Studies in Natural Language Processing

Natural language parsing

Psychological, computational, and theoretical perspectives

Edited by
David R. Dowty, Lauri Karttunen,
and Arnold M. Zwicky

Natural language parsing

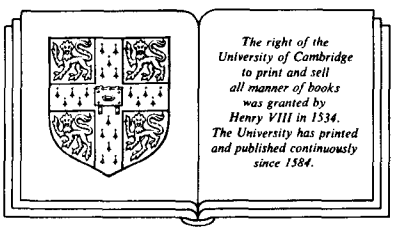
Psychological, computational, and theoretical perspectives

Edited by

DAVID R. DOWTY
Department of Linguistics, Ohio State University

LAURI KARTTUNEN
Artificial Intelligence Center,
SRI International, Menlo Park

ARNOLD M. ZWICKY
Department of Linguistics, Ohio State University



CAMBRIDGE UNIVERSITY PRESS
CAMBRIDGE
LONDON NEW YORK NEW ROCHELLE
MELBOURNE SYDNEY

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
32 East 57th Street, New York, NY 10022, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1985

Printed in the United States of America

Library of Congress Cataloging in Publication Data

Main entry under title:

Natural language parsing.

(Studies in natural language processing)

Includes index.

1. Parsing (Computer grammar) I. Dowty, David R.
II. Karttunen, Lauri. III. Zwicky, Arnold M. IV. Series.
P98.N3 1985 415 84–4360
ISBN 0 521 26203 8

Studies in Natural Language Processing

This new series will publish monographs, texts, and edited volumes within the interdisciplinary field of computational linguistics. Sponsored by the Association for Computational Linguistics, the series will represent the range of topics of concern to the scholars working in this increasingly important field, whether their background is in formal linguistics, psycholinguistics, cognitive psychology, or artificial intelligence. *Natural language parsing* is the first volume to appear.

Contributors

Greg N. Carlson, Department of English, Wayne State University
Stephen Crain, Department of Linguistics, University of Connecticut
Alice Davison, Department of Linguistics, University of Illinois, Urbana-Champaign
David R. Dowty, Department of Linguistics, Ohio State University
Elisabet Engdahl, Department of Linguistics and Phonetics, Lund University
Janet Dean Fodor, Department of Linguistics, University of Connecticut
Lyn Frazier, Department of Linguistics, University of Massachusetts
Aravind K. Joshi, Department of Computer and Information Science, University of Pennsylvania
Lauri Karttunen, Artificial Intelligence Center, SRI International, Menlo Park
Martin Kay, Xerox Palo Alto Research Center
Richard Lutz, Department of Linguistics, University of Illinois, Urbana-Champaign
Fernando C. N. Pereira, Artificial Intelligence Center, SRI International, Menlo Park
Mark S. Seidenberg, Department of Psychology, McGill University
Mark Steedman, School of Epistemics, University of Edinburgh
Michael K. Tanenhaus, Department of Psychology, University of Rochester
Arnold M. Zwicky, Department of Linguistics, Ohio State University

Acknowledgments

Earlier versions of the papers by Davison and Lutz, Engdahl, Crain and Fodor, Frazier, Joshi ("Tree Adjoining Grammars"), Karttunen and Kay, Crain and Steedman, and by Tanenhaus, Carlson, and Seidenberg were originally presented at a conference, *Syntactic Theory and How People Parse Sentences,* held at the Ohio State University on May 14–16, 1982. This conference was supported by the Linguistics Department of that university with assistance from the Ohio State University Graduate School.

Two papers, by Joshi ("Processing of Sentences with Intrasentential Code Switching") and by Kay, were originally presented at a conference on parsing held at the Cognitive Science Center of the University of Texas in 1981, supported by a grant from the Alfred P. Sloan Foundation.

The editors would like to thank the following referees for their advice regarding revisions of these papers for publication: Janet Fodor, Stephen Isard, William Marsh, Ellen Prince, Geoffrey Pullum, Jane Robinson, Kenneth Ross, Richmond Thomason, and Ellen Woolford. Others who provided suggestions directly to the authors are acknowledged by the authors in their individual papers.

The final preparation of the manuscript was supported by a grant from Ohio State University College of Humanities.

Contents

Introduction

LAURI KARTTUNEN and ARNOLD M. ZWICKY

1. Parsing in traditional grammar

Like so many aspects of modern intellectual frameworks, the idea of parsing has its roots in the Classical tradition; (*grammatical*) *analysis* is the Greek-derived term, *parsing* (from *pars orationis* 'part of speech') the Latin-derived one. In this tradition, which extends through medieval to modern times,

(1) parsing is an operation that human beings perform,

(2) on bits of natural language (usually sentences, and usually in written form),

(3) resulting in a description of those bits, this description being itself a linguistic discourse (composed of sentences in some natural language, its ordinary vocabulary augmented by technical terms);

(4) moreover, the ability to perform this operation is a skill,

(5) acquired through specific training or explicit practice, and not possessed by everyone in a society or to equal degrees by those who do possess it,

(6) and this skill is used with conscious awareness that it is being used.

Parsing, in the traditional sense, is what happens when a student takes the words of a Latin sentence one by one, assigns each to a part of speech, specifies its grammatical categories, and lists the grammatical relations between words (identifying subject and various types of object for a verb, specifying the word with which some other word agrees, and so on). Parsing has a very practical function:

> It is not generally realized, even in the schools, how difficult it is for anyone to control the expression and interpretation of language, and that control is as difficult to teach as it is to achieve. The traditional means of teaching control, to pupils at all levels, in their own language as well as in foreign languages, is the set of analytical procedures called grammar.
>
> (Michael 1970:1)

In other words,

(7) the reason for a discipline of parsing is to increase one's mastery over expression in language.

Another important part of the tradition is a separation between grammar and logic. Parsing is analysis for the purposes of grammar; quite a different sort of analysis is appropriate in the study of argument. Although the distinction between grammatical form and logical form has been drawn in a number of ways, not always clearly, it plays a role in linguistic discussions from Aristotle through Port Royal to Chomsky. Here we must stress the fact that, in its traditional sense, parsing is in no way an extraction of properties and relations that are of direct *semantic* relevance. In rather modern phrasing,

(8) the descriptions in (3) are grammatical in nature; that is to say, they describe facts relevant to the co-occurrence of and alternation between units in a particular language.

Note that (8) does not specify any particular theory of grammar; one can parse sentences with respect to any given theory. Indeed, much of the history of parsing until a few decades ago can be understood as the direct consequence of the history of (partial) theories of grammar. Changes in the list of parts of speech, in the list of grammatical categories, or in the list of grammatical relations carry with them changes in what has to be said in parsing a sentence.

We now summarize these eight characteristics of parsing in the Western grammatical tradition. Characteristic (1) says that parsing is done by human beings, rather than by physical machines or abstract machines. Characteristic (2) specifies that what is parsed is a bit of natural language, rather than a bit of some languagelike symbolic system. Characteristic (3) specifies that the analysis itself is a bit of natural language, rather than a bit of some languagelike system, and characteristic (8) that the analysis concerns grammatical rather than logical properties. Characteristic (4) tells us that parsing is heuristic rather than algorithmic, characteristic (5) that it is learned by certain people and not "given" within a society. According to characteristic (6), parsing is overt rather than covert. Characteristic (7), finally, says that the function of parsing is pedagogical.

2. New notions of parsing

In this century the word *parsing* has come to be extended to a large collection of operations that are analogous in some ways to the traditional one just described, but differ from it in one – or usually more – of the eight characteristics. These changes result from a series of new conceptualizations, partially independent of and partially interconnected with one another, in theoretical linguistics, formal language theory, computer science, artificial intelligence, and psycholinguistics. Although the historical roots of these ideas are in some cases fairly deep, they flower together only about the middle of this century, in the 1950s and early 1960s.

3. Parsing in formal linguistics

In linguistics the first of these changes was to view the rationale for parsing not as pedagogical, but rather as scientific – in other words, to emphasize the descriptive, rather than the prescriptive, side of characteristic (7). This shift in emphasis was largely the work of structuralist linguistics, and in its train came a significant change in characteristic (3), as a concern for precision in grammatical descriptions led increasingly to their formalization. The end of this movement away from the informal and discursive descriptions of traditional grammar was a view of these descriptions as completely formal objects – in particular, as *constituent structures* (assigning words to parts of speech and describing which adjacent constituents can combine with one another, in what order they combine, and what phrase category the combination belongs to).

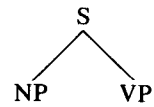
This particular formalization of grammatical descriptions is bought at some cost, for the information coded in constituent structures is considerably less than that supplied in traditional parsing. For example, in such structures heads and modifiers are not systematically marked, discontinuous constituents are not recognized, the relationship between the determined and determining constituents in government and agreement is not generally indicated, and only some of the applicable relations between NPs and Vs are noted. It is striking in this regard to compare the "coverage" of an elaborated traditional approach to parsing, such as Reed and Kellogg diagrams (see Gleason, 1965:142–51 for a succinct presentation), with that of formalized constituent structures (for instance, Harris, 1946 and Chomsky, 1956). Succeeding developments in grammatical theory can, in fact, be seen as attempts to devise fully formalized types of grammatical descriptions with something approaching the coverage of traditional grammars.

Before turning to these developments, however, we must comment on a further conceptual change set off by the move to formalized grammatical descriptions (in particular, constituent structures) as the output of the parsing operation. It is now possible to view parsing as algorithmic rather than heuristic. That is, it is now possible to see the parsing operation as the application of a series of language-particular principles, *phrase structure rules* like NP + VP = S and V + NP (+NP) = VP, to (a representation of) a sentence, in such a way that all the appropriate grammatical descriptions for that sentence, and no others, will be obtained.

Once such a change in characteristic (4) of parsing has been made, the way is open to view principles like NP + VP = S either analytically, as instructions for assigning structures to given sentences in a language, or synthetically, as instructions for composing the sentences of a language. That is, the full set of such principles constitutes a *formal grammar* for the language, which can be seen, indifferently, as having an analytic or *parsing function*, or a synthetic or *generative function*. (Both interpre-

tations appeared early in the history of formal grammatical theory – the generative interpretation most prominently in Chomsky's early work, the parsing interpretation in Hockett's "grammar for the hearer" [1961] and in "dependency grammar" [Hays, 1964; Robinson, 1970; among others].) Indeed, phrase structure rules can also be viewed neutrally, as having a *checking function*, an idea first mentioned in the linguistic literature by McCawley (1968) and developed recently in such work on generalized phrase structure grammar as Gazdar, 1982.

On its analytic or parsing interpretation, the phrase structure rule NP + VP = S licenses the grouping of a constituent known to be (or suspected of being) an NP along with an immediately following constituent known to be (or suspected of being) a VP, into a single constituent of type S; an acceptable constituent structure is then one that is headed by S and can be constructed from entirely by a sequence of such groupings. On its synthetic or generative interpretation, the rule is a *rewrite rule*, customarily formalized as S → NP VP, licensing the replacement of the symbol S in a line of a derivation by the string NP VP, or, equivalently, licensing the branching of a node labeled S in a constituent structure tree into two ordered nodes, labeled NP and VP, respectively; an acceptable constituent structure is then one that can be constructed from the symbol S by such rewriting, or from a node labeled S by such branching. On its neutral or checking interpretation, the rule is a *node admissibility condition*, stipulating that the subtree



is well formed; an acceptable constituent structure is then one that is headed by S and contains only admissible branchings.

When it is recognized that there is more than one way to view the function of phrase structure rules, then it is no longer necessary (though it is not barred) to think of parsing, or for that matter generation or checking, as something human beings do. Instead, these operations can be viewed abstractly, as performed by an idealized device – a change in characteristic (1) of parsing, one that makes characteristics (5) and (6) simply irrelevant.

The consequence of all these reconceptualizations is a distinct second notion of parsing, associated with formal theories of grammar. In this notion

(9) parsing is an operation performed by an abstract device,

(10) on (representations of) sentences in a natural language,

(11) resulting in a formal representation of sentence structure;

(12) this operation is algorithmic.

The next development is for the parsing, generative, and checking functions of a formal grammar for a natural language to be separated. It is a consequence of the particularly simple form of principles like NP + VP = S that they can be interpreted as steps in parsing, as rules for generation, or as node admissibility conditions. But if the steps, the rules, or the conditions are not of this simple, technically *context-free*, form, there is no guarantee that parsing operations, generative operations, and checking operations can be matched in a one-to-one-to-one fashion, and we must contemplate the possibility that the *parser*, the *generator* (sometimes referred to simply as the *grammar*), and the checking device or *acceptor* are three separate devices.

Historically, just such a separation, of parser and generator, followed on the development of transformational grammar as a particular generative framework. And more recently the development of generalized phrase structure grammar has required that generator/parser and acceptor be distinguished. In the first case, the perceived limitations of context-free generative grammar motivated a syntactic theory with at least two distinct components. What is relevant here is that in the new theory constituent structures are not generated directly by rewrite rules, so that a parser cannot be merely a generator run backward. In the second case, the intention was to rehabilitate context-free grammar as a plausible theory of syntactic structure. Part of this program depends on the fact that an acceptor making reference to local context accepts a set of constituent structures generable by a context-free generator (or parsable by a context-free parser); context-sensitive acceptors are thus not simply context-sensitive generators or parsers viewed in a different light.

In general, then, changes in the shape of syntactic theory carry with them consequences, often profound, with respect to the role and nature of a parser. In *monostratal theories* there is only one sort of syntactic representation, and it is the parser's business to assign just the right representations to any given sentence. The representations themselves will of course vary from one theoretical framework to another; the representations of arc pair grammar (Johnson and Postal, 1980) are graphs of a type quite different from the tree structures of classical constituent structure grammar, while the graphs of generalized phrase structure grammar are trees, but trees with node labels decomposed into sets of features (including a slash feature indicating a missing constituent). In *polystratal theories*, with two or more distinct levels of syntactic representation posited, the parser must either construct one level of representation (a *surface*, or *final, structure*) and then translate that into another (a *deep, basic*, or *initial structure*), or it must reconstruct the appropriate initial structures directly, thus operating in a fashion that bears no visible relationship to the operation of the generator.