

PARSING NATURAL LANGUAGE

King, Margaret (editor)

London: Academic Press, 1983 [second printing, 1986], x+308 pp.

Hardbound, ISBN 0-12-408280-7, \$21.00, £12.60

Reviewed by
Klaus Schubert
BSO/Research

This book falls into three parts, an introductory, a syntactic, and a semantic one. In a rapidly developing branch of applied science like parsing, it is not at all surprising that the authors have more to say about the fundamentals than about syntactic applications, and more about syntax than about semantics. This is so at least if measured by the number of articles: five introductory, four syntactic, and three semantic contributions.

Especially the first part appears to be well edited, so that the five articles form a readable and worthwhile whole. Starting from Anne De Roeck's definitions of parsing, recognizing, grammar, and various parsing strategies, the introduction proceeds to presentations of a procedural parsing system (by Margaret King) and a declarative parsing system (by Michael Rosner). Special attention is paid to different versions of transition networks (by Roderick Johnson) and to charts (by Giovanni Varile) as a powerful data structure for partial parsing results, applicable in parsers of various types.

In the syntactic part, Geoffrey Sampson discusses a deterministic parser for which psychological validity is claimed. Eugene Charniak's parser for both grammatical and ill-formed utterances has a similar orientation, although Charniak admits that psychological validity has not been attained. His parser is special, as compared to other systems, because it assigns a grammaticality figure rather than providing a yes/no judgement on the correctness of a sentence. In his second contribution, Sampson discusses (and discards) many arguments against context-free grammars, suggesting that they can be made powerful enough for describing human languages. Steve Pulman presents an implementation of Chomsky's (1981: 55ff) trace theory for long-distance dependencies and similar notorious problems. Psychological validity is again at issue when semantic parsing is taken up. Graeme Ritchie especially focuses on the problems of local decidability of semantic choices, taking as proven by psychology that humans understand much in a word-by-word way. He admits that a computer system that would perform a semantic analysis in this fashion has not been devised yet. Yorick Wilks discusses deep and superficial methods of semantic analysing, arriving at the conclusion that the differences are not as significant as had been supposed. Steven Small describes in much detail a semantic word-

expert parser used for determining the meaning of words in a given context.

The articles come from a tutorial held at Lugano in 1981. Reviewing them today, it may be interesting to ask how the book can be assessed now, in view of these six years' developments. Much could be said; I choose three observations.

First, it strikes me to what extent Augmented Transition Networks seem, at that time, to have been the most well-known parsing formalism, referred to by almost every author. The Lugano tutorial was held at a time when two powerful competitors of ATNs were being brought to a wider public's attention, but probably had not been recognized as such: Definite Clause Grammar (Pereira and Warren 1980) and Lexical-Functional Grammar (Kaplan and Bresnan 1982). Another development has a forerunner in this book: Sampson's "context-free parsing" is based on work by Gerald Gazdar and others that later was to become known as Generalized Phrase Structure Grammar (Gazdar, et al, 1985). (On all these formalisms cf. Schubert 1987: 211ff.)

Second, for all the authors the most natural language seems to be English. I have found only four or five sample sentences from other languages. This is a severe shortcoming in a book where keywords such as "psychological validity" and "syntactic universals" are used. I feel that computational linguistics has become slightly more international since then, but wide-scoped publications are still rare.

Third, there is a certain lack of alternative thinking in another respect. There are two approaches for analysing grammatical systems, both feasible to parsing: the constituency and the dependency approaches (Schubert 1987: 17ff). On the syntactic level, this book is entirely constituency-minded, although an introductory series of articles would be an appropriate place for at least mentioning that one has made a deliberate choice by presenting only one of the two approaches. Dependency syntax seems to have spread and gained ground in computational applications since 1981.

REFERENCES

- Chomsky, Noam. 1981 *Lectures on government and binding*. Dordrecht: Foris.
- Gazdar, G.; Klein, E.; Pullum, G.; Sag, I. 1985 *Generalized phrase structure grammar*. Oxford: Blackwell.
- Kaplan, Ronald M. and Bresnan, Joan. 1982 *Lexical-Functional grammar: A formal system for grammatical representation*. In Bresnan, J., Ed., *The mental representation of grammatical relations*. Cambridge [USA]: MIT Press.
- Pereira, Fernando C. N. and Warren, David H. D. 1980 *Definite clause grammars for language analysis—a survey of the formalism and a comparison with augmented transition networks*. *Artificial Intelligence*, 13: 231–278.
- Schubert, Klaus. 1987 *Metataxis. Contrastive dependency syntax for machine translation*. Dordrecht and Providence: Foris.
- Klaus Schubert holds a doctoral degree in General, Scandinavian, and Slavic Linguistics from the University of Kiel (Federal Republic of

Germany). Since 1985 he has been with the DLT machine translation project at the BSO software house in Utrecht (The Netherlands), where he is responsible for the overall linguistic system design. Schubert's address is: BSO/Research, Postbus 8348, NL-3503 RH Utrecht, The Netherlands. E-mail: schubert@dlt1.uucp.

**MACHINE TRANSLATION: THEORETICAL AND
METHODOLOGICAL ISSUES (STUDIES IN NATURAL
LANGUAGE PROCESSING)**

Nirenburg, Sergei (editor)

Cambridge, England: Cambridge University Press,
1987, xv+350 pp
Hardbound, ISBN 0-521-33125-0, \$49.50; Paperback,
ISBN 0-521-33696-1, \$17.95 [20% discount to ACL
members]

Reviewed by
Harold Somers
UMIST

This title in the ACL-sponsored "Studies in Natural Language Processing" series is a collection of seventeen papers on Machine Translation (MT). Twelve of the papers are revised versions of presentations at the 1985 Colgate conference, and four of the papers are completely original. The appearance of this book is further evidence of the massive renaissance of MT, along with which comes the serious risk of surfeit of books and articles covering the same material. It is important, therefore, that publications in this field have something new to offer the experienced reader. Happily, this is the case with Nirenburg's collection, since most of the papers either focus on theoretical and experimental approaches to the problem, or else address some of the less commonly considered aspects of MT. Notable in this respect are Weischedel and Ramshaw's discussion of ill-formed input, McDonald on generation, and Walker's description of tools for extracting information from large databases.

Like the book itself, the remainder of this review is divided into six parts, as reflected in the section titles.

1. The state of the art

The editor's introduction is a rather heterogeneous collection of introductory thoughts. First, some of the flavor of MT is presented through the discussion of the types of knowledge involved in an MT system and the ways in which problems can be addressed, notably by restricting the input, or by involving the human in the translation process. The final section of this chapter gives a useful overview and summary of the remainder of the book. Tucker's contribution is a revised version of his 1984 *ARIST* article, and concerns strategies for MT together with brief sections on sublanguage and evaluation. Various MT systems are reviewed: these are divided into the "operational" systems SYSTRAN, SPANAM, TAUM-METEO, and METAL, and experimental projects EUROTRA, Mu, SUSY, DLT, and

TRANSLATOR. Most of these will be familiar names, except perhaps the last two: TRANSLATOR is Tucker and Nirenburg's own system, while DLT is a project under way in Utrecht, which envisages the use of Esperanto as an interlingua. Tucker is justly critical of this project, and one wonders why it merits two pages of discussion when other more worthy systems are not mentioned.

2. MT and linguistic theory

Raskin discusses the relationship between linguistics and NLP. He begins with a catalog of the various elements of linguistics and gives examples of problems in each domain which are relevant to NLP. He says that linguistic treatments are never complete; furthermore, they are rarely available in a coherent form acceptable for immediate implementation. Therefore, NLP projects must have linguists on their staff who know about, and can gain access to, linguistic materials. But theoretical linguistic work is not always useful for NLP, as we are shown (pp. 52-3) in an interesting point-by-point analysis of the different needs of theoretical linguistics and NLP.

Kittredge's excellent contribution is on the significance of sublanguage for MT. "Sublanguage" is defined informally as a linguistic system used in a particular domain of discourse, and is characterized by specific recurring structures and vocabulary. Although a sublanguage is a proper subset of some natural language, it will not necessarily be a subset of the general variety of that language. For example, the English of weather bulletins (as in METEO) has sentence patterns which are not generally found in standard English (e.g., omitted articles and lack of tensed verbs):

In a sublanguage, the rules for constructing sentences may be quite different from (and even contrary to) the rules for sentences in the 'standard' language. (p. 63)

The main attraction of a given sublanguage for the purposes of MT is the extent to which it can be described by a significantly smaller grammar than that required for the full general language, and the extent to which lexical ambiguities are reduced by the exclusion of non-domain-relevant alternatives. Some sublanguages are not so "well-behaved" in this respect, permitting "seepage" from general language (p. 63).

Kittredge next considers the choice of sublanguages as suitable candidates for MT, noting that not all sublanguages are necessarily good in this respect. This was the experience of the TAUM-AVIATION project, where some of the characteristics of the aircraft hydraulics manual sublanguage were particularly unsuitable for MT (e.g., complex NPs). Finally, he offers some guides to estimating the suitability of candidate sublanguages for MT. These include comparing vocabulary size in texts of different lengths: a vocabulary growth curve which tends to flatten is a good indicator of a constrained vocabulary. Estimating the computational tractability of the grammar is more difficult. Kittredge gives