ANALYZING LANGUAGE IN RESTRICTED DOMAINS:

SUBLANGUAGE DESCRIPTION AND PROCESSING

edited by
Ralph Grishman
Richard Kittredge

LEA

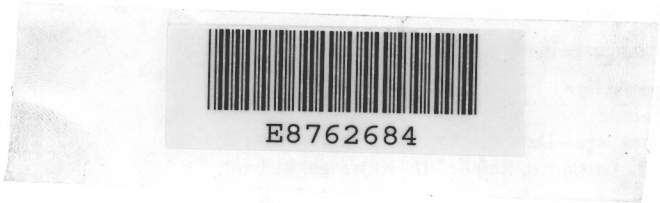
Analyzing Language in Restricted Domains:

Sublanguage Description and Processing

edited by

Ralph Grishman
New York University

Richard Kittredge
University of Montreal



LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1986 Hillsdale, New Jersey London

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Library of Congress Cataloging-in-Publication Data
Main entry under title:

Analyzing language in restricted domains.

Bibliography: p.
Includes index.
1. Sublanguage—Data processing—Addresses, essays, lectures. I. Grishman, Ralph. II. Kittredge, Richard, 1941–
P120.S9A53 1986 001.53'5 85-12850
ISBN 0-89859-620-3

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Analyzing Language in Restricted Domains:

Sublanguage Description and Processing

List of Contributors

Ralph Grishman
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012

Richard Kittredge
Département de Linguistique
Université de Montréal
C. P. 6128
Montréal
Quebec, Canada H3C 3J7

• • •

Robert A. Amsler
Bell Communications Research
445 South Street
Morristown, NJ 07960

Joan Bachenko
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

George Dunham
Laboratory of Statistical and Mathematical Methodology
Division of Computer Research and Technology
National Institutes of Health
Bethesda, MD 20205

Timothy W. Finin
Dept. Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA 19104

Eileen Fitzpatrick
695 Washington Street
New York, NY 10014

Carol Friedman
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012

Bonnie Glover
Logicon, Inc., Operating Systems Division
6300 Variel Avenue
Woodland Hills, CA 91367

Don Hindle
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

Lynette Hirschman
System Development Corporation, A Burroughs Company
Research and Development Division
PO Box 517
Paoli, PA 19301

Jerry R. Hobbs
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

John Lehrberger
56 89th Avenue
Chomedy, Laval
Quebec, Canada H7W-3G8

Elaine Marsh
Navy Center for Applied Research in Artificial Intelligence
Code 7510, Naval Research Laboratory
Washington, DC 20375

Christine Montgomery
Logicon, Inc., Operating Systems Division
6300 Variel Avenue
Woodland Hills, CA 91367

Naomi Sager
Courant Institute of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012

Jonathan Slocum
Microelectronics and Computer Technology Corporation
9430 Research Blvd.
Echelon Building #1, Suite 200
Austin, Texas 78759

Donald Walker
Bell Communications Research
445 South Street
Morristown, NJ 07960

Preface

Successful computer processing of natural language requires detailed knowledge of the language at many levels: lexicon, syntax, semantics, discourse, etc. Providing the linguistic knowledge for an entire language is truly a staggering task. In fact, no single human language has yet been fully described in a form usable by computers. Fortunately, many language processing problems are effectively restricted to the language used in a particular domain, where all this knowledge can be more readily obtained. The variety of language used in a given science or technology not only is much smaller than the whole language, but is also more clearly systematic in structure and meaning. These considerations have motivated linguists and computer scientists to collaborate in studying the properties of such specialized languages, which have come to be called *sublanguages*.

The term sublanguage has gained widespread acceptance among computational linguists only recently, and a lively debate is now underway concerning its proper definition. For most of the authors represented in this collection, the term suggests a subsystem of language that behaves essentially like the whole language, while being limited in reference to a specific subject domain. In particular, each sublanguage has a distinctive grammar, which can profitably be described and used to solve specific language-processing problems.

It is the overriding concern with grammatical subsystems and their use in language processing that distinguishes current sublanguage research from earlier investigations of domain-restricted language. Most of the work reported here was done by computational linguists faced with the need to write grammars that accurately and efficiently describe the language of science and technology in such fields as aeronautics, electronics, pharmacology and medicine.

Although the authors have all worked extensively in applied natural language processing, they approach sublanguage from the varied perspectives of artificial intelligence, information science and linguistics. Within artificial intelligence, sublanguage study offers a linguistic basis for describing many kinds of domain-dependent knowledge, such as type hierarchies. Furthermore, sublanguage grammars encode the kinds of domain-dependent semantic distinctions needed for sentence disambiguation, lexical transfer for machine translation, and many other tasks within applied systems. In the particular case of expert systems that communicate with specialists, sublanguage grammars are needed to guide the analysis of queries and the synthesis of explanations with the appropriate linguistic usage.

Within information science, sublanguage techniques have proven effective for the analysis, formatting, dissemination and retrieval of textual information. And it is within sublanguages that some progress can be envisioned in the difficult area of automatic abstracting.

In theoretical linguistics, the importance of sublanguage is only now beginning to be appreciated. Even if sublanguage grammars can be related to the grammar of the full standard language, sublanguages behave in many ways like autonomous systems. As such, they take on theoretical interest as microcosms of the whole language. In particular, the theoretical problem of relating linguistic form to communicative function comes into sharper focus when individual sublanguages are examined.

More than one of the authors in this volume has mentioned the relationship between sublanguage research and current investigations into language for special purposes (LSP). Although LSP work has historically had different practical goals, such as language pedagogy and document design, many possibilities exist for a symbiotic relationship. Sublanguage research stands to gain from the data and insights provided by LSP endeavors; in exchange it offers to strengthen the theoretical foundation of LSP through a better understanding of sublanguage grammars.

ORGANIZATION OF THIS VOLUME

The papers collected in this volume were presented at the Workshop on Sublanguage, held at New York University on January 19-20, 1984. The purpose of the workshop was to bring together leading North American researchers in the field of computational linguistics who have substantial experience in one or more of the following areas:

The general theory and description of sublanguage as a linguistic phenomenon,

the linguistic description and computer processing of particular sublanguages,

the automation of procedures for discovering and describing the particular syntactic, semantic and lexical properties of individual sublanguages.

Although most of the papers do not deal exclusively with only one of these areas, our grouping in this volume (and at the conference) reflects the primary concern of each author's contribution.

The first chapter in the theoretical section of this volume is Naomi Sager's keynote talk, "Sublanguage: Linguistic Phenomenon, Computational Tool." Sager's major concern is with using linguistic methods to reveal the close correspondence between grammatical organization of a sublanguage (including facts about word distribution) and the information-bearing properties of that same sublanguage. Her definitions of sublanguage and analysis methods are based on more than 15 years' experience in sublanguage research, especially in the fields of medicine and pharmacology. One of the significant contributions of the work reported here is a method for measuring the amount and complexity of information contained in a sublanguage. By way of example, Sager compares two sublanguages from the medical domain with respect to their information density and complexity, using such parameters as the frequency and distribution of operator words in sentences.

The second author concerned primarily with sublanguage theory is John Lehrberger. "Sublanguage Analysis" presents a rather abstract and algebraic view of the field, focusing primarily on the relation of individual sublanguages to the standard language. One of Lehrberger's primary analysis tools is the paraphrase relation, and the analysis of the ways in which sublanguage sentences may be paraphrased in the standard language. Another concern here is with the notion of "formattability" of sublanguages (and its use as a means of comparing sublanguages). A third topic discussed is the distinction between natural and artificial or constructed sublanguages, with particular concern for the sublanguages that are found in the domains of mathematics and computer science. Lehrberger's discussion of these topics poses a large and intriguing set of questions that should be of special interest to theoretical linguists.

Another chapter with a theoretical perspective is "The Status of Telegraphic Sublanguages" by Fitzpatrick, Bachenko and Hindle. The primary question this chapter raises is whether sublanguages can be considered as relatively independent syntactic systems with internal consistency, or whether sublanguage syntax must be described in reference to the standard language. The authors use two phenomena in the telegraphic sublanguage of equipment casualty reports to support the former hypothesis. This contrasts with the more frequently espoused view, which can be discerned in a number of other papers in this volume. Clearly this question is of both theoretical and practical importance and constitutes one of the major issues to which we return later.

The fourth and final paper from the theoretical session is Jerry Hobbs' "Sublanguage and Knowledge." This chapter represents a point of view on sublanguage markedly different from most of the other papers in this volume. It does not take as primary data linguistic usage, but rather the knowledge behind that usage. In particular, Hobbs shows how the linguistic presuppositions that must be used to understand certain sentences can guide the selection of facts to be added to a knowledge base. The knowledge base is part of an information retrieval system that matches the semantic structure of queries to relevant parts of a semantically coded medical reference text. Hobbs argues that the kind of axiomatization approach he proposes is necessary for stating critical relationships between facts (in the knowledge base). Although some axioms of knowledge have direct counterparts in terms of sublanguage selection statements, certain more complex statements seem to require the expressive power of first-order logic, which has no obvious equivalent in sublanguage description. Furthermore, argues Hobbs, sublanguage description is often complicated by cases of metonymy (see also the chapter by Hirschman), which must be regularized by an appeal to common sense or expert knowledge.

The second, and largest, group of papers are united in their primary concern with linguistic data in the context of applied language processing systems. Although important theoretical questions are raised in this section as well, their overriding preoccupation with practical problems has dictated this grouping.

First among the "applied" papers is the report by Walker and Amsler on "The Use of Machine-Readable Dictionaries in Sublanguage Analysis." Here, the authors report on a system that determines the subject domain of newspaper articles by using the semantic codes available in a large machine-readable dictionary. To the extent that the article represents a single domain, it is usually possible to determine the domain (and hence disambiguate most polysemous words) by statistical procedures. Extensions of this approach to technical sublanguages and refinements in the semantic marking system are discussed. Although many problems remain in providing and refining the data necessary to apply this approach on a high-volume basis, it is almost certain that near-term applications exist.

Carol Friedman's "Sublanguage Text Processing – Application to Medical Narrative" gives a very detailed account of the use of distributional techniques for setting up a sublanguage grammar and for converting text (in the domain of clinical records) to a structured data form. Each sentence pattern is mapped to a separate information format, with the result that the formatted natural language data can be queried in relational database systems. This chapter will be particularly useful for readers who are not already familiar with the details of the method of information formatting practiced by NYU's Linguistic String Project.

One chapter included in the applied section is actually a generalization of work carried out in at least two domains. Elaine Marsh reports on "General Semantic Patterns in Different Sublanguages," drawing on the experience of the NYU group in analyzing medical discharge summaries, as well as her more recent participation in a project for processing equipment casualty reports. After giving a detailed account of the semantic patterns found in reports on equipment failures, she compares those patterns with the medical reports on human health maintenance, aiming to describe a higher level "domain of failures." She also notes some similarities to a special-purpose programming language used to describe requirements for automatic test equipment. Marsh concludes by comparing the semantic patterns of sublanguages with the frame and script structures used in knowledge representation work. It may be interesting for the reader to compare Marsh's view of sublanguage structures as knowledge representations with Hobbs' claims, cited previously.

In "A Sublanguage for Reporting and Analysis of Space Events," Christine Montgomery and Bonnie Glover describe a sublanguage that functions as a medium for communicating both factual reports about space vehicle events (launches, re-entries, etc.) and analytical comments about those events. These two levels of space event discourse, the report level and meta level of commentary on the reports, have distinct linguistic properties, giving rise to the question of whether one or two sublanguages are present. The authors also deal extensively with the problem of representing different levels of meta-event in the space event reports, where different degrees of confidence may be assigned to statements, depending on whether they constitute direct or indirect reports.

Tim Finin's chapter on "Constraining the Interpretation of Nominal Compounds in a Limited Context" is a continuation of work carried out on the UNCLE system, in which semantic interpretations were built for nominal compounds in a technical domain. Here, Finin introduces the use of discourse context as a way of filtering the many possible interpretations for compounds. Whereas earlier work focused on three types of local rules that use lexical information to interpret nominals, Finin now proposes to add rules that treat nominals as referring expressions and use discourse constraints on reference to establish the most likely connection between previously mentioned objects and a newly introduced object.

George Dunham presents an unusual sublanguage case study in "The Role of Syntax in the Sublanguage of Medical Diagnostic Statements." In this written sublanguage, typical syntactic constructions are limited in usage, but certain sublanguage-specific devices are prominent. The sublanguage is telegraphic (lacking true verbs, for example) and makes heavy use of nominalizations. This case is particularly interesting, Dunham says, because it shows how the typical role of syntax as a framework for semantic interpretation is

taken over by pragmatically based discourse rules. He goes on to sketch some of the semantic rules for this sublanguage in a higher-order logic.

The third topic area of this book involves discovery procedures for sublanguages and sublanguage properties. First, Jonathan Slocum presents his quantitative work on identifying and adapting to a sublanguage's syntactic properties. Slocum compares two distinct varieties of text (German technical manuals and sales brochures) by measuring the number of rule applications made by the analysis grammar of a large automatic translation system. Measures based on automatic rule applications are also used to check the amount of variation among different samples of the same type of text. Slocum proposes to use the characteristic frequency of rule applications in each type of text to assign weights to syntactic rules differently when processing different sublanguages, thus increasing the likelihood that the preferred parse for the analyzer is the correct one. These same data on frequency can serve as a basis for dynamically "guessing" the type of text in which an analyzer begins to parse.

A more general approach to the discovery problem is outlined by Lynette Hirschman in "Discovering Sublanguage Structures." Hirschman considers the "portability problem" (adapting existing systems to new domains) to be critical, now that limited-domain systems have had some initial success. If some of the characteristics of a new sublanguage can be discovered automatically, then the considerable effort of describing a new domain can be reduced. Hirschman reviews some recent work in the automatic acquisition of sublanguage semantic classes using clustering techniques and points out some of the unsolved problems and limitations of a statistical approach. She discusses the problem of circularity, which arises from the attempt to automate discovery procedures (i.e., we need the grammar to fully automate the discovery process, whose goal is to establish the grammar). The solution must involve "bootstrapping" through successive iterations. It is important to note that Hirschman's view of sublanguage grammar is much more semantic than Slocum's, concerning as it does semantic word class co-occurrence.

Some Major Issues in Sublanguage Theory and Applications

A number of theoretical and practical issues cut across the grouping just described. Perhaps the most general of these is how sublanguages, viewed as linguistic systems, relate to the much larger system of standard language. Only Lehrberger discusses this question directly and at length. But several authors, including Sager, Fitzpatrick et al., Marsh, Slocum and Hirschman, provide some relevant insights. Most authors take the view that sublanguages, despite the occurrence of specialized structures, can be viewed as derivable (through general types of deletion, etc.) from the standard lan-

guage, at least on the syntactic level. If Fitzpatrick et al. are correct in their opposing view, then organizational principles for each sublanguage must be considered as possibly specialized. This would give a more pessimistic outlook on the possibility of "porting" sublanguage systems to new domains.

A second theoretical problem with practical consequences for computational linguistics is the relationship between sublanguage syntax and semantics. In many of the sublanguages discussed, syntactic information seems to be too impoverished to play a primary role in analysis (cf. Dunham, Finin, Hobbs). The problem of compound nominals is a case in point. These authors stress the need for using not only semantic class information, but also information from discourse context, to establish the proper meaning relationship between words found in text. Many of the other authors (cf. Sager, Marsh, Fitzpatrick et al., Friedman), while not discounting the interest of other information, seem to indicate that the data from word distribution (essentially syntactic information) can reveal enough semantic information (through word grouping) to clarify most of the problems of meaning relationship. To some extent, the conflicting views may be a function of the authors' different processing goals.

DISCUSSION QUESTIONS

During the discussion period of the conference, participants were asked to express their views on a number of questions (selected in advance) of practical importance.

1. There was general agreement that, except for highly circumscribed sublanguages such as weather reports, we are currently not able to obtain correct sentence analyses with high reliability. (By this, we mean the correct determination of all operator-operand and host-adjunct patterns for at least 90% of the sentences in a text.) Reaching this goal is important if we are to develop useful applications involving more complex sublanguages. Is there any clear evidence of progress in this direction?

The discussion of this question raised a number of points:

There are a growing number of sublanguages that can be analyzed with some success, if not high reliability—medical discharge summaries, equipment failure reports, maintenance manuals, intelligence reports. Slocum, in particular, cites performance in the 80% range on telephone system manuals.

There are problems of scale in achieving this objective. Substantial work in analyzing the lexical, syntactic and semantic properties of sublanguages has been required to achieve even the current measure of success. Much more work will be needed to develop these systems further for complex domains.

The steady improvement in "question-answering systems" (natural language database interfaces) gives hope of similar progress in sublanguage text analysis; however, text analysis is more difficult, and so success here will be slower in coming.

Even when no complete sentence analysis can be obtained, useful processing (e.g., generation of information formats) can be done using sentence portions that can be analyzed.

Progress in sublanguage analysis cannot be gauged by success in parsing alone. There has been much greater progress in the last few years, for example, in understanding how to represent the information in a sublanguage text in a database.

2. What information about a sublanguage and its domain is needed to attain the goal of reliable correct sentence analysis?

The general answer of the discussants was "everything," including the lexical, syntactic, semantic, and discourse properties of the sublanguage. The papers of the workshop focused largely on syntactic and semantic properties; much of the exchange in the discussion session was on the use of discourse properties. Among the issues raised were: (a) the need for discourse analysis to handle anaphora; (b) the need for an understanding of discourse structure in order to generate text; (c) representation of different levels of discourse structure: paragraph structure, "top level" macrostructure, etc.; (d) different types of discourse constraints: continuity of focus, parallelism between sentences, etc.

It was agreed that although discourse properties will be important for sublanguage analysis, they are as yet poorly understood by comparison with syntactic and semantic properties. In particular, it is unclear how to use feedback from discourse analysis in order to increase the reliability with which we understand individual sentences. The one suggestion in this regard was that texts with different sections in different styles (e.g., manuals with descriptive sections and sets of instructions) have different grammars for different styles.

The discussion also turned to the issue of which properties might be shared by different sublanguages. It was suggested that texts concerning the same domain should share vocabulary and semantic patterns, while those involving the same function (e.g., telegraphic messages, sets of instructions) should share syntactic characteristics.

3. Can sublanguage characteristics be discovered in an automatic or semi-automatic fashion for a new domain?

The chapters by Slocum and Hirschman describe initial efforts at developing such discovery procedures. One problem brought out in the discussion of these papers concerns circularity: one needs good sentence analyses in order to compute sublanguage properties, and one needs to know sublanguage

properties in order to obtain good sentence analyses. It was suggested that this circle might be broken by identifying "unproblematic" sentences within a corpus—say, those that can be unambiguously parsed without sublanguage information—and using them to start the discovery process.

The discussion of discovery procedures also brought out the trade-offs between simpler constraints and more powerful ones. Simpler constraints, such as word selection constraints, are relatively easy to discover for new domains and are also straightforward to represent and use in processing systems. But the representations and processing strategies used for these simpler constraints are too weak to deal with such phenomena as sublanguage-dependent presupposition. Presupposition requires more complex representation schemes (such as the first-order logic proposed by Hobbs) and computationally more expensive processing strategies. Much more experience is needed before the roles and relative advantages of various formalisms can be assessed.

ACKNOWLEDGMENT

The Workshop on Sublanguage Description and Processing, at which the papers in this volume were presented, was supported in part by a grant to New York University from the National Science Foundation, Division of Mathematics and Computer Science (grant NSF-MCS-83-01197).

Contents