from many other writings on the subject, is its discussion of the problem of evaluating MT systems. The proposed methodology is decomposed into three distinct areas: (i) evaluation by the system's designer; (ii) cost/benefit evaluation by the user; and (iii) linguistic evaluation by the user. This delineation serves as the framework for a more detailed and impressive, though by no means final, study contained in Appendix A.

In the conclusion, Lehrberger and Bourbeau discuss the feasibility of MT, its future prospects, and the impact of evaluation methodology on those prospects.

To summarize: I thought that this book was very well written and intended for the mature MT researcher. The impact of the book would be even greater had it been published earlier in the decade.

References

CMU-CBT 1989, KBMT-89, Technical Report, Carnegie Mellon University, Center for Machine Translation.

Hutchins, W.J. 1986 *Machine translation: Past, present, future.* Chichester, England: Ellis Horwood Limited.

Nirenburg, Sergei (ed.) 1987 *Machine translation: Theoretical and methodological issues.* Cambridge, England: Cambridge University Press.

Rita McCardell is a Ph.D. candidate at the Computer Science Department of the University of Maryland, Baltimore County. Her interests include machine translation and natural language generation. McCardell's address is: 111 Rodeo Circle, Baltimore, MD 21220 E-mail: rita@nl.cs.cmu.edu or mccardel@umbc3.umd.edu

---

Medical Language Processing: Computer Management of Narrative Data

Naomi Sager, Carol Friedman, and Margaret S. Lyman
(New York University)

Reading, MA: Addison-Wesley, 1987, xiii + 348 pp.
ISBN 0-201-16810-3, $41.95 (hb)

Reviewed by
Nicoletta Calzolari
University of Pisa

The book under review builds on and is an extension of the New York University Linguistic String Project system applied to medical language processing. The system analyzes free text and converts the information 'hidden' in it, the syntactic and semantic regularities, into an informationally equivalent structured form, which is best suited for information retrieval and automatic summarization. From the computational linguistics point of view, the main interesting results consist on the one hand of the demonstration that a real world text processing application of linguistic analysis is possible (i.e., the processing of real textual input), and on the other hand in the fact that the methodology and the techniques used here and described for medical language are by and large also applicable to other, completely different, environments. The work also has links to knowledge representation, given that a method for representing and processing semantic information is provided, and the data supplied could be a testbed for knowledge-based systems.

In Chapter 1 a general overview is given of the problems, the methodology, and the theoretical support involved in processing natural language and sublanguage in particular. It is by syntactic clues that a set of semantic statement types are individuated, and therefore semantic results are achieved, but the main methods of analysis are dictionary look-up and pattern matching.

Chapter 2 is of less relevance for linguistics; it is mainly concerned with the medical aspects of the project, and with its practical applications purely from the physician's point of view.

Chapter 3 describes the types of information structures that are typical of the sublanguage of medical narrative and the way in which they are mapped into computer representations, i.e., rather simple information formats. Grammatical paraphrase, deletion of redundant words, and regularization procedures are some of the main procedures used to obtain the information formats from the surface grammatical structure. These format structures, although resembling 'classical' frames, are specifically designed to "reflect the linguistic regularities observed in sublanguage texts, and therefore differ from most uses of frames in artificial intelligence applications." Whether they really differ is perhaps questionable: they both try to capture similar types of regularities and formalize underlying grammatical relations into predicational structures. In my opinion, the real difference is in their suitability as to their application to (and empirical derivation from) real texts.

It is interesting to learn that only six types of information formats, plus seven types of modifier formats, are sufficient for representing information in clinical narrative texts. How many would be necessary if dealing with other types of sublanguage texts? How many for general language? An evaluation of them in other fields and a comparison would be interesting.

Chapter 4 describes how the system uses lists of sublanguage word subclasses, with constraints on the syntactic relations occurring between them, in order to accomplish some linguistic tasks, e.g., to rule out inappropriate prepositional phrase attachments (a typical problem unsolved with pure syntactic analysis) and to select the attachments permitted in the domain. The same method, i.e., checking against a list of well-formed word class patterns, is used for homograph disambiguation. An essential tool is therefore the possibility of classifying lexical entries into a well-defined set of semantic word classes, for which it is possible to state a number of syntactic and semantic properties in the sublanguage being analyzed. These entries do all the work. This approach, which gives good results, is of

interest to me because it stresses the importance of the lexicon in NLP systems and also encourages lexical analysis in general language. Many parts and components of the system are in fact essentially lexically driven.

Another typical problem, the analysis of conjunctions, is handled with far better results by means of sublanguage selection: conjunction is restricted only to members of the same or similar subclasses (i.e., belonging to the same equivalence class). The list of word classes is also used in computing the semantic subclass of a phrase from the subclasses of the head noun and the adjunct.

Procedures that use this classification in lists of sublanguage word classes occurring in particular syntactic relations (i.e., sublanguage co-occurrence patterns) not only structurally improve the parse tree, but also allow adding to the parse tree semantic information that:

1. provides semantic characterization to the words and phrases;
2. solves word sense and syntactic ambiguity (for adjuncts);
3. identifies combinations of basic statement types;
4. determines the underlying structure of compound noun phrases.

All these are advantages of dealing with a sublanguage. We must in fact say that these patterns are highly dependent on the particular domain, so that the lists would obviously have to be redefined for another sublanguage domain, but the same procedure can still be used. All these pattern-matching procedures, and the transformational machinery, make it possible to reduce the various linguistic means of expressing the medical information to standard forms.

A crucial step in ensuring the quality of the system to be designed is obviously a preliminary linguistic analysis of sample sublanguage documents. The main results of this analysis must be to determine:

1. the special well-formed syntactic structures in the domain;
2. the specialized word subclasses;
3. the basic statement types, based on recurring syntactic patterns of sublanguage word class co-occurrence.

All these must be stated in relation to general standard English.

The overall system in fact is made up of components (a processor and a grammar) for parsing and regularizing general English, which form the shell of the system, and other modules that are sublanguage specific. However, the latter procedures are ultimately highly portable and applicable to other domains, since their constraints are mainly list-driven. This is a good feature of the system, also due to its modularization.

In Chapter 5, various experiments using existing database management systems are described, showing both advantages and disadvantages. But too many details of the implementation—the actual organization, storing, and content (in terms of types of data) of the tables of the database and the query types and procedures to implement retrievals—would be more interesting for database people than for computational linguists, and would seem more appropriate for a manual for the system user than for a scientific book dealing with a research program and results.

The final output form is a tabular parenthesized structure which, "although structurally flat, represents the complete hierarchical linguistic structure of the original narrative," and is suitable for successive querying and retrieval in a database of relational type.

Chapter 6 describes the dictionary as a very important component of the system. It is rather large compared to the lexicons of usual CL systems: about 12,000 entries of common English words, subclassified for approximately 150 syntactic properties. A subset, i.e., the specialized medical dictionary, contains additional subcategorization, and is subdivided into 45 medical sublanguage classes, which ultimately allow correct syntactic analysis, determine the correct semantic statement type, and map the words into the correct format slots. Incorrect parses are ruled out after testing for syntactic and semantic compatibility of co-occurring text words. Very useful to the processing is the classification of verbs in terms of noun subclasses disallowed as subjects and/or objects, and of permissible or admissible object strings. This "makes it unnecessary for the program to compute all the possible object options (66 in the English grammar) for every verb."

"The medical subclasses are derived by examining the cooccurrence patterns in medical documents," a method that has proved useful by analyzing a number of sublanguages in order to extract word classes for structuring the textual information. This same method of analyzing distributional similarity of words in particular syntactic relations has proved advantageous in a different domain, i.e., in the analysis of standard dictionary definitions, with the purpose of formalizing their semantic informational content; the difference is that in these research projects the analysis is not manual, but is done mainly in a semi-automatic way. The usefulness and cost-effectiveness of applying similar methodologies to general English should be investigated and evaluated by the analysis of a very large corpus of texts.

Chapter 7 deals with the creation of the dictionary. It has been prepared manually because of the unsuitability of the information contained in large dictionaries in machine-readable form, judged not sufficiently detailed for processing sublanguage documents. However, a fully automatic dictionary coding procedure for lexical entries has been designed and implemented, taking advantage of the fact that in the medical vocabulary there is a substantial number of words of Greek and Latin extraction that exhibit a very high degree of morphological regularity connected to a high degree of semantic regularity, i.e., words that are morphologi-

cally and semantically transparent. This application of morphosemantic analysis to the creation of fully specified entries is certainly peculiar and best suited to a sublanguage, where complex words are often perfectly compositional and therefore the category and subclass can be inferred from the constituents; but it can find application also in general English. The description of the morphology program in all its steps is perhaps too detailed and rather obvious, being a rather simple scanning strategy looking for a match in both the suffix and prefix dictionaries. And the results do not seem extraordinary.

Chapter 8 deals with the parsing algorithm, top-down, syntax-driven; the syntactic structures are specified in the grammar in BN formulas, divided into types based on linguistic string analysis (Harris 1964). The parser is thus enabled to "use a small set of linguistically motivated procedures." The computer grammar of English has been adapted to the clinical sublanguage by means of the addition of new options and deletion of rare or unlikely usages in medical narrative.

Chapter 9 deals with some interesting issues in processing temporal information, using linguistic clues as adverbial expressions, verb tenses, coordinate and subordinate conjunctions, and "narrative time progression," i.e., the implicit temporal ordering of events suggested by their sequencing in the text. The result is a directed acyclic graph that represents the partial order of events that can be obtained from the text. The transitive closure of this graph adds further information, and more time relationships than those explicit in the text can be inferred. "The time graph generated can be used in conjunction with the database to answer time-related queries."

Chapter 10 has as its main topic the sublanguage grammar, whose main characteristic is to capture the typical co-occurrence restrictions. The patterns of word co-occurrence provide a set of semantic structures for representing subfield information. Everything in the sublanguage-grammar hypothesis is based on the application of the methods of descriptive linguistics to a corpus of texts in the selected subfield of science, thus establishing word classes on the basis of co-occurrence similarity. The extracted word classes have been found to correlate with recognizable semantic classes, and a clustering program obtained the main semantic classes only from the syntactically analyzed sentences. The grammatical structure of the sublanguage can be presented as a prototype sentence form, and the hierarchical organization of a sentence can be displayed in flattened form. "The overall structure of the format is given by English grammar," while particular sublanguage word classes in certain positions characterize particular subtypes of sentences with a particular semantic character, determining the sublanguage grammatical relations.

One of the characteristics of a study of this type is that although each text in a single specialized discipline "brings in some new features," these texts are "sufficiently similar so as to fit into an overall structural characterization. The repeated structures, implemented as information formats, are then a powerful tool for organizing the information in subfield texts."

There are in this section a number of repetitions and also some too-obvious considerations.

Chapter 11 again describes in detail, in other types of scientific texts, the technique of sublanguage analysis (performed manually). Once again the important step is the "establishment of equivalence classes of words based on equivalence of word environments (Harris 1963, p. 8)", after a partial normalization (by transformational decomposition) of the text, consisting of an operator-argument analysis of the sentences (Harris 1982). This part is too repetitive; the same concept is repeated again and again.

After storing the results in the database, queries can be mapped onto the set of sublanguage formats. The database is searched for sentences:

1. matching sentence types, or
2. matching members of the word categories.

In this way queries can be answered on a much more detailed level than that afforded by existing information retrieval systems (with simple keyword-identified material), given that it allows asking for a specific piece of information in the form of a relation between rather complex facts.

Appendix A would really seem more suitable for a system manual than for a book.

The book is addressed to the computational linguist. The approach described successfully combines practice with theory. Computational linguistics methods are not so frequently applied with success to a real-world problem that is not too narrowly restricted either in the lexicon or in the grammar. That it is done here is very positive. Even though the success is influenced by the fact that it operates in a specific sublanguage, it deals with the sublanguage in an extensive way without imposing too-severe limitations: the vocabulary is quite large, and the range of phenomena of language handled is quite reasonable. Of less interest for the linguist are the chapters or paragraphs where medical problems and considerations are dealt with too extensively.

Another drawback is that, perhaps for reasons of clarity, there is sometimes too much detail on too-specific steps, or there are too many repetitions of similar arguments in different chapters. This happens also because the linguistic methodology applied is the same to deal with different types of phenomena, e.g., for syntactic and semantic ones. The basic method is pattern matching. For many of the pattern-matching tasks, the program procedure is largely lists of patterns, i.e., lexical items that fail to conform to rules. But the overall value of the book outweighs the drawbacks of its repetitiveness.

With reference to principles of system design, the linguistic considerations of the problem are given prior-

ity over the technical or implementational point of view. In this regard, it cannot be criticized as designed in a too rigid way from the implementational viewpoint and not adaptable to new situations and unforeseen phenomena. The separation of data structures from the procedures and the modularity of the system are features that are essential to the extendability of the system to other domains.

In general, the work is a good example of:

1. the necessity of creating extensive lexicons, where "extensive" must be intended both in breadth (i.e., in quantitative terms) and in depth (i.e., from a qualitative viewpoint, as to the types of information associated with the entries);
2. the necessity of working with large textual corpora, both for obtaining linguistic data and for testing systems.

This is encouraging for a trend that is in recent years showing up, and having, for example, in Europe, great success also in projects sponsored by national and international organizations.

References

Harris, Z.S. 1963 *Discourse analysis reprints*. The Hague: Mouton.

Harris, Z.S. 1964 *String analysis in sentence structure*. The Hague: Mouton.

Harris, Z.S. 1982 *A grammar of English on mathematical principles*. New York: Wiley-Interscience.

Nicoletta Calzolari is a researcher at the Department of Linguistics of the University of Pisa and at the Institute of Computational Linguistics of CNR, Pisa. Her main research areas are in the field of computational lexicography and lexicology. Calzolari's address is: Dipartimento di Linguistica, Universitá di Pisa, Via S. Maria 36, I 56100 Pisa, Italy. E-mail: glottolo@icnucevm.bitnet

---

Information-based Syntax and Semantics. Vol 1: Fundamentals

Carl Pollard and Ivan A. Sag
(Carnegie Mellon University and Stanford University, resp.)

Stanford: Center for the Study of Language and Information, Stanford University, 1987, x + 227 pp.
(CSLI lecture notes no. 13)
Distributed by the University of Chicago Press
ISBN 0-937073-23-7, $39.95 (hb); ISBN 0-937073-24-5, $17.95 (sb)

Reviewed by
Edward P. Stabler, Jr.
University of Western Ontario

This book is an introductory text in linguistics, a very pleasant and readable introduction to head-driven phrase structure grammar (HPSG). HPSG will be of particular interest to computational linguists who have wanted to see situation semantics integrated with a unification-based phrase structure syntax. However, computational linguists should be warned, in the first place, that the book is truly introductory, focusing on the preliminaries to a sophisticated account of the language. Many of the serious problems to be faced are not discussed at all. In other places good, hard problems are posed, only to reward the reader's anticipation of a resolution with a promissory note about the forthcoming Volume 2. There are so many promissory notes at crucial places that it becomes clear that Volume 2 will be the real test of the framework. The two volumes are apparently organized not by topic, but by difficulty. All the difficult material on a whole range of topics—syntactic and semantic—is left for the second volume. The second warning for readers of this journal is that this book does not consider the computational properties of HPSG at all. No standard characterization of the HPSG-definable languages, no algorithm for unification or for parsing, and no complexity results are presented. As one expects in all but the most superficial or artificial approaches to human language, the grammar is incomplete in both its universal and its language-specific components. The grammatical principles, rules, and lexical entries are feature-based, where feature values can be complex (i.e., lists or sets), and computationally oriented readers might wonder how many features are needed and whether the set of possible values of syntactic features is finite, but we are not told.

Pollard and Sag describe HPSG as a "synthetic and eclectic" theory that draws on the insights of GPSG, LFG, GB, FUG, categorial grammar, situation semantics, and other approaches to language, which makes the title rather puzzling. How is HPSG "information-based"? Even when acquainted with the contents of this volume, I was still puzzled: I am not attuned to the Californian sense of "information." HPSG is "an information-based (or unification-based) theory of language," and it fundamentally regards "the objects that make up a human language as bearers of information within the community of people who know how to use them." To call HPSG information-based for *both* reasons, because it unifies the partial information structures called features *and* because utterances bear information, strikes me as a pun. But if smoke meaning fire is very much the same as *fire* meaning fire, then it is no doubt natural to think that *fire* meaning fire is quite a lot like a feature's being a partial specification, an element of a meet semilattice under subsumption where unification corresponds to the greatest lower bound.

HPSG includes a rather complex array of different kinds of propositions. We are given, in the first place, some basic facts about what types of linguistic objects there are. For example, there are two mutually exclusive types of signs, lexical and phrasal. We are told that "in general, such facts about relationships among types of linguistic objects are obvious, and we will not explicitly state them," but then in later pages we are infor-