## Advances in Probabilistic and Other Parsing Technologies

**Harry Bunt and Anton Nijholt (editors)**
(Tilburg University and University of Twente)

Dordrecht: Kluwer Academic
Publishers (Text, speech and language
technology series, edited by Nancy Ide
and Jean Véronis, volume 16), 2000,
xv+267 pp; hardbound, ISBN
0-7923-6616-6, $112.00, £71.00,
Dfl 230.00

*Reviewed by*
*Chris Brew*
*The Ohio State University*

This book is an edited selection of papers presented at the Fifth International Workshop on Parsing Technologies, held at MIT in September 1997. Several of the papers are already well-known and others should be. The book could easily be used as the basis for a graduate-level advanced course on parsing. The title is unwieldy, but appropriate: most but not all of the papers have a strong probabilistic flavor.

My favorite papers are Erik Hektoen on "Probabilistic parse selection based on semantic co-occurrences," Jason Eisner on "Bilexical grammars and their cubic-time parsing algorithms," and Chris Manning and Bob Carpenter on "Probabilistic parsing using left corner language models." I like these papers because they step back from the details of parsing technology and consider its wider significance.

Manning and Carpenter offer both detail and overview. They provide a series of probabilistic models that relax the context-freeness assumption of probabilistic context-free grammars, measure performance in the usual way, draw appropriate conclusions, then provide the kicker in the form of a brief section explaining "Why parsing the Penn Treebank is easy." As Manning and Carpenter point out, in the particular case of the Penn Treebank, the currently accepted PARSEVAL metrics (Grishman, Macleod, and Sterling 1992) are actually quite easy to do well on, even if the system makes systematic errors on such things as prepositional-phrase attachment. If systems are to be deployed into situations where such deficiencies might matter, it might be necessary to find more appropriate evaluation methods. This issue has subsequently been addressed by others (Carroll, Briscoe, and Sanfilippo 1998; Carroll, Minnen, and Briscoe 1999), who argue for more obviously task-related evaluation schemes involving predicate argument structure and/or dependency information.

Hektoen's contribution is in the same vein; it takes seriously the notion that parsing is often simply a device for getting at an underlying semantics. Under his scheme, parse selection relies on the ability to collect statistics over semantic forms. Following this path leads Hektoen into a careful exposition of a *Bayesian-estimation* approach to parse selection, which appears to be "a sufficient response to the high degree of sparseness in the lexical co-occurrence data without the blurring associated with smoothing and clustering" (p. 162). Hektoen's approach appears to work well; of course, it does require a broad-coverage parser capable of generating semantic representations, which may be an obstacle for many. The exposition of the method is very clear and the comparison with previous approaches is enlightening.

Mark-Jan Nederhof's "Regular approximation of CFLs: A grammatical view" is similar to Eisner's contribution in that its focus is primarily mathematical. It describes an attractive approach to finite-state approximation of regular grammars. The essential idea is to characterize properties that make grammars non-regular, and to develop schemes for systematically removing such properties. This helps to keep the approximation process perspicuous. Experimental work with this approximation scheme is absent from the current article, but is reported elsewhere (Nederhof 2000).

In "Probabilistic GLR parsing," Kentaro Inui, Virach Sornlertlamvanich, Hozumi Tanaka, and Takenobu Tokunaga provide a careful analysis of the process of LR parsing. This leads to a probabilistic parsing scheme having the desirable property, not previously achieved for LR parsers, that the sum over all parses of the probability is unity. Once again experimental work is not present here but is reported elsewhere (Sornlertlamvanich, Inui, Tokunaga, Tanaku, and Takezawa 1999).

Eisner's paper does not report experiments either, but addresses a problem with profound practical significance. It analyses the computational properties of grammars in which potentially idiosyncratic word-to-word relationships play a key role. The framework used is general enough to capture the essence of many recent statistical parsers and clean enough to make it easy (and interesting) to compare one with another. I like Eisner's paper for the insight it provides into the options available to the lexically minded probabilistic modeler. This aspect is also present in "Encoding frequency information in lexicalized grammars," where John Carroll and David Weir, using lexicalized tree adjoining grammar (LTAG) as an example, analyze the problem of providing practically useful estimates of the large number of parameters that are potentially present in lexicalized grammars. Similarly, in "Towards a reduced commitment, D-theory style TAG parser," John Chen and K. Vijay-Shanker describe an approach to TAG parsing whose goal is to delay attachment decisions. This is a design sketch, not an implemented parser, but the design is well fleshed out, and looks worth testing.

Several articles do have extensive evaluation data. Joshua Goodman contributes "Probabilistic feature grammars," developing an implemented and efficient stochastic feature-based grammar formalism. The key idea, prefigured in, for example, Stolcke's (1994) doctoral dissertation, is to choose a feature formalism that does not impede dynamic programming implementations of the usual inside, outside, and Viterbi probability calculations. Goodman includes extensive quantitative evaluation, which is greatly to be welcomed. "A new parsing method using a global association table" by Juntae Yoon, Seonho Kim, and Mansuk Song, is a description and evaluation of a semi-deterministic parsing algorithm designed to exploit the fact that Korean is an SOV language with many surface cues to syntactic dependency. Extensive evaluation is provided. Bangalore Srinavas's "Performance evaluation of SuperTagging for partial parsing" exploits the author's SuperTagging idea (i.e., employing part-of-speech–tagger technology to "almost parse," using the elementary trees of lexicalized tree adjoining grammar) for the now-standard task of partial parsing. Given the title, the plethora of interesting performance figures is to be expected. For example, connecting to the discussion of the Penn Treebank above, Bangalore reports that 35% of the sentences tested have no dependency-link errors, while 89.8% have three errors or less.

Two papers give evaluations that are based on the measurement of run-time behavior. In "Parsing by successive approximation," Helmut Schmid describes an efficient parsing technology that is nonetheless able to process grammars that make significant use of features. The efficiency of this algorithm is demonstrated by appeal to a range of empirical performance statistics. Udo Hahn, Norbert Bröker, and Peter Neuhaus

take a similar approach to evaluation. Their contribution describes "Message-passing protocols for object-oriented parsing," and shows how to derive different heuristically guided parsing algorithms from variations in the communication patterns in an object-oriented parser. They report a variety of performance statistics for a set of 41 challenging-looking sentences from German computer magazines.

Since a version of the material of the book has already been presented at a workshop with proceedings (Bunt and Nijholt 1997), it is relevant to ask what has been gained (or lost) in the transition to (an expensive) book form. The articles average 20 pages—longer than the original conference presentation—and several authors have made good use of the opportunity to update and revise their work. The editors have selected an interesting group of papers, and provide a clear introduction with useful summaries of the chapters, pointing out some interesting relationships between the different lines of research.[1] On the other hand, despite the high price of the book, there is no evidence that a competent professional copy editor was involved in the process of publication. This is a shame, since several of the contributions (especially Hektoen's) deserve to be more widely known.

**References**


Bunt, Harry and Anton Nijholt. 1997. *Proceedings of the Fifth International Workshop on Parsing Technologies*. Massachusetts Institute of Technology, Boston, MA.

Carroll, John, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser evaluation: a survey and a new proposal. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 447–454, Granada, Spain.

Carroll, John, Guido Minnen, and Ted Briscoe. 1999. Corpus annotation for parser evaluation. In *Proceedings of the EACL-99 Post-Conference Workshop on Linguistically Interpreted Corpora (LINC-99)*, pages 35–41, Bergen, Norway.

Grishman, Ralph, Catherine Macleod, and J. Sterling. 1992. Evaluating parsing strategies using standardized parse files. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 156–161, Trento, Italy.

Nederhof, Mark-Jan. 2000. Practical experiments with regular approximation of context-free languages. *Computational Linguistics*, 26(1): 17–44, March.

Sornlertlamvanich, Virach, Kentaro Inui, Takenobu Tokunaga, Hozumi Tanaka, and Toshiyuki Takezawa. 1999. Empirical support for new probabilistic generalized LR parsing. *Journal of Natural Language Processing*, 6(2): 3–22.

Stolcke, Andreas. 1994. *Bayesian Learning of Probabilistic Language Models*. Ph.D. thesis, University of California at Berkeley.



*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 psycholinguistics and in natural language generation. Brew's address is: Department of Linguistics, Oxley Hall, 1712 Neil Avenue, Columbus, OH 43210; e-mail: cbrew@ling.ohio-state.edu.


1 In some cases, Bunt and Nijholt seem to be going out of their way to convince themselves that essentially symbolic work is founded on a probabilistic approach. The papers by Chen and Vijay-Shanker, and by Hahn, Bröker and Neuhaus, in spite of the editorial claim that they fall under "the development of strategies for efficient probabilistic parsing", do not go into detail on this issue.