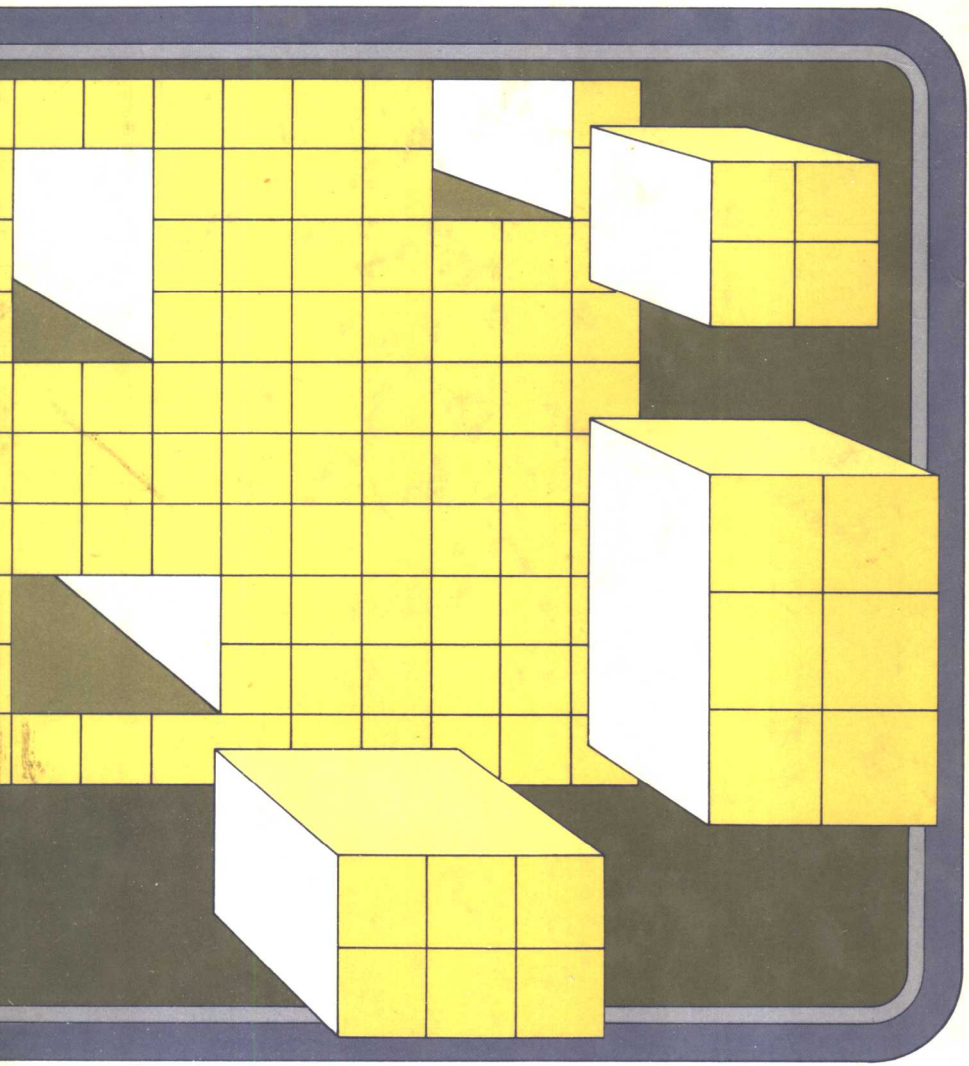
Ellis Horwood Series ARTIFICIAL INTELLIGENCE

Automatic Natural Language Parsing

edited by K.SPARCK JONES and Y. WILKS



AUTOMATIC NATURAL LANGUAGE PARSING

ELLIS HORWOOD SERIES IN ARTIFICIAL INTELLIGENCE

Series Editor: **Professor John Campbell,** University of Exeter

COMPUTER GAME PLAYING: Theory and Practice
M. A. BRAMER, The Open University, Milton Keynes

MACHINE INTELLIGENCE 8: Machine Representations of Knowledge
Edited by E. W. ELCOCK, University of Western Ontario, and D. MICHIE, University of Edinburgh

MACHINE INTELLIGENCE 9
Edited by J. E. HAYES, D. MICHIE, University of Edinburgh, and L. I. MIKULICH, Academy of Sciences, USSR

MACHINE INTELLIGENCE 10: Intelligent Systems: Practice and Perspective
Edited by J. E. HAYES and D. MICHIE, University of Edinburgh, and Y-H. PAO, Case Western Reserve University, Cleveland, Ohio

IMPLICATIONS OF COMPUTER INTELLIGENCE
Edited by M. YAZDANI and A. NARAYANAN, University of Exeter

AUTOMATIC NATURAL LANGUAGE PARSING
Edited by K. SPARCK JONES, University of Cambridge, and Y. WILKS, University of Essex

COMMUNICATING WITH DATA BASES IN NATURAL LANGUAGE
M. WALLACE, ICL, Bracknell, Berks

AUTOMATIC NATURAL LANGUAGE PARSING

Edited by
KAREN SPARCK JONES
Senior Research Associate
Computer Laboratory
University of Cambridge
and
YORICK WILKS
Professor, Cognitive Studies Centre
University of Essex

ELLIS HORWOOD LIMITED
Publishers · Chichester

Halsted Press: a division of
JOHN WILEY & SONS
York · Brisbane · Chichester · Ontario

First published in 1983 by

ELLIS HORWOOD LIMITED
Market Cross House, Cooper Street, Chichester, West Sussex, PO19 1EB, England

The publisher's colophon is reproduced from James Gillison's drawing of the ancient Market Cross, Chichester.

Distributors:

Australia, New Zealand, South-east Asia:
Jacaranda-Wiley Ltd., Jacaranda Press,
JOHN WILEY & SONS INC.,
G.P.O. Box 859, Brisbane, Queensland 40001, Australia

Canada:
JOHN WILEY & SONS CANADA LIMITED
22 Worcester Road, Rexdale, Ontario, Canada.

Europe, Africa:
JOHN WILEY & SONS LIMITED
Baffins Lane, Chichester, West Sussex, England.

North and South America and the rest of the world:
Halsted Press: a division of
JOHN WILEY & SONS
605 Third Avenue, New York, N.Y. 10016, U.S.A.

British Library Cataloguing in Publication Data
Automatic natural language parsing. –
(Ellis Horwood series in artificial intelligence)
1. Language data processing
I. Sparck Jones, Karen II. Wilks, Yorick
418 P98

Library of Congress Card No. 83-8601

ISBN 0-85312-621-6 (Ellis Horwood Limited)
ISBN 0-470-27460-3 (Halsted Press)

Typeset in Press Roman by Ellis Horwood Limited.
Printed in Great Britain by R. J. Acford, Chichester.

Table of Contents

IV – CONTEXT-FREE PARSING

V – SEMANTIC PROBLEMS IN PARSING

VI – SEMANTICS-DIRECTED PARSING

Preface

This book is aimed at research students and research workers interested in current views on the problems and techniques of automatic natural language parsing. Parsing is a key element of natural language processing as a whole, and the design of natural language processing systems is an important area on the one hand of information technology, and on the other of language studies. Information technology oriented research, concerned with both language-based systems for purposes like summarising and with language subsystems as components of, for example, expert systems, needs effective parsing procedures. Linguistics oriented research needs effective parsing models. The book is intended to exhibit the state of the art in automatic natural language parsing, at the intersection of these two concerns. Research and development in language processing over the last decade has explored specific approaches to parsing in some depth, has consolidated practical experience, and has emphasised some trends, for example towards phrase structure grammar and deterministic parsing, and towards the closer integration of syntax and semantics. The papers collected here, spread across the whole area of parsing, represent the present thinking of active workers in the field on issues and possibilities stemming from the last decade's work.

ACKNOWLEDGEMENT

This book is based on papers presented at a Workshop on Automatic Natural Language Parsing held at the University of Essex in April 1982. The Workshop was supported by the United Kingdom Science and Engineering Research Council, and we wish to thank the Council for funding the meeting and so encouraging work in this area.

Contributors

B. K. BOGURAEV	Computer Laboratory, University of Cambridge, Cambridge, UK.
E. J. BRISCOE	Department of Linguistics, University of Cambridge, Cambridge, UK.
A. CATER	Department of Computer Science, University College Dublin, Dublin, Republic of Ireland.
E. CHARNIAK	Department of Computer Science, Brown University, Providence, RI, USA.
T. CHRISTALLER	Research Unit for Information Science and Artificial Intelligence, University of Hamburg, Hamburg, Germany BRD.
G. GAZDAR	Cognitive Studies Programme, University of Sussex, Brighton, UK.
R. L. JOHNSON	Centre for Computational Linguistics, UMIST, Manchester, UK.
M. KAY	Xerox Palo Alto Research Centre, Palo Alto, CA, USA.
C. S. MELLISH	Cognitive Studies Programme, University of Sussex, Brighton, UK.
D. METZING	Department of Linguistics and Literary Studies, University of Bielefeld, Bielefeld, Germany, BRD.
S. G. PULMAN	Linguistics, School of English and American Studies, University of East Anglia, Norwich, UK.
C. K. RIESBECK	Department of Computer Science, Yale University, New Haven, CT, USA.
G. D. RITCHIE	Department of Computer Science, Heriot-Watt University, Edinburgh, UK.
K. SPARCK JONES	Computer Laboratory, University of Cambridge, Cambridge, UK.
M. STEEDMAN	Department of Psychology, University of Warwick, Coventry, UK.

J. I. TAIT — Computer Laboratory, University of Cambridge, Cambridge, UK

H. THOMPSON — Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK

Y. WILKS — Department of Computer Science, University of Essex, Colchester, UK

PART I

The field: structure, relations and applications

Introduction: a little light history

Y. Wilks and **K. Sparck Jones**, University of Essex; University of Cambridge

The papers in this book present, or comment on, recent ideas in automatic natural language parsing, using "parsing" to cover both the syntactic *and* the semantic analysis of a text in order to build a meaning representation for it. The purpose of this introduction is to set the papers in their historical context, and to motivate the way in which they are grouped.

It would be convenient, and intellectually tidy, if *automatic* natural language parsing could be discussed without any extended reference to the history of theoretical linguistics, as if computational analysis had its own autonomous life; and it is true that at times artificial intelligence workers have disregarded the concerns of contemporary theoretical linguistics. For others, however, autonomous parsing must be intellectually disreputable, mere amateur posturing: for them computational parsing ought to be the application of (well-founded) linguistic theory.

In fact, the historical relations between computational analysis and theoretical linguistics have been much more complex, with periods in which the two have essentially gone their own ways, and others where the connection has been much closer. Even in the latter, the flow of ideas has not always been in the direction theoretical linguists might take to be the natural one. It is true that the concerns of some research on parsing have been theoretical *ab initio* (for example the work on Generalised Phrase Structure Grammar). But in other cases (of which the Augmented Transition Network and Marcus's deterministic approaches *may* be examples (Woods, 1970; Marcus, 1980)), it is possible that the much publicised connection between the practical parsers and the theoretical results is *post hoc:* the procedural idea motivating the parsers preceded the search for theoretical justification. If this interpretation is correct, and we believe it is, then

work on automated natural language parsing has a logic and motivation of its own, and one may as well look for evidence of the influence of parsing practice on linguistic theory as the reverse.

There are indeed good reasons for looking at work on automatic natural language processing in its own right, not merely as a source of practical tips, but as an intellectual tradition. From the earliest days of machine translation in the nineteen-fifties, attempts to build parsers have adopted a 'procedural view of language', not simply as a consequence of the fact that one is writing programs, but in the more fundamental sense foreshadowed by Wittgenstein's advice: don't ask for the meaning, ask for the use. The distinctive emphasis on the activity of parsing as opposed to its formally possible products has had unexpected consequences, leading to views of grammar and linguistic knowledge very different from those of many linguists. It would be surprising if the by now extensive work on automatic natural language processing had had no impact on theoretical linguistics, and some influence is indeed detectable.

Research on automatic natural language processing has gained in strength and self-confidence in the last ten years, as it has been clearly shown that working systems can be built, even if they are only very modest ones. We have suggested that automatic parsing ideas have influenced theoretical linguistics; but theoretical linguistics has also manifestly affected automatic parsing. However, this has sometimes emerged as negative rather than positive affect: automatic parsing research has developed and emphasised ideas in opposition to those of theoretical linguistics. The semantic parsers of Riesbeck (1975) and Wilks (1975), for example, were largely motivated by a rejection of the assumptions of the dominant theoretical orthodoxy.

The rest of this introduction will look in a little more detail at the development of automatic natural language parsing, and hence at the sources of the specific interests and 'schools' represented by the groups of papers in this book. For this purpose, and to make explicit what has been tacitly assumed so far, we will simply state, without further elaboration, that parsing is formally a computational process, and hopefully an actual working program on some (non-human) computer, that takes sentences in a natural language (but preferably texts) and converts them by rules to some representational structure useful for further processing as might be required, for example, for translation or question-answering. Parsing is thus not confined to syntactic operations, but can include semantic ones: indeed the definition does not exclude operations applying pragmatic information (whether "pragmatic" is used to refer, in the manner familiar to linguists, to e.g. speaker/hearer features, or, in the manner familiar to workers in artificial intelligence, to e.g. real world facts). This 'definition', moreover, makes no commitments about the 'depth' of the resulting representation, using "depth" as a shorthand for a whole gamut of possibilities covering the types of concept used for the representation, the degree of abstraction of the representation from the text, the amount of explicit filling-in of implicit items, and so on.

Increasing depth may be viewed as a progression from a purely syntactic representation to an extended situational description, not explicit at all in the given text, However, this progression is only valid as a very crude approximation, since e.g. syntactic representations, or semantic ones, may themselves be more or less abstract, and may not necessarily exhibit uniform abstraction: for example a representation may be more abstract in its treatment of input lexical items than in its treatment of the input sentence structure. The definition makes no commitment either about whether the parsing process is reversible, i.e. whether it could be run the other way round to produce sentences from representations.

Strictly, the nature of the output representation produced by a parser is not relevant to any discussion of parsing. Parsing is a formally defined process, and the proper concerns of parsing are therefore the abstract mechanisms for applying structure-determining rules to input strings. However, it has been found very difficult to confine the discussion of parsing and the development of parsing techniques to those proper concerns, and views of parsing have been heavily influenced by beliefs about the substantive content of the rules applied, and even more by beliefs about the nature of the text representation to be supplied by the parser.

The 'neutral' definition of parsing given is intentional: it brings all the current approaches within the scope of the book. But it has not been adopted merely for editorial convenience: as the papers show, what the scope of parsing is or ought to be is a matter of argument, or at least discovery. The characteristic inputs, processes and outputs of parsing are objects of research; the nature of the representations to be built is a major problem and issue in the field. The book is thus as much about what workers in the field believe are the problems to be tackled and promising-looking ways of approaching them as about tried, tested and agreed solutions to problems. However, as the chapters show, work in automatic natural language parsing in the last twenty years has provided research workers and practitioners with a range of implementational and investigative techniques, at both the programming and the analytic levels.

Without wishing to be aggressively archaeological, it is necessary, in looking at the background to the current state of parsing research, to make a brief reference to the Stone Age of automated parsing represented by the early work on machine translation. In a general way, for the lack of an autonomous tradition of its own, this research on translation *de facto* tended to be within the then dominant structuralist paradigm of theoretical linguistics. It was predominantly syntactically, and surface, oriented, though inevitably, as translation was the task, the problem of how to tackle text meaning determination and specifically lexical sense selection could not be neglected. The solutions proposed by those most directly involved were felt at the time to be distinctly *ad hoc:* for example, assuming given universes of discourse (e.g. chemistry), or providing special-purpose interpretive routines as parts of lexical entries. It is therefore of some interest to note that somewhat more sophisticated versions of these

ideas (in the form of frames, procedural semantics, word expert parsing) have since appeared and have been accorded more respectful attention.

However, the main line of work was being attacked long before the public assault that the ALPAC Report of 1964 constituted, and both from within and from without. Some of those engaged in machine translation research were maintaining that semantics should be given a much more dominant role in parsing, effectively as the driver rather than as the finisher and polisher, and semantic primitives and semantic preference were already being advocated as parsing devices, though they were applied somewhat simplistically. At the same time, the phrase structure approach to syntactic description was being subjected by the theoretical linguists to the full weight of the Transformational Grammar juggernaut.

The relation of parsing to theoretical linguistics is brought out very clearly by reference to Transformational Grammar (TG) in its then Standard Theory (ST) form. From the beginning, for those concerned with parsing, the problems with the transformational approach were decidability and parsability: indeed these concepts are closely related in that, putting it crudely, only decidable systems can be reliably parsed. In its initial form, TG had neither of these properties. Much of the subsequent work within the TG paradigm (e.g on constraints) can be seen as an attempt to limit the power of TGs so that they would be decidable (at least as regards the class of structures generated, if not of strings).

Practical parsing workers were somewhat sceptical about these efforts to 'control' TG, because they were well aware that theoretical decidability and parsability in no way guarantee that a system will actually parse real sentences it is offered within a medically finite lifetime. Moreover, irrespective of the relation between decidability and parsability as a matter of principle concerning grammars, the intrinsic generative character of ST meant that, in a much more important and real, engineering sense, transformational grammars were not for parsing (nor, to be fair, had they been offered as such by linguists). This is indeed shown by the contortions and compromises found necessary in the building of ST parsers seeking a more respectable linguistic foundation for their efforts than any hitherto plausibly on offer.

It was here that the augmented transition network (ATN) was so important: here was a formalism of the same power as a transformational grammar, as a means of characterising sentence structures and structural relationships, but one whose operational claims could be clearly stated, and which, in addition, provided a 'surface grammar', i.e. provided a means of obtaining a sentence surface structure from the given input string.

This parsing capacity was crucial, for ST grammars could never be parsing systems in their standard form simply because they operated on and generated phrase markers: ST provided no mechanisms by which a parser could map, in the reverse direction, from sentence strings to deep structures. Although ST was claimed to be non-directional, it had, from the beginning, a deep-to-surface

virtual machine associated with it. In the transformational parsing systems that were built this was the crux: *ad hoc*, heuristic links between surface trees and candidate deep sources had to be provided to guide the parsing-by-generation process.

In relation to TG, therefore, parsing workers found themselves in the awkward position of being offered a linguistic theory of manifestly superior status to anything they had tried to apply before, but one which was of such an abstract, descriptive character that it could not itself be applied to a processing task in any perspicuous way. Thus to those parsing workers concerned with syntax, the main consequence of ST was that it presented, in a well-founded way, an important general idea, namely that of deep structure, supported by a large body of specific structural observations and analyses. The felt semantic inadequacies of TG were major sources of concern to adherents of what may be described as the alternative tradition, which gives priority in parsing for discourse interpretation to semantic rather than syntactic processing, as will be seen below. But the problem of getting a procedural grip on the abstractions of TG was a challenge even to the adherents of the syntactic tradition in parsing, and so for them, ATNs were manna in the wilderness.

The ATN idea has had considerable staying power, and it has provided a useful tool largely because of the attractive way in which it explicates left-to-right processing, and because of the way in which, while largely separating the parsing interpreter from the grammar being applied, it encourages the grammar writer to think about his rules in a procedural way. More specifically, ATNs, in contrast to ST, provided a set of structure-building actions offering a procedurally neat way of linking the surface structure of the network path with the deep structure held in some structure register(s).

There have been subsequent developments of ATNs, for example island parsing, and cascading, which have extended the range of application of ATN parsers. There are also problems with ATNs: as linguistic description systems they fail to capture certain generalisations, for example in relation to conjunctions; and as programmable systems they suffer, for example, from a lack of properly scoped variables. The papers on ATNs in this volume illustrate some of the concerns for those who find the ATN model a useful one. It is worth noting that, although ATN parsing has been primarily applied in syntax-driven mode, it has also been successfully applied in semantically-driven analysis, illustrating its hospitality as a formalism. The ATN tradition is still a currently active one in automated parsing.

More recently, developments of TG itself, from the Standard Theory to the Extended Standard Theory, have found procedural explication in parsing implementations of TG, notably Marcus's PARSIFAL. Marcus claims that his deterministic parser is a theoretically-motivated implementation of the Extended Theory, and not merely a programmed hack, in the sense in which earlier attempts to implement transformational parsing were, theoretically speaking, unmotivated.

To be that, PARSIFAL must show in its operation the kind of universal constraints on the theory of grammar that Chomsky advocates. For him, though, these constraints are declaratively expressed and a component of the theory of grammar rather than procedurally explained.

The interest aroused by Marcus's 'determinism hypothesis' has been focussed more on the psychological claim being made than on the practical application of the model. The clear challenge is whether, within a syntactic framework, parsing can be done deterministically, given the definition of determinism Marcus offers, in terms of no building of (ultimately) unused structures. This is obviously a matter of importance for those who find it convenient to do syntactic processing before semantic processing. However, it has been suggested that the deterministic strategy will sometimes fail on sentences other than 'garden paths', and will require semantics in support of syntax. Marcus concedes the latter, and this has given comfort to those who advocate a larger role for semantics: both those who argue for its necessity as an aid to syntax (and so deny any useful autonomy to syntax), and those who have claimed effective determinism for semantic parsing. What is therefore of interest for computational parsing, rather than psycholinguistics, is the extent to which Marcus-style syntactic parsing can be usefully combined with semantics for a more powerful parser overall: this is currently an open question.

An important theoretical development in recent years has been the astonishing change in the intellectual climate implied by the re-evaluation of Chomsky's long-accepted argument that Finite State Grammars (FSGs) and Phrase Structure Grammars (PSGs) are inadequate for natural languages. Some theoreticians (for instance Harman, 1963) never accepted these arguments, at least as far as PSGs are concerned; computationalists were often unconvinced (see Wilks, 1967), and formalists (like Joshi and Levy, 1982) produced convincing counter-arguments. In the last few years a noticeable revival of interest in PSGs has occurred, most obviously connected with Gazdar's work on Generalised Phrase Structure Grammars (GPSGs) (see this volume).

GPSG has many attractions as a linguistic theory, notably the fact that phrase structure offers a firm anchoring of underlying structures in the realities of actual text, which in turn implies a potentially firm grounding for surface semantics as well as syntax. More to the point in the present context, it is easy to see the attractions of GPSG for computational parsing: unlike TG, GPSG can easily be given limitations to make it decidable, and so parsable; and the formal semantic theories which at least some computationalists believe offer the right kind of tools for their purposes assume a correspondence between syntactic and semantic rules of a sort that is readily available in GPSG. It is not surprising, therefore, that GPSG should currently be an object of intensive investigation by both linguistic and computational communities.

It is hardly surprising either, given all this, that some should want to reverse the historical process yet further and re-examine the relation between PSG and