the urge to one-sidedly topicalize this ultimately not very fruitful issue. The greater part of Chapter 1 is devoted to an overview of the place of corpus-based research in various fields, ranging from grammar- and dictionary-writing to language pedagogy, and taking in historical linguistics and contrastive analysis on the way. The treatment here is necessarily cursory, but it serves the purpose of highlighting the wide range of the possible applications of corpora and of stimulating further interest in corpus linguistics in readers of many different linguistic persuasions.

Chapter 2 is concerned with the planning stage of corpus construction. Meyer stresses the importance of careful initial planning in setting up the criteria for collection, which are determined by the future uses of the corpus, while at the same time retaining flexibility for adjustment in the compilation process. The chapter presents a comprehensive and clear discussion of the following compilation criteria: size of corpus, genres, length of text samples, number of texts, range of speakers, time frame, native vs. non-native speakers, and sociolinguistic variables (age, gender, dialect, education). Throughout the discussion, alternative approaches are evaluated and problematic points highlighted, e.g. the difficulties probability sampling can present (p. 43f.). However, not all of the aspects are treated as thoroughly as one might wish, a case in point being the question of the inclusion of complete texts or of text samples. Discussion of this aspect is biased towards the latter solution, without a clear statement of the potential advantages of using complete texts, among them the uneven distribution of linguistic features throughout texts as well as the general consideration that text-linguistic studies (beyond register comparison) should also be possible with corpora. The chapter uses the *BNC* as its example for illustrating the various criteria, which does not seem to be the most logical or useful choice: how many *beginning* corpus linguists would start with compiling a corpus of that scale – and thus have corresponding problems? It might also have been helpful to list more clearly those corpora that are in some way representative in their treatment of one or the other criterion discussed, so that the interested reader could have a closer look for her/himself at corpus linguistic problems and solutions.

Chapter 3 deals with the practicalities of collecting and computerizing samples of spoken and written English. This is done in a very down-to-earth and helpful way, with close attention paid to technical points (e.g. recording and transcription equipment, OCRs), procedural aspects (e.g. record keeping, materials storage) and ethical/legal issues (recording permission, copyright). Some of

the information given here may become outdated fairly fast (e.g. technical aspects), but raising awareness of the menial and mundane aspects of corpus linguistics is a very necessary and laudable thing to do. However, the chapter could have been more detailed and comprehensive in some respects. Written texts are admittedly less problematic than spoken ones; none the less the treatment they receive here is somewhat too brief and neglects the challenges they potentially represent. A possible reliance on electronically available texts is presented in a rather optimistic light and scanning is too much taken for granted, the latter perhaps due to the double bias resulting from thinking mostly in terms of printed and modern texts. Hand-written modern texts (e.g. letters, student essays) are not mentioned at all, while older texts, and manuscripts especially, are touched on only briefly. The discussion of computerizing speech is more detailed and necessarily shades into annotation matters when intonation is mentioned. What is not mentioned here is the possibility of sound files accompanying the transcription (as is the case with *COLT* and the *Santa Barbara Corpus of Spoken American English*) and alignment of text and sound, a practice which, with increasingly available computer space, might – indeed should – become more common.

Annotation of various types, namely structural markup, part-of-speech tagging and parsing, is the topic of Chapter 4. According to Meyer, annotation is necessary for a corpus to be "fully useful to potential users" (p. 81), which seems to be putting things too strongly. First, there are numerous features which are (fairly) easily retrievable without (grammatical) annotation and many linguistic questions to be pursued which are not affected by the surface features of the text (layout etc.). Secondly, it is not sufficiently highlighted that any form of annotation, but especially grammatical annotation, is already an interpretation (although cf. Meyer's own remark that "tagsets reflect differing conceptions of English grammar", p. 90) – an interpretation, moreover, that might ultimately contribute to obscuring a feature an individual analyst is looking for. A good solution for the corpus creator might actually be to provide both an annotated and a 'bare' text version of a corpus. As to structural markup, this receives rather too brief a discussion; in consequence, the aims and potential linguistic usefulness of this type of mark-up does not become clear. Furthermore, the main example is SGML as used in the *ICE* project, which might not be the best choice, because it is merely SGML-conformant and predates the TEI guidelines. The *BNC* would have served as a better illustration here. Moreover, a more detailed one of the SGML/XML/TEI complex would have been an advantage, in particular as it is the only comprehensive system with aspirations to become a standard. In view of the fact that the book is also intended for the corpus user

(and not only the compiler), a discussion, however brief, of earlier and/or related but supplemented annotation systems (e.g. COCOA, RET) might have been included. The chapter also includes a treatment of speech/intonation annotation. A point that might have been mentioned in that context is that (some) intonation markup conventions can actually make analysis – especially automatic computer analysis – harder, e.g. forms such as *ti=me* in the *SBC* example on page 85. The corpus user perspective is somewhat neglected throughout Chapters 2-4; they would also have profited from a greater number of examples, e.g. showing different annotation systems (for the same text, perhaps) and texts at different stages of annotation. This would have been very useful for the novice corpus linguist in particular.

Corpus analysis, i.e. the user perspective, is the focus of Chapter 5 and is exhaustively illustrated with a single well-chosen case study, Meyer investigating the occurrence of pseudo-titles in the press sub-corpora of seven *ICE* corpora. The comparative approach provides the opportunity to look again in more detail at corpus compilation, representativeness, and available annotation, this time from the analyst's perspective. The chosen feature is one that is not automatically retrievable in an untagged/unparsed corpus (six of the seven corpora used). This may not be very typical of corpus linguistics methodology as a whole, but the choice highlights the point that automatic retrievability should naturally not be a guide to what is being researched. Unfortunately, Meyer does not comment on the manual retrieval procedure and its results, merely mentioning it (p. 119); there are certainly degrees of manual retrieval, and the process can also turn up findings at odds with those of automatic retrieval, as well as findings the researcher did not expect. Meyer argues for combining quantitative and qualitative aspects in the analysis of corpus data, a very important point as the balance can easily become tilted towards the former in corpus linguistics. The chapter works through the whole process of analysis step by step, thoroughly comparing options and motivating the decisions to be taken, and linking the aspect in hand to more general questions wherever possible. The whole research procedure thus becomes highly accessible and comprehensible even for readers with little to no experience in the field.

In conclusion, the work under consideration here is a very welcome addition to the range of corpus linguistic publications. It offers the beginner a brief yet valuable introduction to the basic aims and – especially – the research procedures of corpus linguistics and thus serves a real need. Perhaps the content of the book could have been more clearly reflected in the title in order to attract the attention of its intended readership. It can be argued that certain aspects have not been treated with sufficient explicitness and detail or in adequate depth, in par-

ticular for readers with little previous knowledge (cf. the remarks above), but remedying this point would have considerably increased the length of the book. However, one helpful addition would have been a 'further reading' section after every chapter.

References

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Kennedy, Graeme. 1998. *An introduction to corpus linguistics*. London and New York: Longman.

McEnery, Tony and Andrew Wilson. 1996/2nd ed. 2001. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Stubbs, Michael. 1996. *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford and Cambridge, Mass.: Blackwell.

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

Randi Reppen, Susan M. Fitzmaurice, and **Douglas Biber** (eds.). *Using corpora to explore linguistic variation*. Studies in Corpus Linguistics. Amsterdam and Philadelphia: John Benjamins Publishing Company, 2002. xii +274 pp. ISBN 90-272-22279-7. (Eur.) / 1-58811-283-7 (US). Reviewed by **Nelleke Oostdijk**, University of Nijmegen.

Using corpora to explore linguistic variation opens with an introductory chapter by the editors. They describe the linguistic scene, providing the background that ties together the papers presented in the present volume. They also explain the organizational principles adopted. In all, the introductory chapter is very informative and highly illuminating, as it clarifies what the book is about, how it is organized, and why it comprises the papers it does. The editors characterize the book as "a collection of papers that illustrate ways in which linguistic variation can be explored through corpus-based investigation" (p. viii). The organization of the book has been guided by the primary research questions addressed in the respective papers. Thus each of the papers in part I "focuses on the use of a particular linguistic feature (a single word, a set of related words, a grammatical construction, or the interaction between particular words and grammatical structures", while the papers in part II typically focus "on the overall characteristics of language varieties, either a single dialect or register, or the similarities and differences among a range of dialects/registers." (p. viii). In the third and final part, the same perspectives are applied in a historical context.

Below I briefly describe the contents of each of the papers in the three parts, before I go on to discuss the book along more general lines.

Part I: Exploring variation in the use of linguistic features

1. Deanna Poos and Rita Simpson, 'Cross-disciplinary comparison of hedging. Some findings from the Michigan Corpus of Academic English'

While Lakoff (1975) claims that hedging is one of the qualities of feminine speech, others have failed to find evidence for this claim. In their paper, Poos and Simpson investigate to what extent hedging is related to gender differences. More specifically, their research focuses on the use of *kind of* and *sort of* as prototypical examples of hedging devices in academic spoken English. Their analysis shows that academic discipline is a stronger predictor for the occurrence than gender, while the functions of these devices are rather diverse. Thus, apart from

expressing inexactitude, *kind of* and *sort of* may be used, for example, to soften the force of a stance or opinion, or to mitigate a criticism or request.

2. Fiona Farr and Anne O'Keefe, 'Would as hedging device in an Irish context: An intra-varietal comparison of institutionalised spoken interaction'

Like Poos and Simpson, Farr and O'Keefe are also concerned with hedging. Their perspective, however, is rather different, as they look at the socio-cultural context as a factor in explaining why speakers hedge in discourse. Following an analysis of the hedging involving the use of *would* that occurs in two institutional face-to-face interactions in an Irish setting, they arrive at a tree-tiered model for the analysis of spoken interaction.

3. Michael McCarthy, 'Good listenership made plain: British and American non-minimal response tokens in everyday conversation'

McCarthy examines listeners' responses in exchanges between speakers in everyday conversations. In his research, he focuses on the role of adjectives and adverbs "which typically occur at points of speaker change in every talk, and which either account for the whole of the listener response or are the first item in the listener response" (p. 49). An examination of two varieties, viz. British and American spoken English, shows that 'good listenership' involves that the listener takes on an active role not only in acknowledging what the speaker says, but also in investing in what McCarthy describes as the relational aspects of discourse, creating and maintaining sociability and affective well-being in their responses.

4. Graeme Kennedy, 'Variation in the distribution of modal verbs in the British National Corpus'

Kennedy's large scale study of the distribution in the BNC of modal verbs and the verb phrase structures they occur in, confirms the findings of earlier studies which were based on smaller and/or less representative corpora. The analysis of some 1.45 million occurrences of modals shows that there is great deal of variation in their distribution in different genres and media. The use of different modals varies, depending on the meaning the modal carries, the texts and the genre it occurs in (spoken or written), the structure of the verb phrase, and whether the verb phrase is affirmative or negative. At the same time, however, the use of modals in complex verb phrase structures is found to be quite stable.

5. Ferdinand De Haan, 'Strong modality and negation in Russian'

In his study of modality and negation in Russian, De Haan examines the relation between the scope of negation and modality on the one hand and syntactic position on the other hand. The modal system in Russian is not as grammaticalized as in English and also the sentence structure is different. Modality in Russian is defined by its meaning, rather than the syntactic characteristics. On the basis of the results obtained De Haan reaches the conclusion that "it would appear that languages go from a syntactic approach (where placement of the negation in the sentence determines its scope) to a semantic approach (where the scope of the negation is determined by the modal verb)" (p. 108).

6. David Okey, 'Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines'

Over the years, the existence of ready-to-use strings (referred to as prefabricated strings, lexical phrases, etc.) has been acknowledged in many studies. Okey in his paper undertakes to "provide a clearer, less intuitive insight to these units" (p. 111). He uses a subset of the BNC to investigate the use of the lexical phrase *it is/has been (often) asserted/believed/noted that X* as it occurs in academic writing in the fields of social science, medicine and engineering. Apart from the topic priming function, four other discourse functions are identified that are associated with this lexical phrase.

7. Viviana Cortes, 'Lexical bundles in Freshman composition'

Lexical bundles as defined in Biber et al. (1999) are extended collocations, i.e. sequences of three, four, five or six words that statistically co-occur in a register. Cortes investigates the occurrence of four-word lexical bundles in the writing of freshman university students. Her findings do not confirm her working hypothesis, which predicts that the bundles used by the students probably resemble more closely the bundles found in conversation than those found in academic prose. Instead, students seem to "closely imitate" the most frequent bundles encountered in academic prose. However, a careful analysis of the findings reveals that there are pervasive differences in the way that freshman students use these bundles.

8. Charles Meyer, 'Pseudo-titles in the press genre of various components of the International Corpus of English'

In his paper, Meyer presents an analysis of the occurrence of pseudo-titles across seven different regional varieties of English. Finding its origin in Ameri-

can English press reportage, the use of pseudo-titles has spread to other varieties of English, including British English and New Zealand English. While the use of pseudo-titles in American English is considered unmarked, in British English it is stigmatized (pseudo-titles are found to occur mainly in tabloids; in more formal newspapers they are generally prohibited). Meyer's findings lead him to observe that "the spread of pseudo-titles in press writing not only shows that a grammatical construction can be borrowed from one variety to another but that once the construction is borrowed, the constraints on its usage can change, leading to new forms." (p. 148).

9. Susan Hunston, 'Pattern grammar, language teaching, and linguistic variation: Applications of a corpus-driven grammar'

Following a concise introduction to the principles of pattern grammar, Hunston presents an interesting discussion on the merits of this type of grammar and its application to the study of language variation on the one hand, and language teaching on the other. It is claimed that pattern grammar is "an approach to language which maintains the generalising characteristics of grammatical descriptions while prioritising the behaviour of individual lexical items" (p. 167). The discovery of patterns – a pattern is defined as "a sequence of grammar words, word types or clause types which co-occur with a given lexical item" (p. 169) – benefits from the availability of large corpora such as COBUILD, although, as Hunston is careful to point out, intuition also comes into play in this, as the co-occurrence of lexis and pattern is not random but is associated particularly with meaning, while this association is not predictive. Moreover, there is evidence that patterns change over time.

Part II: Exploring dialect or register variation

10. Chandrika Rogers, 'Syntactic features of Indian English: An examination of written Indian English'

Rogers investigates three syntactic features that have previously been identified as characteristic features of Indian English. They are: use of the progressive with stative verbs, use of the present tense and the past perfect, and use of prepositional verbs. Her present study, which is based on the use of the stative verbs *have*, *know*, *want*, *like*, *hear* and *look* in an 800,000 word corpus of written Indian English, does not confirm earlier findings. In comparison with British and American English, in the Indian English data the progressive is more frequent in general, i.e. not specifically with stative verbs. The corpus comprises

insufficient data to draw conclusions on the use of the present and past perfect. Rogers suggests that an investigation of spoken data might yield rather different results. With respect to the use of prepositional verbs and patterns of preposition use, the data show Indian English to be markedly different from British and American English.

11. Eniko Csomay, 'Variation in academic lectures: Interactivity and level of instruction'

Csomay sets out to investigate the linguistic characteristics of academic lectures as they actually occur in real settings (as opposed to experimental settings which have been used in earlier studies). The present study involves 23 features that have been identified in Biber (1988) as characteristic of academic prose and conversation. An analysis of data from 176 lectures taken from the T2K-SWAL Corpus brings to light two situational parameters that have an effect on the linguistic features present in the lectures, viz. the degree of interactivity and the level of instruction.

Part III: 'Historical variation'

12. Susan Fitzmaurice, 'The textual resolution of structural ambiguity in eighteenth-century English: A corpus linguistic study of patterns in negation'

Within a context in which two grammatical systems for the formation of negative clauses co-exist, Fitzmaurice investigates whether this co-existence potentially gives rise to ambiguity and, if so, how speakers deal with this. The older of the two systems is the *do*-less one in which the main verb is followed by *not*, while the newer system is the one that uses *do*-support. The older system is understood to be recessive. An in-depth study of the different patterns in which the negative can occur reveals that the two systems occur side by side without the older system getting in the way of the newer one.

13. Christer Geisler, 'Investigating register variation in nineteenth-century English: A multi-dimensional comparison'

Geisler follows in Biber's footsteps in his multi-dimensional analysis of the development of English registers through the nineteenth century, a period which in other studies so far has largely been neglected. Adopting the sets of co-occurring grammatical features identified in Biber (1988) and the four dimensions associated with these, Geisler investigates the development of seven registers over three time periods: 1800-1830, 1850-1870, and 1870-1900. His findings

show that some of the registers are rather heterogeneous. The results obtained in this study only in part confirm the findings for other time periods.

Using corpora to explore linguistic variation is a book that clearly belongs in the tradition of what can be characterized as 'the Biber school', although in some contributions also the influence of Sinclair's work is apparent. With one or two exceptions, all papers build upon and extend previous research carried out by Biber and others (especially his work published in Biber 1988 and 1995, but also the joint publications with Finegan on historical English, incl. Biber and Finegan 1989, 1992 and 1997), while frequent reference is also made to the *Longman grammar of spoken and written English* (LGSWE 1999). Biber's research on the dimensions of linguistic variation and the linguistic features that characterize these, together with the findings published in the LGSWE with respect to the frequency and distribution of linguistic structures, lexical bundles, etc. is highly influential and pervasive in the research presented in the papers in this volume, as much in the research questions that are being investigated as in the (methodological) approach adopted.

Central to all papers is their use of corpus data. Some of the research reported on in this volume clearly benefits from the availability of (relatively) novel resources, such as the Michigan Corpus of Academic Writing (MICASE), the T2K-SWAL Corpus and CONCE (Corpus of Nineteenth-Century English). Although some researchers explore well-known corpora such as the BNC, the International Corpus of English (ICE) or COBUILD, others for their specific research find a need to compile special data collections. This seems to support the claim that, although there are so many corpora available already, still more corpora are needed.

While the editors in their introduction describe the methodological challenges that researchers encounter in analysing the influence of contextual factors on linguistic variation, the authors of the individual papers should be complimented on their work: without exception, they appear to have a strong awareness of the methodological sanity of what they are doing. They are quite ready to point out any limitations of the data they have used, or to identify possible flaws or defects of their investigative approach. All the papers in this volume report on substantial work. There is ample reference to the linguistic literature, and much care is taken to relate the present findings to results obtained in earlier studies.

Most of the authors of the papers included were present at the Second North American Conference on Corpus Linguistics and Language Teaching held at Northern Arizona University in Flagstaff, Arizona in the spring of 2000. This

might explain why many papers also pay (some) attention to the implications their results (may) have for pedagogical applications and teaching methods. For the time being, I think, the papers contribute to raising an awareness of different aspects of linguistic variation. Before the results presented here can be put to any practical use, however, much more work will yet have to be done.

References

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

Biber, Douglas. 1995. *Dimensions of register variation*. New York: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1989. Drift and evolution of English style: A history of three genres. *Language* 65: 487-517.

Biber, Douglas and Edward Finegan. 1992. The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen, and I. Taavitsainen (eds.). *History of Englishes: New methods and interpretations in historical linguistics*, 688-704. Berlin: Mouton de Gruyter.

Biber, Douglas and Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In T. Nevalainen and L. Kahlas-Tarkka (eds.). *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253-275. Helsinki: Societé Néophilologique de Helsinki.

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

Biber, Douglas, Randi Reppen, V. Clark, J. Walter. 2001. Representing spoken language in university settings. The design and construction of the spoken component of the T2K-SWAL Corpus. In R. Simpson and J. Swales (eds.). *Corpus linguistics in North America*, 48-57. Ann Arbor MI: University of Michigan Press.

Lakoff, Robin. 1975. *Language and woman's place*. New York: Harper and Row.