#56

1 Sublanguage: Linguistic Phenomenon, Computational Tool

Naomi Sager
Linguistic String Project
Courant Institute of Mathematical Sciences
New York University

ABSTRACT

A sublanguage is characterized by distinctive specializations of syntax and the occurrence of domain-specific word subclasses in particular syntactic combinations. The Linguistic String Project of New York University has studied several sublanguages in detail over the past 15 years and developed computer methods for obtaining the relevant word classes and relations from samples of syntactically analyzed domain sentences. The methods are illustrated in application to articles in the lipoprotein literature. It has also proved possible to measure such features as the quantity, density, and complexity of information in the sentences of contrasting sublanguages.

The special word-classes and relations of a particular sublanguage provide the basis for a variety of natural language processing applications that would not be practicable in the language as a whole. For example, it is possible (with difficulty) to process full texts in a sublanguage and convert the free-text information into a structured form suitable for fact retrieval and data summarization. The information structures arrived at in such processing are similar in certain respects to data models used in database management systems, and suggest the possibility of adapting such systems for the management of natural language-derived databases.

INTRODUCTION

Without language there would be no culture, no means of recording and transmitting the knowledge necessary to survival, nor the means to register, hence to develop, thought. Here, we consider the ability of language to carry information. In the use of language within specific (primarily technical) do-

From Analyzing Language in Restricted Domains
(R. Grishman and R. Kittredge, eds.)
Lawrence Erlbaum, 1986.

mains (Kittredge & Lehrberger, 1982), it is a highly developed, highly specialized tool, an essential part of the very technology its role is to record. Yet in considering the use of language within specialized areas of knowledge, where it is stripped of literary niceties and totally dedicated to the communication of information, we also have the opportunity to witness in a relatively pure form some of the mechanisms by which language fulfills its broader function of transmitting all verbal culture.

For example, the grammar of the language as a whole has numerous instances of compacting by deletion ("zeroing") where a shorter, less regular sentence form is created that is paraphrastic to the fuller form. (*It is considered to be harmless* ↔ *It is considered harmless*). Various sublanguages make extensive use of this process, presenting so-called ill-formed input to language processing programs, that is, syntactic forms that are not found as sentences in the language as whole. In a sublanguage text it is clear that a dropped word in such a form is the one that would occur in the corresponding position in the fuller sentence using the same words (e.g. *patient* as subject of the occurrence *is on folic acid* in the sublanguage of patient documents). Because the word classes and patterns of word-class occurrence are more narrowly defined in a sublanguage, with clear semantic values attaching to each syntactic position within a form, we see more sharply than in the language as a whole how departures from regularity are made possible by the existence of linguistic expectations as to which words combine to make well-formed ocurrences in the language of the discourse.

Thus, while our concern in sublanguages is with the restricted use of language in narrow domains, our studies also cast light on processes that are at work in the language as a whole.

FUNDAMENTAL CONCEPTS

Definition of Sublanguage

What is a sublanguage; how is it defined? Informally, we can define a sublanguage as the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation (Bross, Shapiro & Anderson, 1972). This corresponds to the way a language is identified as the mutually understandable verbal communication of some community, often geographically bounded. Faced with the difficult problem of defining the boundaries of a given sublanguage, we may take comfort in the fact that even whole natural languages sometimes have fuzzy edges, seen in the shading of dialects across border areas and in the uncertainties as to what constitutes a well-formed sentence within a well-defined language like English in some cases.

A technical approach to language definition defines a language as the set of all sentences than can be generated by its associated grammar. In the case of artificial languages, the grammars can be stated as a set of formal definitions. In the case of natural languages, we may aim for a formal generating grammar similar to those used to define artificial languages; but whatever the formal constructs are, they must generate just those sentences that a native speaker accepts as belonging to the language.

A formal characterization of a whole language requires at least two levels of word classification. Gross syntactic formulas for well-formed sentences are obtained in terms of categories such as noun, verb, adjective (defined in part on morphological grounds); a more refined syntactic characterization uses grammatical subclasses, such as singular, plural, count-noun, etc. The subclasses are grammatical because, regardless of the particular words that satisfy a gross syntactic formula, whether they make a sensible utterance or not, an educated speaker will by and large reject as incorrect an occurrence that violates the stated well-formed combinations of the grammatical subclasses in a given syntactic formula (*The book was interesting, *Book was interesting, *The book were interesting*).

But, as is well known, this two-level grammatical characterization still leaves untreated a major feature of natural language, the fact that for a given sentence form, say a simple NOUN-VERB-NOUN (subject-verb-object) form, not every sentence obtained by substituting a class member for the class symbol in the syntactic formula constitutes an equally acceptable sentence of the language, even if the substituted words satisfy all grammatical subclass constraints (*John loves Mary, Misery loves Company, Company loves misery, Clouds love chocolate, etc.*). This is the phenomenon known linguistically as *selection.*

The distinguishing feature of sublanguage is that over certain subsets of the sentences of the language the phenomenon of selection, for which rules cannot be stated for the language as a whole, is brought under the rubric of grammar. In a sublanguage, selectional word classes have relatively sharp boundaries, reflecting the division of real world objects into classes that are sharply differentiated in the domain. Refined syntactic formulas stated in terms of these classes reflect the types of relations the objects named in these classes can have to each other and thereby provide a semantic characterization of the discourse in the given domain, using grammatical methods of description.

Sublanguage Grammar

The restrictions on word combination on the sublanguage level are grammatical in the sense that a speaker of the sublanguage can say with reasonable assurance whether a given sentence is a *possible* utterance in a discourse in the given area, independently of whether it is true or false. The more structured the knowledge in a given area, the sharper the constraints on what can reasonably be said. Hence, the sublanguage phenomenon is most marked in scientific and technical areas, where the discourse proceeds on the basis of some body of already well-established laws or practices.

Although the sentences of a sublanguage are a subset of those in the parent language, the sublanguage grammar is not a subset of the grammar of the parent language and in fact intersects it, as Harris (1968) points out in the first treatment of sublanguage as a linguistic phenomenon. Some rules (in terms of the special word classes of the sublanguage) are not satisfied by sentences that are part of the parent language (e.g., to use Harris' example, *Hydrochloric acid was washed in polypeptides* is not a sentence of biochemistry though still a sentence of English), while some rules of the parent language do not apply in the sublanguage (e.g. colloquial forms defined for the whole language may not occur in a science sublanguage).

Science sublanguages may be studied using a corpus drawn from the published literature in the field. In this case, the sublanguage sentences are indeed a subset of the sentences of the parent languages. However, the notion of sublanguage has also been applied to informal communications in technical areas, where the "sentences" may be so shortened for rapid communication that they may not qualify as grammatical utterances in the parent language. A modificiation of the orginal definition of *sublanguage* in Harris (1968) may have to be developed to cover these cases.

METHODS OF ANALYSIS

Sublanguage Word Classes

The first step in obtaining a sublanguage grammar is to determine the domain-specific noun classes, and, close upon this task or in conjunction with it, the verb and other linguistic operators that co-occur with them in elementary structures. In some domains, the noun classes are virtually given by established classifications within the science or by the organization of data in a database to which sublanguage sentences (in this case, questions) are to be addressed. It is also possible to obtain the sublanguage word classes by grouping the words of sublanguage texts into classes on the basis of their occurrence in similar environments. For example, it has been shown that the domain-specific noun and verb classes can be obtained simultaneously by a clustering program that operates on a sample of transformationally analyzed sublanguage sentences (Hirschman, Grishman & Sager, 1975). The illustration of sublanguage analysis that is presented here utilized a computer-aided method of obtaining a sublanguage grammar where the main sublanguage noun classes were provided by a scientist working in the field.

The domain of this sublanguage is lipoprotein kinetics. Lipoproteins are the molecules that serve as the transport system for cholesterol and other lipids in the body. The journal literature from which the texts of this sublanguage are drawn treats such topics as cholesterol turnover and the metabolic pathways of the lipoproteins under various conditions of diet, disease, and other factors, as studied both in humans and experimental animals. We undertook this sublanguage study with the aim of arriving at information structures that would help to organize portions of the literature for use by scientists engaged in mathematical modeling in this area (Sager & Kosaka, 1983).

Co-occurrence Patterns

To obtain the domain-specific co-occurrence patterns, we first manually analyzed a sample of the textual material and entered the syntactic sentence trees into the computer in parenthesized list form. The type of analysis used was a modified form of *operator-argument* grammar (Harris, 1982). In the operator-argument sentence tree, every node corresponds to a word or phrase of the sentence (sometimes restored from a zeroed occurrence or morphologically transformed), and the dominance of one node over others immediately below it has the interpretation of predication.

An example of an operator-argument tree for a sentence from this literature (Brown, Kovanen & Goldstein, 1981) is shown in Fig. 1.1. The sentence reads: "For export of triglycerides and cholesterol, the liver incorporates the lipids into VLDL (300–800A)." VLDL stands for *V*ery *L*ow *D*ensity

```
BKG 2.9.1  FOR EXPORT OF TRIGLYCERIDES AND CHOLESTEROL,
           THE LIVER INCORPORATES THE LIPIDS INTO VLDL (300-800A).
```

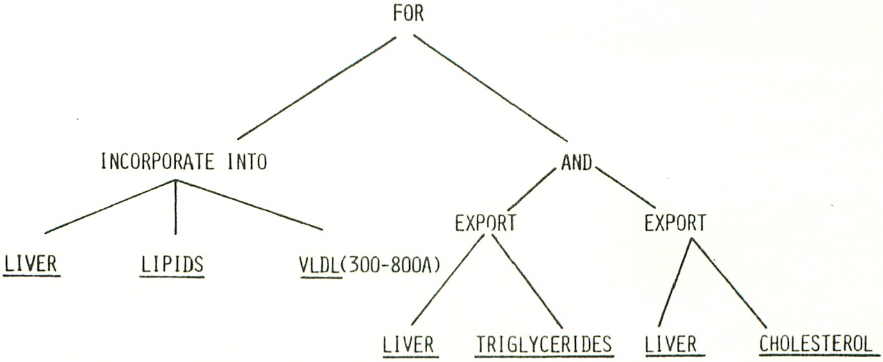


FIG. 1.1 Lipoprotein sentence tree.

*L*ipoproteins. The tree has three operator occurrences whose arguments are sublanguage nouns, corresponding to the elementary sublanguage sentences: (1) *the liver incorporates lipids into VLDL*; (2) *the liver exports triglycerides*; and (3) *the liver exports cholesterol*. The first of these operators with its arguments is connected by *for* to the other two, which are joined by *and*. Thus it is asserted that one activity of the liver is *for* (i.e. in the service of) another: It incorporates lipids into the lipoprotein VLDL in order to export them. In this type of grammatical analysis the syntactic relations obtained for a sentence correspond to the informational relations among the words of the sentence.

Note that in the operator-argument tree of Fig. 1.1 two zeroed occurrences of *liver* have been restored, one (along with the verb *export*) due to zeroing under the conjunction *and*, and one due to the nominalization of *export* and its occurrence (under *for*) as a sentence adjunct to the main clause where *liver* is the explicit subject. These zeroings are grammatically reconstructable. By contrast, *the lipids*, which in this sentence is a referential noun phrase referring to *triglycerides and cholesterol*, has not been replaced by these nouns. A full informational expansion would have four operator occurrences, cross linked (*For export of triglycerides, the liver incorporates triglycerides into VLDL*, and *For export of cholesterol, the liver incorporates cholesterol into VLDL*). This sublanguage-dependent referential resolution cannot be assumed in the analysis that is to produce the sublanguage grammar. Nor, as it turns out, is it needed for producing the sublanguage grammar, since *lipids* (as classifier) and *triglycerides* and *cholesterol* are all in the same sublanguage noun class with regard to their occurrence under particular classes of operators.

It will be noticed that in Fig. 1.1 the sublanguage nouns (*liver, lipids, VLDL, triglycerides, cholesterol*) occupy the bottom-most nodes of the operator-argument tree. It is a striking fact that when sublanguage text sentences are syntactically analyzed in operator-argument form, the sublanguage vocabulary is always found at the bottom nodes of the tree. Fig. 1.2 illustrates this fact by showing the location of the sublanguage vocabulary in the operator-argument trees of three consecutive sentences from Brown et al. (1981). (The second tree of Fig. 1.2 corresponds to the sentencee tree in Fig. 1.1; the articles and parenthesized modifier are not represented.) The squares and circles show the placement of words of the sublanguage vocabulary: white squares for the class of ORGAN/CELL words, black squares for the class of LIPIDS, white circles for LIPOPROTEINS, and black circles for ENZYME words. Nouns of the sublanguage vocabulary always occur in the bottom-most tree nodes and never in the top-most nodes. Sublanguage-specific verbs or predicates are the immediate operators on sublanguage nouns.

Having the operator-argument trees in the computer, we entered separately a list of the main sublanguage noun classes and their members, some of the main ones being:

AP	apoprotein	e.g.	apoprotein B, apoprotein C
LI	lipid	e.g.	cholestrol, triglyceride, cholesterol ester
LP	lipoprotein	e.g.	chylomicron, VLDL (very low density lipoprotein), HDL (high density lipoprotein)
EZ	enzyme	e.g.	LPL (lipoprotein lipase), LCAT
OR	organ/cell	e.g.	liver, peripheral tissue, fibroblast cell
HO	hormones	e.g.	insulin
RC	receptors	e.g.	LDL receptors

It was then straightforward for a program to substitute class names for class-member occurrences in the sentence trees and to make a table of the operator-argument tuples, sorted alphabetically by operator or by the sublanguage class of each argument.

Figure 1.3 shows a portion of a sublanguage operator-argument table generated from our data. In the first column are the operators, mostly verbs, or verbs in nominal form, that relate ORGAN/CELL words to LIPID words; for example, for ARG1, an ORGAN-CELL word, we have the operator-argument tuples corresponding to the elementary sublanguage sentence oc-

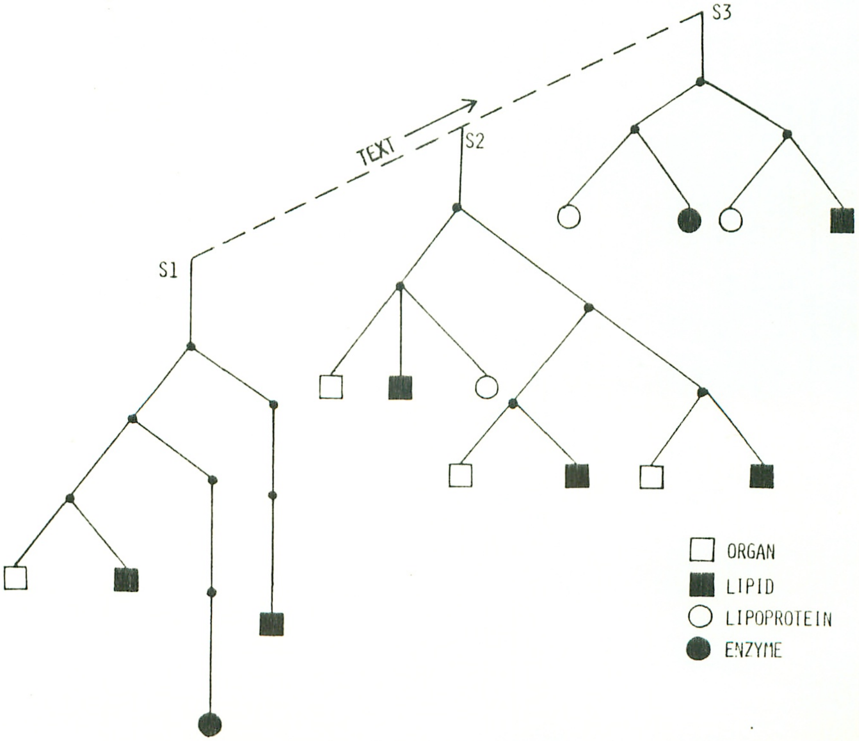


FIG. 1.2 Location of sublanguage vocabulary in sentence trees.

OPERATOR	ARG1	ARG2
DEMAND OF-FOR	/OR = LIVER	/LI = CHOLESTEROL
EXPORT OF	/OR = LIVER	/LI = CHOLESTEROL
TAKES UP	/OR = LIVER	/LI = CHOLESTEROL
SYNTHESIZES	/OR = LIVER	/LI = CHOLESTEROL
RELY ON	/OR = EXTRAHEPATIC CELLS	/LI = CHOLESTEROL
USE	/OR = CELL	/LI = CHOLESTEROL
EXPORT OF	/OR = LIVER	/LI = TRIGLYCERIDES

FIG. 1.3 Portion of operator-argument table.

currences: *The liver has a demand for cholesterol, The liver exports cholesterol, The liver takes up cholesterol, The liver synthesizes cholesterol, The liver relies on cholesterol, The cell uses cholesterol, The liver exports triglycerides,* and so on. The same argument word-classes appearing in the opposite order define a related class, as shown in Fig. 1.4, where the data is drawn from the same file as in Fig. 1.3. Here the operators are *are delivered to, are transported to, originate in, excretion into, is delivered to, the exit of-from.*

We can combine operator classes that have the same number and type of arguments but a different order of arguments by establishing a primary order and marking those members that require a different order as inverses. Thus, a single ORGAN-LIPID operator class is obtained from the data illustrated in Figs. 1.3 and 1.4.

The computer-generated tables of subclass co-occurrence patterns provide the material for the most elementary (linguistically, the *kernel* or K-level) portion of the sublanguage grammar. From the lipoprotein kinetics literature

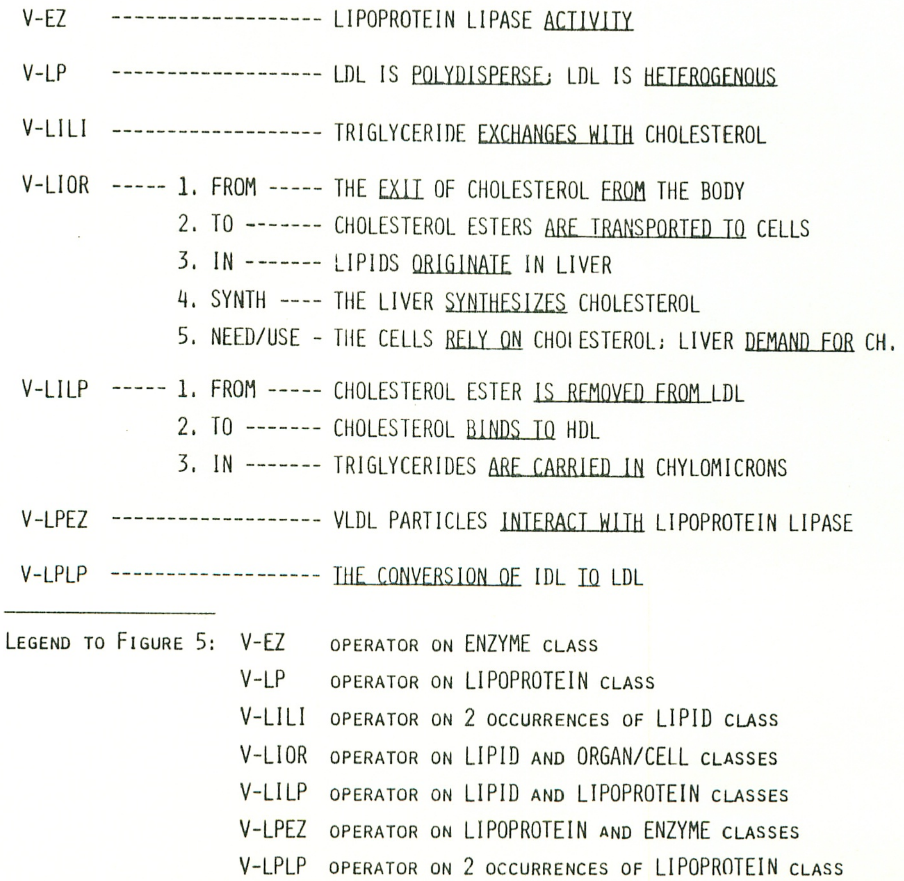
OPERATOR	ARG1	ARG2
ARE DELIVERED TO	/LI = TRIGLYCERIDES	/OR = ADIPOSE TISSUES
ORIGINATE IN	/LI = LIPIDS	/OR = LIVER
ARE TRANSPORTED TO	/LI = CHOLESTERYL ESTERS	/OR = CELLS
EXCRETION INTO	/LI = STEROL	/OR = BILE
IS DELIVERED TO	/LI = CHOLESTEROL	/OR = LIVER
THE EXIT OF-FROM	/LI = CHOLESTEROL	/OR = BODY

FIG. 1.4 Operator-argument table (continued).

as represented in our file, we have obtained some 40 kernel-types, where a kernel type (K-type) is defined as a sublanguage operator class and its argument classes in terms of about a dozen sublanguage noun classes. Some of these kernel level relations are illustrated in Fig. 1.5. Figure 1.5 also illustrates that, within each operator class, semantically distinct subsets can be distinguished.

Sublanguage Sentence Types

While the co-occurrence patterns of sublanguage-specific word classes (the K-types of the sublanguage grammar) are the primary means of distinguishing and characterizing the sublanguage, these forms are often a part of larger sentential structures that are regular throughout the sublanguage texts, and they constitute a second level of sublanguage description.



```
V-EZ     ------------------- LIPOPROTEIN LIPASE ACTIVITY

V-LP     ------------------- LDL IS POLYDISPERSE; LDL IS HETEROGENOUS

V-LILI   ------------------- TRIGLYCERIDE EXCHANGES WITH CHOLESTEROL

V-LIOR   ----- 1. FROM ----- THE EXIT OF CHOLESTEROL FROM THE BODY
               2. TO ------- CHOLESTEROL ESTERS ARE TRANSPORTED TO CELLS
               3. IN ------- LIPIDS ORIGINATE IN LIVER
               4. SYNTH ---- THE LIVER SYNTHESIZES CHOLESTEROL
               5. NEED/USE - THE CELLS RELY ON CHOI ESTEROL; LIVER DEMAND FOR CH.

V-LILP   ----- 1. FROM ----- CHOLESTEROL ESTER IS REMOVED FROM LDL
               2. TO ------- CHOLESTEROL BINDS TO HDL
               3. IN ------- TRIGLYCERIDES ARE CARRIED IN CHYLOMICRONS

V-LPEZ   ------------------- VLDL PARTICLES INTERACT WITH LIPOPROTEIN LIPASE

V-LPLP   ------------------- THE CONVERSION OF IDL TO LDL
```

```
LEGEND TO FIGURE 5:  V-EZ     OPERATOR ON ENZYME CLASS
                     V-LP     OPERATOR ON LIPOPROTEIN CLASS
                     V-LILI   OPERATOR ON 2 OCCURRENCES OF LIPID CLASS
                     V-LIOR   OPERATOR ON LIPID AND ORGAN/CELL CLASSES
                     V-LILP   OPERATOR ON LIPID AND LIPOPROTEIN CLASSES
                     V-LPEZ   OPERATOR ON LIPOPROTEIN AND ENZYME CLASSES
                     V-LPLP   OPERATOR ON 2 OCCURRENCES OF LIPOPROTEIN CLASS
```

FIG. 1.5 Some lipoprotein sublanguage relations.

For example, in the medical sublanguage of patient documents, we can distinguish a large number of co-occurrence patterns in the elementary subject-verb-object and host-modifier relations. These patterns combine to form a small number of inclusive sentence types that represent the main types of events in the course of a patient's illness and treatment (Friedman, this volume). The patient document can be seen to be composed of a sequence of these sentence-type occurrences with explicit linguistic connectives between them and with associated time expressions that provide the chronology of the events (Hirschman & Story, 1981). The sublanguage grammar, stated up to the level of these event-representing sentence types, provides the structural units of the discourse.

In science sublanguages, where the linguistic data are given in journal articles or other published texts, the existence of characteristic sentence types involving more than the kernel-level co-occurrence patterns is even clearer. In the first such sublanguage that we analyzed in some detail, using texts from the literature on the mechanisms of pharmacological action of digitalis and digitalis-related drugs, the kernel-level elementary sentence types had characteristic adjuncts stating certain conditions on the observations, often also quantity operators, and a distinguished noun class (the drug words) that occurred with a causal verb operating on one of the kernel sentences (Sager, 1972). Texts were found to be composed primarily of sequences of such occurrences (sometimes in reduced form) under a complicated structure of conjunctions and co-reference.

The structure of the larger sentence types of the lipoprotein kinetics sublanguage shown schematically in Fig. 1.6, involves the addition to the K-types of "slots" for sublanguage material that occurs variously in adjunct status (i.e. as modifiers) or as the subject of operators on the kernel sentences. Sublanguage word classes on this level include classes for diseases, elements of diet, human/animal subjects and particular physiological variables.

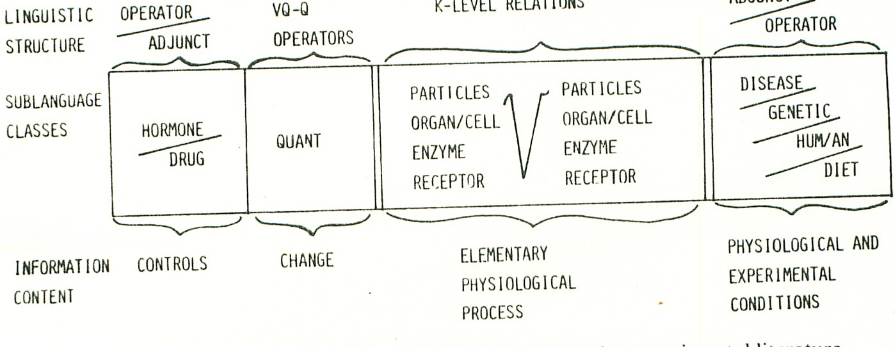


FIG. 1.6 Schematic of sentence types in lipoprotein kinetics experimental literature.

These classes occur linguistically most often as adjuncts and can be interpreted in such texts as given conditions for the observations reported. However, where the reported investigation varied these conditons, members of these classes may occur as operators, as illustrated in the text analyzed in the next section. Several other sublanguage classes, principally the hormone and drug class, also occur linguistically as either adjunct or operator, but more frequently as operator, corresponding to their role as controls on the elementary physiological processes described by the K-level relations.

In addition, there are quantity operators on nouns and verbs that may introduce an intermediate structure between a sublanguage operator (or English operator) and its arguments. Further, sentences in this sublanguage are rich in local modifiers, which are introduced as adverbial phrases or as relative clauses (mainly, reduced) and compound nouns. The main logical structure of each sentence is most clearly seen when local modifiers are associated with the element they modify and are not made a part of the main branching structure of the sentence tree.

SUBLANGUAGE SENTENCE TYPES IN DISCOURSE

To illustrate the role of sublanguage sentence types in an informational representation of sublanguage texts, the first paragraph of an abstract from the lipoprotein literature (Grundy, 1975), which is shown in Fig. 1.7, is shown in analyzed form in Fig. 1.8.

> Studies were carried out on the effects of polyunsaturated fats on lipid metabolism in 11 patients with hyptertriglyceridemia. During cholesterol balance studies performed in eight patients, the feeding of polyunsaturated fats, as compared with saturated fats, caused an increased excretion of endogenous neutral steroids, acidic steroids, or both in most patients. Increases in steroid excretions were marked in some patients and generally exceeded the decrement of cholesterol in the plasma compartment. The finding of a greater excretion of fecal steroids on polyunsaturated fats in hypertriglyceridemic patients contrasts to the lack of change in sterol balance previously reported for patients with familial hypercholesterolemia: however, other workers have found that polyunsaturated fats also enhance steroid excretion in normal subjects.

FIG. 1.7 Sample Text.

In Fig. 1.8 the entire paragraph is seen to consist of successive occurrences of a particular case of the lipoprotein sentence type, as follows:

DIET | V | VQ ‖ LI V OR ‖ HUMAN | DISEASE
‖ LI V ‖

S#	C	Meta	C	C	DIET	V	VQ	LI	V	OR	HUMAN	DISEASE
1.1.1		Studies carried out			puf	affect		lipid	metab-olism		in 11 patients	with HTG
1.1.2a					feeding of puf	caused	increased	endogenous neutral steroids	excretion (into)	(feces)	in most patients (of 8 pts)	(")
b				,	"	"	"	e. acidic steroids	"			
c				or	"	"	"	both e. neutral and acidic steroids	"			
d			as com-pared with		(feeding of) sf	(not caused)	(increased)	(")	(")	(")	(")	(")
e	during	studies performed						cholesterol	balance		in 8 pts	(")
					on puf		increases marked	steroid	excretions (into)	(")	in some pts (of 8)	(")
1.1.3a	and				(")		(")	(")	(")	(")	(")	(")
			generally exceeded		(")		the decre-ment	cholesterol	in	the plasma com-partment	(")	(")

1.4a		the finding of			on puf			steroids	excretion	fecal	in pts	HTGic
b			greater (than)		(not on puf)			(")	(")	(")	(")	(")
c	contrasts to	previously reported			(on puf)?		lack of change in	sterol	balance		for pts	with FHC
d	; however	other workers have found that			puf		enhance	steroid	excretion	(")	in subjects	normal

Abbreviations:

S#	=	Sentence number	HTG	=	Hypertriglyceridemia
C	=	English conjunction	FHC	=	Familial hypercholesterolemia
Meta	=	Scientist-verb (metalanguage)	LI	=	LIPID class
puf	=	Polyunsaturated fats	OR	=	ORGAN/CELL class
sf	=	Saturated fats	V	=	verb form
VQ	=	Quantity			

FIG. 1.8 Sublanguage sentence types in sample text.

The kernel-type (mainly a LIPID-ORGAN relation but in some cases a LIPID word with a predicate) is enclosed between double bars; and the outer columns contain the classes that extend the kernel, occurring either with an operator (V) or as an adjunct. VQ stands for a quantity operator such as *increase, decrease.* The logical structure of the text is carried by the English language connectives shown in the conjunction (C) columns.

INFORMATIONAL PROPERTIES OF SUBLANGUAGES

Different sublanguages clearly differ in their gross syntactic structure and in the structure of their characteristic sentence types, stated in terms of sublanguage word classes and operators. In addition to these qualitative differences, it is possible to compare sublanguages quantitatively in regard to properties they all share to some degree, such as the amount and complexity of the information in sublanguage sentences.

The operator-argument representation provides a basis for such measurements. Because every operator represents a predication on its argument(s), the number of operators in a sentence provides a rough measure of the amount of information in the sentence, not including the operators that would appear if resolved references were substituted for the referential expression. The new, or manifest, information given by the sentence is what is measured by counting the number of operators in the sentence.

Informational complexity of sentences is also related to the number of operators contained in the sentence, but the structure among operators is also important. For example, a sentence composed of a linear sequence of simple (K-level) assertions connected by *ands* is intuitively less complex than a structure that has the same number of operators but several of them logical operators on K-structures (compare *A1 and A2 and A3 and A4 and A5* (Ai = simple assertions) with *If both A1 and A2 then A3*). To account for this feature, we include the maximum depth-of-nesting as a factor in the measure of complexity.

Two very different sublanguages (clinical reporting in patient records and lipoprotein kinetics in the experimental literature) were compared (Gordon & Sager, 1985) using in part the variables defined in Table 1.1, where the results of the comparison are given in terms of ratios of the average values of the variables in each sublanguage corpus. Literature sentences were about three times as long as record sentences (a ratio of 3.21), with somewhat less than three times as many operators (a ratio of 2.37), making for an almost equal density of information (a ratio of 1.24). However, the reductions that made for compactness were not the same in both cases. With regard to the measure

TABLE 1.1
Informational Properties of 2 Sublanguages Compared

LITERATURE CORPUS = 62 SENTENCES
RECORDS CORPUS = 113 SENTENCES

			LITERATURE : RECORDS RATIO OF AVERAGE VALUES
W	=	NUMBER OF WORDS IN SENTENCE	3.27
O	=	NUMBER OF OPERATORS IN SENTENCE, MEASURES THE AMOUNT OF MANIFEST INFORMATION	2.37
W/O	=	RATIO OF WORDS TO OPERATORS, MEASURES THE INVERSE OF INFORMATIONAL DENSITY	1.24
D	=	MAXIMUM DEPTH OF NESTING, OVERALL	2.10
D_{LOC}	=	LOCAL MODIFIER TREES, MAXIMUM DEPTH OF NESTING	13.0
O * D	=	PRODUCT OF O AND D, MEASURES THE SENTENCE INFORMATIONAL COMPLEXITY	4.17

of sentence complexity, the most striking difference is in the complexity of the local modifiers, with the literature showing an average maximum depth of nesting in adjuncts thirteen times as great as the records. The overall complexity as measured by the product of the number of operators and the maximum depth of nesting was over four times greater in the literature than in the records, that is, one third more than the ratio of sentence lengths. As sentences get longer the increase in complexity of information is not a linear function of length.

SUBLANGUAGE ANALYSIS AS A COMPUTATIONAL TOOL

Sublanguage analysis has great utility in natural language processing and its applications. Sublanguage co-occurrence patterns help to resolve syntactic ambiguity. Sublanguage sentence types lead to the formulation of target structures for semantic representation. A sublanguage grammar may eliminate ambiguity in some structures of the parent language. Sublanguage structures may also suggest larger patterns that include representations of subfield knowledge not explicit in the sublanguage text.

Sublanguage analysis provides a bridge between sentence analysis and discourse analysis. It provides forms in terms of which sublanguage discourses are seen to have a repeating structure. It may be that further work on sublanguage texts will bring into the sublanguage grammar some of the large informational patterns that now seem special to each discourse. For example, in patient documents it may be possible to identify certain event sequences, occurring under certain types of connectives or time-order relations, that correspond to more complex events, such as the patient response to treatment. As a linguistically identified, patterned occurrence such a higher level structure could be added to the sublanguage grammar.

At the same time, it is clear that extensions of sublanguage grammar toward discourse patterns and the study of discourse processes go hand in hand. In order to recognize the regular occurrence of sublanguage sentence types in a discourse, we often have to perform discourse-level operations, such as reference resolution and special transformations whose justification, in part, is the discourse regularity obtained. Conversely, without a repertoire of sublanguage sentence types to help in reference resolution and to provide the skeleton of discourse regularity, it is difficult indeed to extract such patterns from a text.

To bring together the many aspects of language analysis into a single working system is a challenging task. Sublanguage analysis is a welcome, new, powerful tool to aid in this endeavor.

ACKNOWLEDGMENTS

This research was supported in part by National Library of Medicine grant number 1-RO1-LM03933, awarded by the National Institutes of Health, Department of Health and Human Services, and in part by National Science Foundation grant number IST81-15669 from the Division of Information Science and Technology.

REFERENCES

Bross, I. D. J., Shapiro, P. A., & Anderson, B. B. (1972). How information is carried in scientific sub-languages. *Science, 176,* 1303–1307.

Brown, M. S., Kovanen, P. T., & Goldstein, J. L. (1981). Regulation of plasma cholesterol by lipoprotein receptors. *Science, 12,* 628–635.

Gordon, D., & Sager, N. (1985, August). A method of measuring information in language, applied to medical texts. *Information Processing & Management, 21* (4), 269–289.

Grundy, S. M. (1975). Effects of polyunsaturated fats on lipid metabolism in patients with hypertriglyceridemia. *The Journal of Clinical Investigation, 55,* 269–282.

Harris, Z. S. (1968). *Mathematical structures in language.* (Sec. 5.9). New York: Wiley (Interscience).

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

Hirschman, L., Grishman, R., & Sager, N. (1975). Grammatically-based automatic word class formation. *Information Processing and Management, 11,* 39–57.

Hirschman, L., & Story, G. (1981). Representing implicit and explicit time relations in narrative: In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI 81), 1* (pp. 289–295).

Kittredge, R., & Lehrberger, J. (Eds.). (1982). *Sublanguage: Studies of language in restricted semantic domains.* Berlin: de Gruyter.

Sager, N. (1972). Syntactic formatting of science information. *AFIPS Conference Proceedings, 41* (pp. 791–800). Montvale, NJ: AFIPS Press. (Reprinted in R. Kittredge & J. Lehrberger (Eds.) (1982), *Sublanguage: Studies of language in restricted semantic domains* (pp. 9–26). Berlin: de Gruyter.)

Sager, N., & Kosaka, M. (1983). A database of literature organized by relations: In R. Dayhoff (Ed.), *Proceedings of the Seventh Annual Symposium on Computer Appliations in Medical Care (SCAMC 7)* (pp. 692–695). Silver Spring, MD: IEEE Computer Society.