Book Reviews

Natural Language Computing: The Commercial Applications

Tim Johnson

London: Ovum Ltd, 1985, 459 pp.
ISBN 0-903969-22-X; $395, or £275 in the UK
[Available from the publisher, 44 Russell Square, London WC1B 4JP, England. Price includes airmail postage.]

> This report aims to identify the areas where NLP (natural language processing) is expected to be successful, and to suggest how people and organisations can make use of the opportunities it will create, as users, suppliers or investors. (p. 4)

Tim Johnson has written a very useful guide to the commercial side of NLP applications. It is a large, comprehensive report that runs 459 double-spaced pages. Johnson is a graduate of Imperial College and has previously written a report on Expert Systems (July 1984), so he is familiar with work in other areas of machine intelligence. This report is better technically than one might expect from a layman. At $395 it is probably most appropriate for managers or researchers who are actively involved in the marketplace and for institutional or library copies.

The major sections of the report are Management Summary, Markets, Technology, Applications, and Company Profiles. The first two sections are of primary interest either to those who are unfamiliar with the major application areas of NLP (described as mainframe and micro database interfaces, dialogue interfaces, content scanning, text editing, machine translation, and talkwriter) or with the business side of projected markets. For example, total US market projections are given as $15M in 1985, $420M in 1990, and $1500M in 1995. These projections are also broken down by application area.

For active researchers in NLP, the technology section will not contain any major surprises. A very brief tour of approaches in syntactic and semantic parsing is followed by a review of current systems in both the mainframe and micro markets. These include Intellect, Ramis II English, Plume, Themis, Easytalk, Clout, NaturalLink, Savvy, Microdata Natural Language, Safeguard Cash-Management System, Logos's Intelligent Translator, ALPS Computer Translation System, Smart Translator, and Weidner MicroCAT. For each of these systems, there is a brief discussion of the product and a summary table that lists pertinent data such as availability, price, computer requirements, implementation language, opportunities for customizing, the software interface, the underlying NL technology, input requirements (menu-driven, ill-formed input, spelling correction), vocabulary size, and dialog management facilities (e.g., ellipsis and anaphora).

The last two sections of the report contain the material of greatest interest to me, since it is not readily available elsewhere. The application section describes user experiences with a number of systems. For example, PPG, an Intellect site, estimates that they were handling 300-500 queries a day in early 1985. In one anecdote, "it took two questions and about five minutes through Intellect to come up with data that would have taken hours by any other route." Johnson also describes the experience of Cognitive Systems with Explorer, a custom system that accesses a cartographic geologic database for oil exploration.

The last section contains profiles of 30 organizations engaged in NLP research or development. The organizations include companies, research labs, and universities. This is the place to find out what is going on at Kurzweil now, what venture capital is supporting the company (Xerox and Wang among others), who has non-exclusive marketing rights to Kurzweil's AI products (Xerox), how many staff they have (50), who their major players are (Dennis Klatt, Francis Ganong, Susumu Kuno and Glen Akers), how they are doing and where they are going. Even those who think they have a fairly comprehensive mental "who's who" of the field should still find this section of interest.

A set of appendices provides additional material, the most useful of which is a listing of organization addresses. My biggest complaint is the lack of an index for a volume this large. As a partial compensation, the table of contents is five pages long; however, it is impossible to quickly find information such as where is Gary Hendrix working (Symantec), which is available from the text.

Mark Jones
AT&T Bell Laboratories
Murray Hill, NJ 07974

Boolean Semantics for Natural Language (Synthese Language Library, 23)

Edward L. Keenan and Leonard M. Faltz

Dordrecht: D. Reidel, 1985, xii+387 pp.
ISBN 90-277-1768-0; Dfl 145,-, $54.00, £36.95

Part of the enterprise of model-theoretic semantics is the construction of formal tools that illustrate and, one hopes, explain linguistic phenomena. That is, a mathematical apparatus is built that models some feature of natural language, typically the entailment relation between (sets of) sentences. The cornerstone of this approach is the Fregean principle of compositionality, which, when used along with a categorial syntax, suggests

that the semantic spaces in which expressions are interpreted are collections of functions of higher type.

As Keenan and Faltz state, *Boolean Semantics for Natural Language* "situates itself squarely in the tradition of model-theoretic semantics". Nevertheless, their work develops an insight absent from other work in the Fregean tradition. As suggested by modern algebra, the study of arbitrary functions from one set to another is often less enlightening than the study of those functions that preserve some primitive underlying structure of the sets involved. Now Keenan and Faltz locate a structural property of natural language and use it as the main feature of their system.

The structural property basic to the book is the closure of syntactic categories under boolean combination. That is, in, for example, *John and Mary walked to school or to the beach*, we see an NP, *John and Mary*, which is itself a conjunction of the proper nouns *John* and *Mary*, and a PP, *to school or to the beach*, which is a disjunction of two PPs. Rejecting analyses of boolean combinations as arising from paraphrastic reduction, Keenan and Faltz endow semantic interpretation spaces with boolean operations. The key point here is that a boolean combination of expressions of a given category can be interpreted in a uniform way as a function of the interpretations of those expressions. This leads to the view that semantic spaces should reflect or preserve a boolean structure. It turns out that in the system of *BSNL*, some categories are interpreted in spaces that are boolean algebras, some in spaces that are homomorphisms, and some others in spaces definable by axioms stated in boolean terms. For example, the semantic space for the category N of common noun phrases is taken to be an arbitrary complete and atomic boolean algebra P.% Now adjectives combine with Ns to form Ns, but the semantic space is not the whole class of maps from P to P. Extensional adjectives satisfy entailment patterns exemplified by *Every small fish is a fish.* In boolean terms, these satisfy the definition of a *restricting map:*

$$\forall\, p \in P,\ f(p) \leq p.$$

It also turns out that the class of restricting maps can itself be made into a boolean algebra.

Following a short overview that highlights some of the underlying principles of the work and some of the main themes, the book is divided into two parts. The first presents the extensional system, a syntactic and semantic system that includes such categories as NP, proper NP, VP (both one-place and n-place), Determiner, Adjective Phrase, and PP. In addition, there is a discussion of how passives, reflexives, and relative clauses (among other constructions) are handled. One of the themes is that the important subcategories can be defined in terms of boolean axioms. There is also a nice technical result called the Justification Theorem which is too complicated to discuss here except to say that it allows, for example, the interpretations of one-place predicates to be determined by their action on a small set of NPs, the interpretations of proper nouns.

The second section extends the system to consider intensional phenomena, including intensional predicates and adjectives, sentential complements (through introduction of bar categories), and de-re/de-dicto ambiguity. The main difference between the two parts is that the first part studies phenomena that can be modeled using set-theoretic resources; thus it does not treat tense and aspect, and the boolean connectives are interpreted classically. The second part extends the first part using possible-world semantics. The combination of the boolean approach with possible-world semantics leads to some challenging data. It also poses technical problems which, as the authors point out, are not fully resolved. I think it is fair to say that the most fruitful application of the boolean approach has been in the extensional system.

Readers of this journal will undoubtedly be interested in the computational significance of boolean semantics. Keenan and Faltz do not address any computational issues directly, but there are a few points that can be made in this regard. First, boolean semantics should be more amenable to actual implementation than systems based on Montague grammar. The primary reason is that the sizes of the semantic spaces are smaller here as we consider special classes of functions. Equally important is the fact that many of the classes have algebraic descriptions in terms of generators, and the generators are a small set. For example, the Justification Theorem mentioned above makes the interpretation of predicates tractable by reducing them to sets of individuals. One can also imagine the development of efficient semantic algorithms based on the algebraic approach for exactly the same reason, and these could be used to account for the relative difficulty in the interpretation of various sentences. (It should be noted that the combinatorial explosion is not completely diffused by the boolean approach, since higher categories such as PP are still interpreted in spaces that grow superexponentially.) Still, this book suggests ways in which computation and semantics can interact.

BSNL is a reworked version of the authors' 1978 monograph, *Logical Types for Natural Language* (UCLA Occasional Papers in Linguistics, No. 3). It should be accessible to anyone with an interest in semantics. The authors compare their proposals with those in the linguistic literature only at a few places, including the treatment of passive and sentential complements, and thus familiarity with this literature is not assumed. A successful reader would need to be comfortable with mathematical ideas and notation. Even though there are many examples and proofs worked in detail, someone with no exposure to algebra would find even the extensional system rough and the intensional system rougher still. The authors use a number of different modes of argumentation, including the examination of data from languages other than

English. This is particularly so in the early part of the book. When making a proposal, they often examine several possibilities which are all plausible in their framework. This should be welcomed by readers with only limited exposure to formal semantics.

The authors address methodological issues at several points, since their main concern is the presentation of a fragment. Nevertheless, it would have been useful for the authors to situate their work more explicitly by indicating what a boolean emphasis has added to semantics, what areas are likely to benefit from a similar approach, and what problems will require new methods. This book shows that the algebraic approach offers much of interest to semantics.

Lawrence S. Moss
Department of Mathematics
University of Michigan
Ann Arbor, Michigan 48109

GENESIS: AN AUTHORSHIP STUDY IN COMPUTER-ASSISTED STATISTICAL LINGUISTICS

Y.T. Radday and H. Shore
with D.Wickman, M.A. Pollatschek, Ch. Rabin, and Sh. Talmon

Rome: Biblical Institute Press, 1985, xx+235 pp.

This book studies a topic in Bible scholarship by means of computer-assisted statistical methods. The book has seven parts:

I Introductory, by Radday,
II On statistics in general and in Genesis, by Wickman,
III Statistical analysis of formal criteria, by Shore and Radday,
IV Vocabulary richness and concentration, by Pollatschek and Radday,
V. An interim postscript, by Radday and Shore,
VI Linguistic aspects, by Rabin,
VII A Bible scholar's evaluation, by Talmon.

The book combines the work of authors from different fields of research: Radday is a Bible scholar, and as a matter of fact, was the first to introduce statistical computational methods into Bible study in Israel; Pollatschek is specializing now in operations research and computer techniques. Both are at the Technion. Shore carries titles in industrial engineering and operations research, as well as in philosophy and psychology. Wickman teaches mathematics and statistics at the Technische Hochschule in Aachen, FRG. Rabin is from the field of classical and modern Hebrew and Arabic linguistics, and Talmon is Magness Professor of Bible Studies. Both these latter scholars are from the Hebrew University of Jerusalem.

The goal of this study was to examine the authorship of the Book of Genesis. Textual and epexegetical difficulties in this book roused theories, such as Wellhausen's, that the Book had not been written originally by one hand, or that a later editor edited material written by several (at least three) previous writers. This "documentary theory" is investigated in the study in three phases:

Phase I general statistics;
Phase II statistics of linguistic data;
Phase III vocabulary statistics.

The text of Genesis was divided into three parts:

(1) Sort of Text:
N (Narrative);
H (Human speaker);
D (Divine speech);
(2) Documentary Source (following Wellhausen's theory):
J (text pieces using the letters JHWH for God),
E (text parts using the word Elohim for God), and
P (representing a priestly writer);
(3) Division I, II, III, according to story-type, namely the first cycle of stories of the creation, the flood etc., the heroic stories of the Fathers, and the cycle of stories about Joseph. The creation story of the first chapter and Jacob's blessing were excluded since they were linguistically too deviant.

Further slicing of each of the above classes yielded text sequences of about 200 words each, which were convenient for statistical analysis. Each sample was statistically examined for 54 linguistic items in the areas of word length (2 to 10 characters); certain nominal and verbal morphology elements; syntactical elements; and frequency of inter-word transfers, such as noun/noun, noun/verb, noun/pronoun, and noun/stop. These features are unique for any writer, and cannot be consciously manipulated. They therefore reveal the writer's individual style and may corroborate assumptions concerning the text's author. Statistical results show minute differences per item; but when all details are collected, consistent facts of certain linguistic features emerge and yield a complex picture of the linguistic structure of the text.

In Phase I the univariate analysis of variance revealed that J and E were indistinguishable from one another, while P was strikingly unlike either. Also, NP was heterogeneous, while NJ and NE were not. The multivariate analysis of variance demonstrated the same pattern of differences among the documents. P appeared to be of an independent source, while J and E bore very close resemblances.

In the analysis of the sorts of discourse, N is completely unlike H and D in any document. Seventeen out of 39 variables were found to be powerful discriminants between N and H+D. All the results of these analyses mean that the Narrator behaves linguistically in a significantly different fashion from the speakers.