name—see Schützenberger (1956). For a summary of the issues, see Ryckman (1986), chap. 5.

[2] Carnap and Bar-Hillel (1952), Bar-Hillel (1952). The present book seems in part responsive to this program, having the same title as Bar-Hillel (1964).

[3] See papers collected in Hintikka and Suppes (1970).

[4] Dretske (1981), Israel and Perry (forthcoming). Peer commentary in Dretske (1983), especially that of Haber, did not accept Dretske's attempted analogies to the metrics of Shannon and Weaver. The notion of "information pickup" implies a pre-established harmony of the world and the mind, disregarding the well-known arbitrariness of language.

[5] While Fodor (1986) does gives a cogent criticism of attempts to locate information "in the world", the alternative "intentional" conception that he advocates relies on questionable assumptions of an "internal code" wherein such information is "encoded". The problem, of course, lies in unpacking this metaphor. Falling into the custom of taking the computational metaphor of mind literally, he resuscitates our old familiar homunculus (in computational disguise as the "executive") to provide a way out of the problem of node labels being of higher logical type than the nodes that they label. A simpler resolution follows from Harris's recognition that natural language has no separate metalanguage. See also Fodor (forthcoming).

[6] See especially Harris (1982), and Harris, Gottfried, Ryckman, et al. (in press).

[7] This thus cuts deeper than the naive rule-counting metrics for adjudication of grammars advocated not so long ago by generativists (see Ryckman 1986).

[8] This work is reported in depth in Harris et al. (in press). These science languages occupy a place between natural language and mathematics, the chief difference from the former being that operator-argument likelihoods are much more strongly defined, amounting in most cases to simple binary selection rather than a graded scale. One of the many interesting aspects of this research is determining empirically the form of argumentation in science. The logical apparatus of deduction and other forms of inference are required only for various uses to which language may be put, rather than being the semantic basis for natural language, as has sometimes been claimed.

[9] This is a refinement of the notion of distributional meaning developed in, e.g., Harris (1954).

[10] The case of zero likelihood is covered by the word classes of the first constraint.

[11] An example is the elision of one of a small set of operators including *appear, arrive, show up*, which have high likelihood under *expect*, in *I expect John momentarily*. The adverb *momentarily* can only modify the elided *to arrive*, etc., since neither *expect* nor *John* is asserted to be momentary. The infinitive *to*, the suffix *-ly* , and the status of *momentarily* as a modifier are the results of other reductions that are described in detail in Harris (1982).

[12] For a computer implementation, see Johnson (1987). I am grateful to Tom Ryckman for helpful comments on an early draft of this review.

The Computational Analysis of English: A Corpus-based Approach

Roger Garside, Geoffrey Leech, and Geoffrey Sampson, (eds.)
(University of Lancaster and University of Leeds)

London: Longman, 1987, xii+196 pp.
ISBN 0-582-29149-6; (sb)

Reviewed by
Michael Lesk
Bell Communications Research

Why is it so remarkable to have a book whose analysis of language is entirely based on actual writing? Professors Garside, Leech, and Sampson have the refreshing view that the analysis of language ought to be based on real language, and have presented 12 papers resulting from their studies using the Lancaster-Oslo-Bergen corpus of a million words of British English. They present studies of spelling correction, part-of-speech assignment, parsing, and speech synthesis based on probability techniques derived from corpus studies. The methods here work on arbitrary texts and with reasonable efficiency.

English includes a great variety of constructions that pose a dilemma for any strict grammar: to include everything and face great ambiguity, or to be extremely prescriptive and reject much. The authors solve this problem by using probabilities to balance both frequent and infrequent constructions, and to emphasize low-level simple algorithms over deep interpretation.

For anyone trying to make practical use of text, this book is extremely enlightening. English is not an inferior substitute for Prolog, and treating it as such is not only a mismatch, but also unnecessary for many tasks. The simple use of probabilities can perform many tasks that at first glance might be thought to require understanding. Methods for doing these are explained clearly in the book.

The most detailed result described is the technique of tagging, or assigning parts of speech statistically. By using both the individual probabilities of different parts of speech for a single word, and the combined probability of sequences of two parts of speech, tagging can be done with 96–97% accuracy. This relatively simple algorithm, relying for performance on statistical data accumulated over a large sample of English rather than upon some kind of model of language, is typical of the results presented in this book. The algorithm runs on any input, from any subject area, and does a useful job without claiming to "understand" natural language. Just as we have learned that computers can play master-level chess by exhaustive evaluation of all possible moves, without any grand strategy or even plausible move selection, it seems that many linguistic tasks do not require understanding or modeling, but merely experience, translated into probability data.

Similar discussions apply to parsing. Fifty thousand words of the corpus have been parsed by hand, and this has been used to make a table of the relative frequencies of different syntactic constructions. Assuming that the correct parse tree is the one made of the most probable constituents (to greatly oversimplify in the interests of saving space), a program was written to parse with about 50% accuracy. Since the preparation of this book, continuing work by Eric Atwell and Geoffrey Sampson at Leeds has greatly improved on this figure, using a simulated annealing technique (see Sampson 1986).

Other chapters of the book discuss the history of corpora in linguistic research, a defense of probabilistic methods, a discussion of speech synthesis and an outline of a sophisticated spelling corrector. Not much has been done on speech synthesis, partly because we do not as yet have good data on the relation between syntax and prosody. The spelling corrector is aimed at errors of word selection, i.e., finding words that, although they appear in the dictionary, should not appear in the particular sentence being studied (e.g., "They kingdom come, thy will be done"). All these tools follow the same model: reliance on statistics from the corpus.

It is a great relief to read a book like this, which is based on real texts rather than upon the imaginary language, sharing a few word forms with English, that is studied at MIT and some other research institutes (see Postal 1988). It is amazing that computers, which are distinguished for their ability to deal with vast quantities of bytes and their incompetence with even simple patterns and models, have been used in linguistics primarily for the implementation of complex logical models. This book is a start on the exploitation of large database methods for linguistic information. It is remarkable for the performance of its methods combined with their simplicity. Unlike many books on linguistics, it is easy to understand; it makes one think of the Molière character who suddenly found out he had been speaking prose all his life.

I heartily recommend this book to anyone who wishes to process language for a useful purpose. Other workers such as John Sinclair (1987) and Yaacov Choueka (1988) have also used large text databases for deriving linguistic information. When I was an undergraduate, one of my professors said that "mathematical intuition means having seen the problem before." Similarly, there is no substitute in linguistics for knowing that a particular construction is likely because it has appeared many times. This book is a testimony to the superiority of experience over fantasy.

References

Choueka, Yaacov 1988 Looking for Needles in a Haystack. In *Proceedings of the RIAO 88*, 609–623.

Postal, Paul 1988 Advances in Linguistic Rhetoric. In *Natural Language and Linguistic Theory* 6:129–137.

Sampson, Geoffrey 1986 Simulated Annealing as a Parsing Technique. In *University of Leeds Working Papers in Linguistics and Phonetics* 4:43–60.

Sinclair, John 1987 *Looking up*. Collins, London, England; Glasgow, Scotland.

Michael Lesk is division manager of computer science research at Bell Communications Research, 445 South St., Morristown, NJ 07960. He uses machine-readable dictionaries in his research on text handling and retrieval. E-mail: lesk@wind.bellcore.com

---

Semantic Interpretation and the Resolution of Ambiguity

Graeme Hirst
(University of Toronto)

Cambridge, England: Cambridge University Press, 1987, xiv+263 pp.
ISBN 0-521-32203-0; (hb) $39.50 [20% discount to ACL members]

Reviewed by
Karen Sparck Jones
University of Cambridge

Hirst's book presents an approach to natural language interpretation, using as his vehicle a description of the experimental system he built. It therefore has to be evaluated as a contribution on how to build NLP systems from both theoretical and practical points of view. It also has to be considered for teaching purposes, since Hirst has vamped up what was originally a thesis with some pedagogic exposition and test exercises, as well as a substantial and useful bibliography.

Hirst is very clear about his aims and very honest about what he has tackled. He presents detail well and provides excellent summaries, so the essential properties of his work are well laid out.

His goal was to build an interpretation system that could handle serious lexical and structural ambiguity, and handle it in a principled way. His concern is thus essentially computational; he does not make any claims for the psycholinguistic relevance of what he is doing, but he is, on the other hand, willing to exploit psycholinguistically derived support for good processing strategies.

The system consists of a syntactic parser, **Paragram,** a semantic interpreter, **Absity,** and two disambiguation processors: the **Polaroid Word** (PW) subsystem for lexical disambiguation and the **Semantic Enquiry Desk** for structural disambiguation. The system builds an explicit meaning representation in the frame language **Frail.**

Hirst's design is motivated by two goals: to allow processes of different sorts to use different kinds of information but to interact to construct a sentence representation; and to do this in the theoretically well-founded way exemplified by Montague's work by doing