Corpus Linguistics and the Automatic Analysis of English

Nelleke Oostdijk
(University of Nijmegen)

Amsterdam: Editions Rodopi
(Language and Computers: Studies in
Practical Linguistics 6, edited by Jan
Aarts and Willem Meijs), 1991,
xii + 267 pp.
Paperbound, ISBN 90-5183-281-8,
$40.00, Dfl 80.00

Reviewed by
Ted Briscoe
University of Cambridge

In a recent paper advocating a corpus-based and probabilistic approach to grammar development, Black, Lafferty, and Roukos (1992) argue that "the current state of the art is far from being able to produce a robust parser of general English" and advocate "steady and quantifiable," empirically corpus-driven grammar development and testing. Black et al. are addressing a community in which armchair introspection has been and still is the dominant methodology in many quarters, but in some parts of Europe, corpus linguistics never died. For nearly two decades, the Nijmegen group led by Jan Aarts have been undertaking corpus analyses that, although motivated primarily by the desire to study language variation using corpus data, are particularly relevant to the issue of broad-coverage grammar development. In distinction to other groups undertaking corpus-based work (e.g., Garside, Leech, and Sampson 1987), the Nijmegen group has consistently adopted the position that it is possible and desirable to develop a formal, generative grammar that characterizes the syntactic properties of a given corpus and can be used to assign appropriate analyses to each of its sentences.

Nelleke Oostdijk's book provides a detailed description of the cumulative development of a grammar capable of analyzing a one million–word corpus of English written texts, drawn from a wide but balanced variety of sources. This task forms a significant component of the wider Tools for Syntactic Corpus Analysis (TOSCA) project being undertaken at Nijmegen. Oostdijk's work provides an excellent example of the strengths and weaknesses of the approach advocated by Black et al. In addition, she discusses issues such as sampling and tokenization of corpus material, as well as the exploitation of the analyzed corpus in studies of language variation. However, in this review I will concentrate on the central core of her book: the development of the grammar and performance of the associated parser, since this is the part that is most relevant to computational linguistics.

Oostdijk begins by locating her work and the TOSCA project within the field of computational linguistics (arguing that it is distinguished by "an interest in language itself as it is actually produced" (p. 2)) and contrasting it to the LSP system (Sager 1981) and Parsifal (Marcus 1980). The comparison is brief and the choice odd since more general broad-coverage grammars, such as DIAGRAM (Robinson 1982), PEG (Jensen et al. 1986) and ANLT (Grover et al. 1989), and more corpus-oriented parsing systems, such as FIDDITCH (Hindle 1983, 1993) or MITFP (de Marcken 1990), have been developed within the field, but are not discussed anywhere. A similar suspicion of isolationism recurs in the sections dealing with the grammatical formalism used;

this is based on (extended) affix grammar (Koster 1971) and, although only described informally, the variant of affix grammar adopted is probably similar in generative and expressive capacity to unification-based formalisms, such as PATR-II (Shieber 1986) or the ANLT formalism (Briscoe et al. 1987), with some interesting extensions making it more adequate to phenomena such as agreement in coordinate structures. Unfortunately, no comparison is offered. More discussion is devoted to comparison with the approach to corpus analysis taken by the Lancaster group (Garside et al. 1987); Oostdijk argues that because their espousal of probabilistic methods and rejection of a rule-based generative approach is not founded on sound empirical evidence, it is impossible to develop a comprehensive generative grammar for a corpus. While I am sympathetic to Oostdijk's position and think that the grammar she goes on to present is impressive enough to bias us towards the opposite conclusion, it is a mistake to accept the assumption that the two approaches are incompatible, as much recent work (including that of Black et al. 1992) has demonstrated the usefulness of combining statistical techniques with rule-based systems.

The core of the book is a description of the grammar developed and analyses adopted for notoriously difficult phenomena, such as nonconstituent coordination, gapping, apposition, partitives, other noun phrase premodifier syntax, and so forth. The grammatical framework adopted is based on a conventional notion of constituency, with nodes assigned categorial labels augmented with functional categories encoding mostly familiar grammatical relations. The commitment to nonelliptical accounts of the full range of coordinate and gapped constructions that occur in the corpus leads to adoption of linguistically nonstandard analyses; for example, grouping noun phrase complements of ditransitive verbs into single constituents. Once again, no reference is made to recent theoretical work addressing similar problems, such as extended categorial or combinatory grammar (e.g., Steedman 1985). Nevertheless, the coverage of the resultant grammar is impressive, and the (computational) linguist who has not developed a substantial grammar from natural data will find enough interesting insights, analyses, and detailed discussion of constructions sometimes ignored in the more mainstream generative literature to be convinced, I hope, of the value of corpus-based grammar development. There are, however, dangers, as well as strengths in this approach; for instance, the commitment to assign an analysis to each sentence of the corpus can easily lead to reification of undesirable decisions in the grammar and consequent propagation throughout the analyzed corpus: a case in point might be the use of ditransitive complement constituents introduced to deal with gapped examples.

Corpus-based development and testing of a grammar requires computational support to be practical and, given the goal of the TOSCA project, a method is needed to select the semantically and pragmatically appropriate analysis from the set licensed by the grammar for each sentence in the corpus. A separate system is used to assign each word of the input sentence an unambiguous and correct lexical category compatible with the grammar developed. This system and the lexical categories are not described in the book but appear to be more fine-grained than the categories assigned by tagging programs (e.g., CLAWS2, Garside et al. 1987), incorporating subcategorization information concerning complementation, for instance. The parsing system then assigns analyses to this unambiguous sequence of lexical categories. Oostdijk does not describe the parser-generator or parser developed for the affix grammar formalism used, but instead concentrates on the issues of parse selection and performance both in terms of coverage and efficiency. Parse selection is done interactively by guiding the parser manually; Oostdijk justifies this approach by arguing that it ensures a high level of accuracy and guarantees parsing efficiency by pre-empting unnecessary

search. An approach in which intervention is limited to selection between predefined legitimate analyses is an improvement on one in which the analyst is able to create new descriptions at will (e.g., Leech and Garside 1991) in that the resulting database of analyses will be consistent and intervention will be simpler and faster. However, other approaches are possible, such as the use of probabilities to guide parse selection, if not grammar induction (e.g., Black et al. 1992). Oostdijk does not consider this possibility, presumably because of her acceptance of the incompatibility of rule-based and statistical techniques. The decision to manually select parses, coupled with the fact that the TOSCA parser (on the hardware available) is not always able to compute all the possible analyses, even starting from unambiguous lexical categories, has the unfortunate side effect that a significant effort has been devoted to removing linguistically motivated ambiguity from the grammar. Earlier, Oostdijk argues for the strict separation of grammatical formalism and parsing algorithm on familiar grounds, but the same arguments tell against the decision, for instance, to stipulate that coordination occurs at specific nodes in case of ambiguity (p. 133), since such distinctions often correlate with differences of semantic scope.

Despite these criticisms—and in a practical project of this type some compromises are inevitable—Oostdijk's achievement is impressive; her book is well written and easy to read, and she manages the difficult task of striking the right level between an exhaustive and exhausting documentation of a substantial grammar and a superficial overview. There are very few typographical errors and the book has been well produced.

References

Black, Ezra; Lafferty, John; and Roukos, Salim (1992). "Development and evaluation of a broad-coverage probabilistic grammar of English-language computer manuals." *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, Newark, DE, 185–192.

Briscoe, Ted; Grover, Claire; Boguraev, Bran; and Carroll, John (1987). "A formalism and environment for the development of a large grammar of English." *Proceedings, 10th International Joint Conference on Artificial Intelligence*, Milan, 703–708.

De Marcken, Carl G. (1990). "Parsing the LOB Corpus." *Proceedings, 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, 243–251.

Garside, Roger; Leech, Geoffrey; and Sampson, Geoffrey (1987). *The Computational Analysis of English: A Corpus-Based Approach*. Longman.

Grover, Claire; Briscoe, Ted; Carroll, John; and Boguraev, Bran (1989). "The Alvey natural language tools grammar (second release)." Technical Report 162, University of Cambridge, Computer Laboratory.

Hindle, Donald (1983). "User manual for Fidditch, a deterministic parser." Technical Memorandum 7590-142, Naval Research Laboratory.

Hindle, Donald (1993). "A parser for text corpora." In *Computational Approaches to the Lexicon*, edited by B. T. S. Atkins and A. Zampolli, Oxford University Press. In press.

Jensen, Karen; Heidorn, George; Richardson, Stephen; and Haas, Norman (1986). "PLNLP, PEG and CRITIQUE: Three contributions to computing in the humanities." Report RC-11841, IBM Thomas J. Watson Research Center.

Koster, C. (1971). "Affix grammars." In *ALGOL 68 Implementation*, edited by John Peck, North-Holland.

Leech, Geoffrey, and Garside, Roger (1991). "Running a grammar factory: The production of syntactically analysed corpora or 'treebanks'." In *English Computer Corpora: Selected Papers and Bibliography*, edited by Stig Johansson and A. Stenstrom, Mouton de Gruyter.

Marcus, Mitchell P. (1980). *A Theory of Syntactic Recognition for Natural Language*. The MIT Press.

Robinson, Joan (1982). "DIAGRAM: A grammar for dialogues." *Communications of the ACM*. **25**(1), 27–47.

Sager, Naomi (1981). *Natural Language Information Processing: A Computer Grammar of English and Its Applications*. Addison-Wesley.

Shieber, Stuart (1986). *Introduction to Unification-Based Approaches to Grammar.* Stanford, CA: Center for the Study of Language and Information.

Steedman, Mark (1985). "Dependency and coordination in the grammar of Dutch and English." *Language*, **62**, 523–568.

Ted Briscoe is a SERC Advanced Research Fellow at the Computer Laboratory, University of Cambridge. His current research interests include robust parsing of naturally occurring natural language. His address is: Computer Laboratory, University of Cambridge, Pembroke Street, Cambridge CB1 3AZ, UK; e-mail: ejb@cl.cam.ac.uk