Reviews

Sidney Greenbaum, *The Oxford English Grammar*. Oxford University Press, 1996. xv + 652. Reviewed by **Peter Collins**, University of New South Wales, Sydney, Australia.

The Oxford English Grammar (henceforth 'OEG') by Sidney Greenbaum (henceforth 'SG') is much more than a grammar as that term is conventionally used and understood. Of the twelve chapters in OEG, only four (3–6) deal directly with syntax and inflectional morphology, these representing approximately one half of the book's contents. It is fitting that the last major publication from a scholar who has devoted a lifetime to researching, teaching and writing about the English language should be so ambitious in its scope. Everywhere the author's formidable knowledge of the history, diversity and structural intricacies of English is in evidence (as for instance in his discussion of personal pronouns, where SG takes us back in time with an explanation of the origins of *'em* as a survival of Middle English *hem*, right up to the present day with a reference to *themself* as a recently introduced gender-neutral singular form, and across the Atlantic with a comment on *y'all* as an informal American combination for the second person plural).

OEG throws down the gauntlet to descriptive grammarians in two ways. Firstly, through the use of numerous citations it achieves an aura of authenticity unparalleled in rival grammars (including the Collins-Cobuild, in which citations are often 'doctored' and are presented without source references). Citations are derived from two sources: ICE-GB (the British million-word component of the International Corpus of English, comprising 60 per cent spoken and 40 per cent written material, and drawing on language used in the period 1990–3: curiously, a stray example from the as-yet-incomplete ICE-USA is introduced on p 375 without comment), and the *Wall Street Journal* (about 3 million words on CD-ROM from the 1989 issues). Secondly, OEG achieves a high level of 'user-friendliness': each chapter is prefaced by a list of contents and a useful point-by-point summary, a glossary of key terms is included, many of the endnotes supply up-to-date references for further reading, and readily comprehensible diagrams are used as an explanatory tool

(notably, throughout the discussion of clausal coordination and subordination).

In this review I shall concentrate mainly on the centrally grammatical chapters (on the assumption that most readers will consult OEG for the information contained therein). However I shall begin with some remarks on the more 'peripheral chapters' (1, 2, and 7–12). The first two chapters 'set the scene' for the grammatical core of OEG. Chapter 1 ('The English Language') details the geographical spread of English around the globe. It includes observations that will be taken as axiomatic by the linguistically-trained reader, but which are certainly worth asserting for the lay reader. For instance, SG observes that the prestige and practical value of English as an international language 'cannot be attributed to the intrinsic superiority of English over potential other candidates' (p 13), which he confirms by noting that, while some features of English may make it easier to to learn than its rivals (few inflections, natural gender, etc), others make it harder (frequency of idioms, mismatches between pronunciation and spelling, etc). Another truism pointed out by SG is that 'correct English' (conformity to the norms of Standard English) is distinguishable from 'good English' (effective or aesthetic use of the language). He gives the notion of 'good English' an interesting contemporary twist by interpreting it to include language that avoids offensive and discriminatory (eg racist, sexist) usages.

Chapter 2 ('The Nature and Scope of Grammar') discusses various types of grammar (prescriptive, descriptive and 'theoretical'). About half the chapter is devoted to a discussion of Chomsky's generative model, even though this has little influence on the grammatical presentation in OEG. Surprisingly, no comment is offered on the grammatical tradition in which OEG is situated.

Chapters 7–12 take up a variety of topics, some of which one might expect to be dealt with in a linguistics textbook rather than a grammar. Chapter 7 ('Text') reflects the commitment of SG and his co-authors in the famous Longman grammars to textual, beyond-the-sentence, phenomena. The scope of the treatment in this chapter is comparable to, though inevitably less detailed than, that in Quirk, Greenbaum, Leech and Svartvik's (1985) *A Comprehensive Grammar of the English Language*. Chapter 8 ('Words and their meanings') deals with such topics as lexical semantics, etymology, meaning change and dictionaries. Chapter 9 ('The formation of words') is about word formation and lexical morphology. The next, short, chapter on phonetics bears the curious title 'Sounds and tunes' (curious, insofar as 'tunes' are not mentioned at all in the

chapter). The last two chapters deal with punctuation and spelling, and will be of interest to readers who might otherwise consult style guides for information on these matters.

The grammatical core of OEG, it must be said, is somewhat disappointing. There is not space here to offer detailed criticisms: the comments that follow are necessarily selective. Firstly, there are several places in which SG's analyses are open to question. For instance indirect objects are, appropriately, described in positional terms (though, there's no mention of the indirect-direct ordering that is possible in some dialects, as in *I gave it her*). Accordingly, there seems to be little justification for classifying instances such as the following as indirect rather than direct objects (p 65): [15]... I promise *you*; [16]...God doesn't tell *us*. Another debatable analysis occurs with so-called 'nominal adjectives' (eg *the handicapped*, *the illiterate*). If, as we are led to believe, the main syntactic property of 'nominal adjectives' is their capacity to function as head of a NP, then almost any gradable adjective will potentially be a member of the class (eg *silliest* in *There were many proposals, the silliest of them being that we should strike*). SG doesn't mention an alternative and widely accepted analysis of phrases such as *the handicapped* as NPs lacking an overt head. Such an analysis gains support from the possibility of premodifying the 'nominal adjective' with an intensifier, as in *The very handicapped require specialised care* (the intensifier *very* characteristically modifying adjectives, not nouns). A third example of descriptive inadequacy is to be found in the discussion of 'basic' sentences, in particular the discussion of 'rearrangements' to the basic patterns. Clefts, extraposed-subject sentences and existentials are referred to as 'drastic' rearrangements, with no explanation offered for the use of this term. It would appear that a relevant distinction here is between non-basic sentences which change the subject ('drastic') and those that do not ('non-drastic'). As a final example of descriptive weakness, consider SG's use of the notion of 'judgement' in his definition of modality. Judgement is clearly relevant to epistemic modality, but less obviously so to deontic modality. Furthermore, ability and intention are included in the same category as permission and obligation, but the 'subject-orientation' of the former surely sets them apart quite markedly from the latter.

Another disturbing feature of the grammatical description in OEG is SG's tendency to present traditional and arguably outmoded analyses. For instance in Chapter 3, we are presented with the familiar tripartite classification of sentences into 'simple', 'complex' and 'compound'. The

usefulness of the notion of 'complex sentence' is open to challenge: we don't after all subclassify NPs according to whether or not they contain an embedded clause. Perhaps a more valid distinction would be between 'clausal sentences' and 'compound sentences'. As a second example of an outmoded analysis consider the treatment of post-head modification in the NP and AdjP. There are one or two allusions to the treatment of certain post-head dependents as complements rather than modifiers (eg *it's an opportunity for Christians everywhere to at least unite in prayer for a speedy end to the war in the Gulf* (p 219); *I was afraid of him* (p 292)), but discussion of this possibility is relegated to an end note on pp 591–592. Given the considerable discussion of the modifier-complement distinction in the linguistic literature, one might have expected it to receive more prominence in OEG.

In a number of places SG's use of particular terms is misleading or at least open to challenge. For instance, we find *refer* used to express not the logico-semantic relationship of reference, but rather the discourse-grammatical relationship of anaphora: 'the initial pronoun *it* in the second sentence refers back to the initial phrase' (p 44); 'the second mention refers back to the previous mention' (p 243). Or again, the major use of exclamative sentences is said to be 'exclamation', but 'exclamative statement' would surely be more precise given that, as SG himself subsequently notes on p 53 'exclamations may take the form of declaratives, imperatives and interrogatives'. Finally, one may note the inconsistency of SG's use of 'genitive' rather than 'possessive' (preferred on the grounds that the meaning signalled is not always 'possession': see pp 113–115), when at the same time 'subjective/objective' are used in preference to 'nominative/accusative' (despite the fact that the 'subjective' and 'objective' cases do not always correspond to the subject and object functions respectively; for instance accusative *me* functions as subject in *for me to go*).

Occasionally OEG fails to discriminate grammatical and semantic categories. Witness the confusion that results from the lack of distinction between the grammatical class of proper noun and the semantic category of proper name, evident in formulations such as (p 97): 'Proper names are treated as common nouns when they do not have unique references'. Surely *Julians* ([2] on p 97) in *I've got a lot of Julians in my class* is a proper noun, rather than a common noun, one that is here not used as a proper name. As a second example, witness the failure to distinguish grammar from logico-semantics in the discussion of negation. On p 56 the presence of two negative words in the sentence *None of the countries*

have no political prisoners is claimed to make it positive. Grammatically, however, we have a negative clause here, as can be seen from the possibility of appending the emotively-neutral positive tag question *do they?*.

A final point relates to the slippage between form and function that one encounters in OEG. For example, 'verb' is used as both a class and function label, and there is no class term corresponding to 'predicate' ('VP' in the generative tradition), even though SG finds occasion to invoke the notion (as for example when discussing ellipsis in Section 6.1, where he is forced to use the circumlocution 'the main verb and its possible complements').

In conclusion, OEG is a book with strengths and weaknesses. The treatment of grammar is hardly, as the cover blurb proclaims, 'up-to-date and ground-breaking'. It is in fact disappointingly traditional and prone to the same types of confusion between grammatical and semantic categories, between form and function, and other descriptive weaknesses that are found in works by grammarians of less stature than SG. Nevertheless, OEG has some very attractive features, including its vigorous use of corpus data (so that it is only rarely that we encounter examples of questionable grammaticality such as *Who do easterly winds bring this extreme cold?* (p 65) and *An official letter of complaint is being sent you (p 69)), the breathtakingly broad scope which covers – as the cover blurb tells us – 'many issues which have not been widely dealt with in previous guides to English grammar', and the careful attention paid to presentation and accessibility.*

Guy Aston and **Lou Burnard**. *The BNC Handbook. Exploring the British National Corpus with SARA* (Edinburgh Textbooks in Empirical Linguistics). Edinburgh: Edinburgh University Press. 1998. xiv + 256 pp. ISBN 0 7486 1054 5 (cased); ISBN 0 7486 1055 3 (paperback). Reviewed by **Pieter de Haan**, University of Nijmegen.

A review of this book must inevitably imply a review of the software system whose use it is intended to demonstrate. Let me begin by saying that I enjoyed reading the book very much, and as the reader is invited to try the various features of SARA him- or herself, this is, of course,

what I did. This means that my personal appreciation of how the program performs 'in real time' is affected by the way it has been set-up on the network in our faculty in Nijmegen, how the network operates, and also how my own PC coped with the various operations. In the limited time I had at my disposal for the review it turned out that not only had my PC fallen seriously ill, but also the network caused considerable problems. In the event I had to work from a friendly colleague's machine, which was fine.

I discovered, however, that the version of the BNC that we have on our network was release 1.0. Although the authors warn the reader (on p 46) that the results of the various searches may be different from the ones described in the book if a different version of SARA (ie other than version 930) and/or the BNC (ie other than release 1.1) is used, I had not expected these differences to be so dramatic. The authors indicate (on p 46) that 'particularly the frequencies reported on several words in version 1.0 of the BNC (the first released version) may be somewhat lower than those for version 1.1.' It turned out that in the very first task I found only one instance of the word *cracksman*, when I should have found four! Surprisingly enough, the phrase query (see below) did not recognise instances in the corpus where *Cracksman* had been spelled with a capital. Again, the search for the second word, *whammy*, yielded 44 examples in 29 texts where I should have found 46 examples in 31 texts. The authors do not explain what causes these differences.

I would have thought that, even if differences were known to exist between the two releases of the BNC, perhaps a more careful selection of words where no differences were found would have been more elegant. But of course I do not know whether the number of copies of release 1.0 that have been distributed is very small, so that most users will not come across these problems anyway.

Let me first outline the content of the book. It is divided into three parts. Part I, entitled *Corpus linguistics and the BNC*, provides background information on the BNC project. Part II is the central part. It is entitled *Exploring the BNC with SARA*, and it guides the reader through the various operations of SARA in ten different tasks. Part III, finally, is a Reference guide. The book is concluded with a bibliography and an index.

The two main chapters of Part I discuss the BNC against the background

of corpus linguistics. In Chapter 1 the authors provide a brief overview of the field of corpus linguistics, and mention the various uses to which corpora have been put in the past and are still being put today. They list the major corpora in existence today and give a brief description of their various peculiarities. They also dwell on the question of corpus design and various types of annotation. Chapter 2 focuses on the BNC itself. Here the reader finds how the corpus was constructed, what considerations played a role in determining the types of texts eventually incorporated in the corpus.

We learn that the written part covers 90,000,000 words and the spoken part 10,000,000 words. Sixty per cent of the written part comes from books, 30 per cent from periodicals, and the remaining 10 per cent from miscellaneous sources. By a different selection feature, the written texts are divided into imaginative text (25 per cent) and informative texts (75 per cent). A third selection feature, time, divides the texts into those produced between 1960 and 1974 (less than three per cent), between 1975 and 1993 (almost 90 per cent), while about eight per cent remains unclassified.

The 10,000,000-word spoken corpus is divided into two roughly equal parts, a demographic part of informal conversations, and a context-governed part of more formal situations, such as lectures, meetings, radio programmes, etc. Roughly 45 per cent of the spoken material was captured in the South, and roughly 25 per cent in the Midlands and the North respectively, while about five per cent remains unclassified. About 75 per cent of all the interaction types are dialogues.

Some attention is paid to encoding and annotation of the corpus. By encoding is meant the information about the texts, as well as about the text structure, for instance, title, author, year of publication, but also division into chapters, sections, paragraphs etc. The term annotation is reserved for part-of-speech tagging. This has been done by means of CLAWS4, an automatic tagger developed at the university of Lancaster. The authors point out that there is an estimated error rate of approximately two per cent, while slightly under five percent of all the words in the corpus have received a so-called portmanteau tag, which is a two-valued tag, applied in cases where the automatic tagger could not decide between two different tags. Obviously, because of the size of the corpus, manual post-editing of the tagged version of the corpus was out of the question. A new release of the corpus will, however, be tagged with a refined version of the tagger, based on a two per cent sample of the BNC which has been manually corrected.

Chapter 3 very briefly discusses the potential of future corpora. The authors envisage a situation in which it will become increasingly possible for researchers to create their own corpora using the growing availability of electronic texts on the World Wide Web, and the increasing sophistication of web search and retrieval facilities. The application of future corpora will range from teaching to forensics and public communication.

Part II, as I indicated above, is the central part. The authors take the reader through the various features and operations of SARA in ten tasks, which the reader is invited to carry out. The authors describe exactly which steps have to be taken in order to achieve the results of the queries in each of the tasks. Moreover, they also provide the results of the queries, so that the reader can check her/his results against those in the book. This makes this part of the book like a tutorial. It is particularly valuable if you work your way through it from the beginning to the end. It is less suitable as a work of reference for looking things up, but the authors have provided a reference guide (Part III) as well as an index, both of which help the more advanced user to find what s/he wants.

Each task is presented in the same way: first the task is described in general terms, and then the various features of SARA that play a role in the task are mentioned listwise. Moreover, the authors indicate for each task which operations and features are assumed to be known. Thus, for the first task, all the reader/user is expected to know is how to double-click a mouse, how to move and resize windows, how to copy things to the clipboard, and how to switch between windows. At the start of the second task the reader is assumed to be familiar with the features that are introduced in the first task, etc. This ensures a gradual extension of the her/his grasp of the various possibilities that SARA offers, but, as I said earlier, it also means that the tasks should be carried out in that order. Each task is concluded with a section in which related linguistic problems are discussed, with a suggestion for specific queries.

The first few tasks familiarise the reader with the simplest ways of using SARA. Most of that is menu-driven, and all the reader is expected to type in is the word or words that (s)he is looking for. Gradually the reader is introduced to the possibilities of formulating more complex queries, involving the part-of-speech tags, the SGML mark-up, and

various combinations of features that can either be formulated by means of the query builder or directly by means of the query language. All possibilities but the last are menu-driven, and make no particular demand on the reader in terms of her/his mastery of the query language. This is particularly useful for the beginning SARA user.

There are seven different ways in which queries can be formulated, viz by means of the:

- word query dialogue box
- phrase query dialogue box
- part-of-speech query dialogue box
- pattern query dialogue box
- SGML query dialogue box
- query builder dialogue box
- CQL (the SARA Corpus Query Language) query dialogue box

Task #1 introduces the reader to the phrase query. I wondered why, in a task that looked for words, the word query was not the most obvious choice, but this was made clear to me in task #2, which introduced the word query. The essential difference between the two queries is that a word query will, in the first instance, list all the words that match the string that has been typed in, enabling the user to make a selection from that list (which, incidentally, will also show the frequency of occurrence of any item in that list, when you click on it). This selection may contain more than one word.

The phrase query, on the other hand, does not allow multiple selections. It will only look for instances of the string that has been typed in. There are, however, more differences. Task #2 ends with a specific discussion of these differences. The most important differences are:

1. phrase query may make a query case-sensitive
2. phrase query may search for multi-word strings
3. phrase query enables the user to look for orthographic words which are not treated as L-words in the index. An L-word is an entry in the BNC word index, and may be a multi-word item, such as *in spite of*, or a clitic, such as *n't*. A word query will list all the L-words that match the string typed in. Thus a search on *corpus* will give *corpus*, *corpus-based*, *corpus delicti*, etc.

Another feature that the reader learns about here is that from a list of

solutions to a query certain solutions may be selected and subsequently discarded, if they turn out not to be relevant to the search question.

In task #3 the reader is introduced to the collocation option, and learns about KWIC format. It turns out that possible collocates to a certain word must be typed in, if they are not adjacent. SARA will then tell the user how often the collocate is found within the span of words that the user has defined. SARA does not automatically generate a list of frequent collocates. Adjacent collocates can be found easily in the KWIC format by sorting the examples by word preceding the query focus or following the query focus. Incidentally the query focus itself can also be sorted, which is useful in cases where a multiple word search is carried out.

Task #4 introduces the anyword wildcard character in phrase query, as well as the query builder. The reader learns that the query builder often provides an elegant way of combining two separate queries, as in the case of, for instance, the forms of the indefinite article.

In task #5 the reader is familiarised with the SGML query. This is done on the basis of the question whether people ever actually say things like *you can say that again* or *good heavens*, or whether they only say it in imagined speech. In other words, the spoken data is compared to part of the written data, viz the dialogue parts in imaginative writing. Information of this kind can be found in the so-called headers to each text in the BNC, which contain, among other things, a category reference, with specific attributes, each of which has a specific value, according to how the text has been categorised. This makes it possible to limit any search to texts with specific values for specific attributes in the text headers.

Task #6 provides further practice with text header information on the basis of how men and women use certain words or expressions. The spoken data in the BNC contains information about the speakers that were recorded, such as, age, sex, dialect, social class, level of education, etc. The BNC is useful for sociolinguistic research.

In task #7 the reader is acquainted with the possibility of looking for words in combination with a specific word class tag. The part-of-speech query enables the user to restrict a search to words with specific grammatical roles only. This is also the task that introduces the CQL query, the one that is not menu-driven, but which requires the user to type in the entire query. If a user uses any of the other query types, SARA will ‘translate’ these into CQL. The CQL formulation of the query can always be made visible.

Task #8 teaches the reader how to formulate more complex queries, both in word query mode, using the pattern option, and in phrase query mode. Also, query components can be combined in any order by means of so-called two-way links in the query builder. The query builder also allows the user to restrict the scope of a search by means of the span scope option. Many of these options are demonstrated in the search for variant forms of the expression *to spring a surprise*, in which inflected forms play a role, and variant word orders (eg in passive verb phrases).

The final two tasks give the reader a flavour of even more sophisticated searches, in which non-verbal and non-vocal information can be used (in the spoken part), where text structure features, such as sentences or conversations, can be used to search for specific words in specific positions, and, finally, where an interesting pragmatic function of language, the definition of terms, is investigated. The latter is done on the basis of the question what the acronym SARA means. As the BNC was collected before SARA was developed, the authors argue, it is hardly likely that any reference to SARA in the corpus will be to the SARA in the BNC handbook. Incidentally, this is the first and only place in the book (p 180) where the acronym is explained: SGML Aware Retrieval Application.

After the reader has worked her/his way through Part II s/he has a very good idea of the various possibilities that SARA offers. Part III, then, can be used later as a quick reference guide to the various commands in the program. The reference guide lists them in the order in which they appear on the main menu bar. This will help the user later on to find the various features of any command. It also lists the part-of-speech codes used in the BNC (the CLAWS5 tagset), as well as the text codes, the dialect codes and a few other codes. The authors create some confusion by referring to the CLAWS4 system and the C5 tagset on pp 34–35.

The strong point of this book is that the authors understand what an uninitiated user of SARA needs in order to be able to work with it. The reader is almost literally taken by the hand in her/his attempts to retrieve information from the BNC. There is no point where I felt that this was overdone. New information is given to the reader in neat doses. In consecutive tasks the reader has to carry out similar operations repeatedly, which only helps her/him to master them better, so this is a very useful approach from a pedagogical point of view. For instance,

the reader/user is repeatedly asked to look at lists of solutions to queries, and then thin this list by selecting solutions and discarding these (or the reverse – discarding all but the selected solutions).

The authors also urge the user to specify the search in such a way as to ensure maximum **precision** and maximum **recall**. By the former is meant that a search will yield *only* the relevant solutions; by the latter that the search will give *all* the relevant solutions. In other words, the authors are not content merely to show readers how to use the tool, but they also encourage them to use the tool sensibly. That they do this on the basis of a number of very recognisable linguistic problems is, of course, a big help.

What I particularly liked was the authors' use of italics. In all the sections of the various tasks, italic font was reserved for general advice relating to the particular point in question, and/or for cross-referencing to other sections. I soon found myself going for the bits in italics, as they were the bits that taught me things, rather than instruct me how to proceed with the task.

I also have a few points of criticism. First of all, the book is not illustrated in any way. I would have thought that a couple of screen displays, especially at the beginning of the first task, would have made life a little easier for the reader. On pp 54–56, for instance, where the reader is told how to change the default query display settings, I would have welcomed some kind of visual illustration. Of course, when you are using the program, it does not matter so much, but this only strengthens the idea that this is, in fact, a SARA tutorial. Another thing is that, right in the beginning of task #1, the reader is confronted with the notion of SGML *entity references* (…, p 52), which might be a bit too much for the uninitiated.

Another point of criticism that must be made here is that the authors occasionally point out shortcomings in the program, as on p 141, where they indicate that it is not possible to search for specific sequences of part-of-speech codes (although I am sure that such a provision would be welcomed by many potential users). They do not comment on these shortcomings, nor do they anticipate that these shortcomings will be remedied in the foreseeable future. In other words, the reader is given no idea of how the new versions of the corpus and software, which are announced on p 47, are going to be improved.

My overall evaluation, as I indicated in the beginning, is positive. The authors have managed to shape what is essentially a tutorial in Part II into a pleasantly readable text. They remind us of the limitations of the use of corpus data, especially in cases where queries yield only few solutions. It must not be forgotten that the tasks have been designed especially with a view to providing illustrations of the various possibilities that SARA offers. The authors also show that they understand users' reactions, as when, with some sense of humour, they advise readers to 'drink[ ] large amounts of coffee while waiting for the results' (p 125) or even, go to bed, (p. 177), in cases of very time-consuming searches.

I think that the book will prove very useful for the beginning corpus user, given its tutorial character. Of course, a more advanced corpus user will also find useful information about SARA in the book, but mainly in Part III. For more specific details about the BNC, the reader is referred to Burnard (1995). What I do find a little odd, though, is that this BNC Handbook has been published in a series which claims to 'provide accessible introductions for students coming to empirical linguistics for the first time' (stated on the back cover). Although the various tasks do touch on linguistic research questions, the book is first and foremost a SARA tutorial.

Reference

Burnard, Lou (ed). 1995. *User's reference guide to the British National Corpus*. Oxford: Oxford University Computing Services.

Roger Garside, Geoffrey Leech and **Anthony McEnery** (eds). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. New York: Addison Wesley Longman Ltd, 1997. x + 281 pp. ISBN 0 582 29837–7 (paperback). Reviewed by **Atro Voutilainen**, University of Helsinki.

Corpus Annotation contains 17 chapters by 24 authors. Of the authors, 18 are from Lancaster University; six represent academic and industrial institutions in Great Britain, France and Spain.

The book addresses corpus annotation from three main angles. Chapters

1–6 describe different types of linguistic information that can be added to text corpora. Chapters 7–12 concern methods for automating corpus annotation. Chapters 13–17 are more varied: environments for corpus annotation are described; two practical applications of annotated corpora are outlined; and standardisation efforts and specifiability of linguistic representations are examined.

The main emphasis of *Corpus Annotation* is on work carried out by the UCREL team at Lancaster University. Recent contributions by other leading research teams are also discussed. UCREL is a team whose contributions to corpus annotation extend over two decades. Their original and very successful work on automatic wordclass tagging was described in *The Computational Analysis of English* (1987, edited by Garside, Leech and Sampson). Their stochastic word class tagging suite, called CLAWS1, initiated a worldwide research effort on statistical tagging and parsing. Another important result described in the earlier book was the word class annotation of the one-million-word Lancaster-Oslo/Bergen Corpus, an annotated text collection that has been widely used in linguistic research in many countries.

Indeed the term annotation is ambiguous. It can refer to a **product**: a static (linguistic) representation as attached to text words or as described in annotators' manuals or style sheets. The other reading denotes a **process**: mainly methods and tools for automatically adding annotations to text corpora. It seems to me that the primary goal of the UCREL team has recently been on the product side. Text corpora have been enriched with different types of linguistic information: automatically if possible, else (semi)manually by expert grammarians. Presently only the lowest levels of linguistic information (parts of speech, morphology) can be added automatically with reasonable success, so producing carefully annotated corpora for most levels of linguistic analysis requires considerable expertise and a very substantial human efforts. Though annotation by hand may not be the most inspiring exercise imaginable, the result is of great value to several uses, eg linguistic studies as well as the development and testing of computational language models for NLP systems (witness the wide-spread use of the recently published British National Corpus of about 100 million words).

Chapter 1 (by G. Leech) is a concise introduction to corpus annotation. Relevant terms are explained carefully; the other topics are motivation for annotating corpora, annotation standards, a brief history of annotation, and possible levels of linguistic annotation (orthographic through stylistic annotation).

Chapters 2–6 (by Leech, Eyes, Wilson, Thomas, Garside, Fligelstone, Botley, McEnery and Wynne) address different levels of annotation (morphology, syntax, semantics, discourse, prosody, pragmatics and style). The structure of these chapters is highly similar: for each level of annotation, the design of the linguistic representation is discussed as well as automatisation efforts and developments by other groups. The chapters clearly make the point that when we move to more abstract levels of linguistic structure, specifying the linguistic representation (or annotation scheme) becomes problematic enough; automating the annotation process itself is generally understood within the research community only at the most concrete (ie most obviously structural) levels of analysis (morphology and low-level syntax).

Chapters 7–9 (by Garside, Smith, Fligelstone, Pacey and Rayson) describe recent developments in the famous CLAWS tagging suite that was first introduced in the early 1980s. The core of CLAWS has always been probabilistic: ambiguities are resolved on the basis of lexical and contextual likelihoods derived from previously annotated corpora. As Garside and Smith suggest, purely statistical systems seem to have reached a performance 'plateau', in practice a correctness level of about 96–97 per cent. Recent improvements to the CLAWS tagging suite, based on linguistic (or knowledge-based) techniques, are described in these three chapters: errors typical of the statistical core are better **anticipated** using the extended multiword recognition component, or they are **corrected** with hand-grafted syntactic pattern rules that follow a template rule formalism. The result is CLAWS4, a hybrid tagger capable of analysing texts from more genres than before with an accuracy somewhat higher than what was possible with earlier versions.

Chapter 10 (by Leon and Serrano) describes porting the Xerox HMM tagger to Spanish. The claimed language independence of the Xerox tagger as well as of probabilistic modeling techniques in general is critically examined. The results are not entirely negative: the accuracy of the resulting tagger is close to what is typical of statistical taggers in different languages.

Chapters 11–13 describe methods and tools for semi-automatic corpus annotation. After a look at different approaches to syntactic annotation, Chapter 11 (by Bateman, Forrest and Willis) describes the treebanking environment used by Lancaster and ATR (Japan). Particular attention is given to the linguistic quality of annotation; human treebankers seem to produce the best (or most consistent) output if they can choose the correct analysis from alternatives provided by the machine, instead of

generating one by hand. Chapter 12 (by Garside and Rayson) considers annotation at higher levels of analysis. Two kinds of tool are outlined for these emerging levels of annotation: an editor for discoursal analysis and a semantic tagger. The editor, called XANADU, is a program constantly re(de)fined according to user needs. No high-quality semantic taggers exist so far; Garside and Rayson outline the architecture and types of linguistic information possibly needed by such a tagger (in addition to human postediting). Chapter 13 (by McEnery and Rayson) outlines a more general environment for corpus annotation, editing and exploitation. Also questions related to multifunctionality, modularity and genericity are addressed. The Sara package, distributed with the British National Corpus, is outlined as an example.

Chapters 14 (by McEnery, Baker and Hutchinson) and 15 (by McEnery, Lange, Oakes and Veronis) turn the focus to applications of annotated corpora. Chapter 14 describes interesting experiments with a corpus-based grammar tutor; the experiments suggest that certain linguistic tasks are more efficiently 'taught' by a computer than by a human. Chapter 15 focusses on terminology extraction from annotated multilingual corpora; the resulting multilingual term tuples can be used for translation (by humans or machines). Several techniques for extracting term-level correspondences are described and evaluated. Current techniques seem to work satisfactorily for term pairs consisting of one or two words; especially longer sequences require better solutions, perhaps higher-level annotation for the source corpora.

The last two chapters address fundamental issues in corpus annotation: the possibility of cross-linguistic standardisation (Chapter 16 by Kahrel, Barnett and Leech) and the specifiability of a linguistic representation or annotation scheme (Chapter 17 by Baker). Chapter 16 discusses work carried out by a team of experts in the EAGLES (Expert Advisory Group for Language Engineering Standards) framework towards cross-linguistic annotation guidelines or standards eg for improving the sharability and reusability of natural language resources. Problems facing standardisation efforts (eg acceptability, relevance for parallel work) are discussed. Finally, proposed EAGLES standards for morphosyntactic annotation and syntactic treebanking are briefly presented. Chapter 17 in turn addresses the controversial issue of the specifiability of linguistic coding systems or grammatical representations. In short, the issue is to what degree it is possible to specify the usage principles of linguistic descriptors for enabling linguists to apply those descriptors consistently. Baker reports on a postediting experiment where four trained subjects

corrected a small corpus tagged with the CLAWS tagger. Their outputs were identical to a much higher degree than what seems possible according to certain other parties in the debate.

To summarise: this book describes central aspects of corpus annotation in a readable and informative manner. There is a little overlap between some of the chapters; a little tighter integration might have made the book even more accessible. In any case, many people involved in corpus annotation and use will probably find this book a valuable introduction to problems and the state of the art in corpus annotation. Recommended.