Book Review

The Cambridge Grammar of the English Language

Rodney Huddleston and Geoffrey K. Pullum
(University of Queensland and University of California, Santa Cruz)

Cambridge, England: Cambridge
University Press, 2002, xvii+1842 pp;
hardbound, ISBN 0-521-43146-8, $150.00

Reviewed by
Chris Brew
The Ohio State University

The *Cambridge Grammar of the English Language* is a comprehensive descriptive grammar of English designed to be accessible to the general reader. Part of the declared goal is to produce a grammar that "incorporates insights from the theoretical literature but presents them in a way that is accessible to readers without formal training in linguistics."

Although the grammar is styled as "descriptive" and eschews complex notation, it is clearly the product of authors for whom formal and theoretical issues and arguments are of significant interest and concern. The challenge to which the authors have risen is to present complex materials in a largely nontechnical manner and to do so without sacrificing precision and clarity. In this they have largely succeeded. The dedicated general reader should be able to glean from the book not only a great deal of detailed knowledge about the way in which English works, but also a sense of how the grammar is organized and an implicit understanding of some of the ways in which formal linguists have learned to make and sustain arguments. But although the influence of formalist approaches is evident, there is also an obvious love for and enjoyment of the language itself.

For computational linguists, the grammar may also evoke the strong impression that its authors are implicitly operating with a covert but precise set of assumptions about the formal mechanisms that underlie the grammar, and some may be drawn to the project of making these assumptions explicit, perhaps by using the book as the basis for a large-scale computational grammar, very likely one in the tradition of Pullum's earlier work (Gazdar et al. 1985). Such a project, although very worthwhile, would probably be too long-term for most of us, so we now turn to other ways in which the availability of the grammar may enhance the practice of computational linguistics.

One obvious role is as a guide to English grammar for people who build grammatical artifacts. Such artifacts include not only large-scale computational grammars (Grover, Carroll, and Briscoe 1993; Copestake and Flickinger 2000) but also treebanks (Marcus, Santorini, and Marcinkiewicz 1994). It is idle to speculate on whether the Penn Treebank or the Alvey Natural Language Tools would have been significantly different if the *Cambridge Grammar* had been around to influence them, but it should become a routine part of the training of future grammar writers and treebank annotators that they absorb as much as is feasible of this grammar. This will not be too onerous: Few annotation manuals are as enjoyable to read as this grammar.

Another role for the grammar is as an organized repository for language data. Although corpora were used in the preparation of the grammar, there are also many constructed examples. Indeed, for system builders, the copious collection of negative examples is perhaps the most significant aspect of the publication. They encode significant linguistic intuition that is not otherwise available in a single package.

One immediate concern for the system builder will be to design a suitable mechanism for accessing, decoding, and deploying the information that is in principle available in the grammar. Indeed, it is not even clear to me how the typical human reader will approach this. It is certainly possible to read whole sections or chapters in strict sequence, and when one does this, one begins to appreciate the subtlety and interconnectedness of the analyses. It is also clear that it will be a wonderful resource for teachers and learners of English. Many of the negative examples resemble the kind of thing that non-native speakers produce (e.g., from page 220):

Example 1

a. *It contains of egg and milk.

b. *He bought it to Pat.

It therefore seems likely that the discussion surrounding these examples will be of benefit. But how will readers who want to obtain this benefit know where to look? The grammar provides a lexical index, so one can look up *contain*, but this does not lead to the relevant example (it does, however, lead to the useful comment that some verbs such as *contain, belong, matter*, and *own* have a strong tendency to resist the progressive form even in examples like the following (page 168):

Example 2

At the moment she owns both blocks, but she's selling one next week.

The impact of lexical and syntactic data on computational linguistics is much greater when it is available in electronic form (and some computational linguists read inverted indices for fun). In discussion on Linguist List, responding to a critical review by Joybrato Mukherjee (LINGUIST 13.1853, July 4, 2002), Pullum and Huddleston explain their position on the use of corpus material:

> It is true that for examples we standardly mined the Brown corpus for American English, the London-Oslo-Bergen corpus for British English, and the Australian Corpus of English for Australian English (we had convenient interactive access to these through the courtesy of Macquarie University), and these total three million words. But these corpora were merely sources of illustrative examples, nearly always edited for expository reasons. (It is one of the errors of strictly corpus-oriented grammars to use only raw attested data for purposes of illustration. We think it is counterproductive to quote a sentence with a subject NP containing a long and distracting relative clause when all we are concerned to illustrate is the order of adjuncts in the verb phrase.) (LINGUIST 13.1932, July 17, 2002)

This is an entirely reasonable position, especially if one envisages the main use of a written book as to be read by human readers. The situation changes when one begins (as computational linguists often do) to think of a book as a resource for exploitation

Table 1
The chapter listing.

1.	Preliminaries (GKP and RH)
2.	Syntactic overview (RH)
3.	The verb (RH)
4.	The clause: complements (RH)
5.	Nouns and noun phrases (John Payne and RH)
6.	Adjectives and adverbs (GKP and RH)
7.	Prepositions and preposition phrases (GKP and RH)
8.	The clause: adjuncts (Anita Mittwoch, RH, and Peter Collins)
9.	Negation (GKP and RH)
10.	Clause type and illocutionary force (RH)
11.	Content clauses and reported speech (RH)
12.	Relative clauses and unbounded dependencies (RH, GKP, and Peter Peterson)
13.	Comparative constructions (RH)
14.	Non-finite and verbless clauses (RH)
15.	Coordination and supplementation (RH, John Payne, and Peter Peterson)
16.	Information packaging (Gregory Ward, Betty Birner, and RH)
17.	Deixis and anaphora (Lesley Stirling and RH)
18.	Inflectional morphology and related matters (Frank Palmer, RH, and GKP)
19.	Lexical word-formation (Laurie Bauer and RH)
20.	Punctuation (Geoffrey Nunberg, Ted Briscoe, and RH)

Note: RH = Rodney Huddleston; GKP = Geoffrey K. Pullum

by computer programs. The authors are under no obligation whatsoever to regard their book in this light. But it remains the case that an exhaustive concordance of examples would also be very useful for the language learner. This might take up more space than is warranted in a printed book but would be highly appropriate in an electronic version, where space is no longer at a premium. The publishers of the grammar also are the creators of the *Cambridge International Dictionary of English* (Procter 1995), which is backed by corpora and has spawned a variety of electronic complements, so this is not new territory for them. Because the examples are edited, it may not be easy to identify the source material unambiguously, but when possible it would be especially useful if the concordance contained explicit back references to corpus material. We know that in many cases the source material will be in one of the three corpora named by Huddleston and Pullum. It would be an excellent summer project to apply fuzzy-string-matching technology to generate automatically a first draft of this concordance.

Since the book is a large team effort, it is worth drawing attention to the management structure: Rodney Huddleston acted as the hub, having a hand in the writing of every one of the 20 chapters; Geoffrey Pullum is also named on the front cover, contributing to five chapters, and a team of 13 other linguists gave time and effort. This works really well: readers get the benefit of multiple-author expertise without the pain of gratuitous stylistic variations. As can be seen from Table 1, the term *grammar* is interpreted broadly to include not only sentence syntax, but also lexis, pragmatics, and punctuation. There are no in-text citations, another deliberate design decision to make the grammar more inviting to the general reader, but pointers to further reading are provided.

Everything about this book is a credit to the authors and the publishers. It is authoritative, interesting, reasonably priced (for a book of this size), beautifully designed, well proofread, and enjoyable to handle. Without training, few owners will have the arm strength to read it on the beach or in the bath, but those who can would

probably enjoy the experience. It is both a modern complement to existing descriptive grammars (Quirk et al. 1985; Biber et al. 1999) and an important resource for anyone interested in working with or finding out about English.

In addition, the book is a very complete and convincing demonstration that the ideas of modern theoretical linguistics can be deployed in the detailed description of a particular language. Pedagogically, it might also work as a tool for sneaking the ideas of formal linguistics into the minds of people initially more interested in the language itself. That is a topic for a different review, but this book is as appropriate for the formally trained linguist wishing to broaden the range of data that a theory covers as for the software engineer wishing to augment NLP skills with a more serious understanding of how the language works.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson ESL, Harlow, England.

Copestake, Ann and Dan Flickinger. 2000. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the Second Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA.

Grover, Claire, John Carroll, and Ted Briscoe. 1993. The Alvey natural language tools grammar. Technical report, Human Communication Research Centre, University of Edinburgh and Computer Laboratory, Cambridge University.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Procter, Paul, editor. 1995. *Cambridge International Dictionary of English*. Cambridge University Press, Cambridge, England.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Chris Brew is an assistant professor of computational linguistics and language technology at the Ohio State University. His recent research has concerned the use of corpus-based methods in speech synthesis and in lexical semantics. Brew's address is: Department of Linguistics, The Ohio State University, Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210; e-mail: cbrew@ling.ohio-state.edu.