Foundations of Communication
Library Edition

Editor
Roland Posner

Sublanguage

Studies of Language in Restricted Semantic Domains

edited by
Richard Kittredge and John Lehrberger

Walter de Gruyter · Berlin · New York
1982

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Sublanguage : studies of language in restricted semant. domains / ed. by Richard Kittredge and John Lehrberger. – Berlin ; New York : de Gruyter, 1982.
(Library edition / Foundation of Communication)
ISBN 3-11-008244-6
NE: Kittredge, Richard [Hrsg.]

Satz und Druck: Arthur Collignon GmbH, Berlin 30
Buchbinder: Lüderitz & Bauer, Berlin

Contents

Introduction

It is well known that the topic of a discourse may affect not only the choice of vocabulary, but the "style" of expression as well. The correlation of grammatical features with discourse in certain fields gives rise to what are sometimes called *subject matter varieties*. The term *register* has also been used to describe the type of language characteristic of discourse restricted to particular subject matter, although subject matter is usually considered to be just one of several factors determining register[1]. Recently the vague notion of a correlation between what is being spoken or written about and how it is expressed in the language has come under close scrutiny by researchers in linguistics and related disciplines. Investigations of scientific articles, technical manuals, legal documents, and even cook books reveal systematic usage and, of course, specialized vocabulary that suggest *sublanguages* within a natural language.

The idea of a sublanguage as part of a natural language, with a grammar of its own, was developed in a systematic way by Zellig Harris in his work on transformations and discourse analysis. Harris gave a precise characterization of the concept in terms of his theory of language structure and carried out research on the analysis of scientific writing. In *Mathematical Structures of Language* (p. 152) he states: "certain proper subsets of the sentences of a language may be closed under some or all of the operations defined in the language, and thus constitute a sublanguage of it." There is no mention here of subject matter restrictions, only closure under certain operations. These operations correspond to transformations within Harris's theory; hence a sublanguage is closed under transformations introduced independently for the grammar of the whole language. This notion of sublanguage is like that of *subsystem* in mathematics. Harris argues that although the set of sentences in a sublanguage is a subset of the set of sentences in the whole language, the grammar of the sublanguage is not necessarily included in the grammar of the whole language; rather the two grammars intersect (see *Mathematical Structures* and chapter 11 of this book). In chapter 11 Harris discusses the relation between sublanguage and discourse and their points of departure out of the grammar of sentences. He

---

1 For a discussion of the use of the term *register* see the article by Zwicky and Zwicky in this book.

also introduces the interesting notion of a language as an envelope of its sublanguages.

Actual instances of sublanguages that have been recognized and studied are the result of discourse in particular subject matter fields. The term *sublanguage* has come to be used not just for any marked subset of sentences which satisfies the closure property, but for those sets of sentences whose lexical and grammatical restrictions reflect the restricted sets of objects and relations found in a given domain of discourse. The central role of restricted semantic domains in the identification of sublanguages is seen throughout the present book.

One of the first major practical applications of sublanguage analysis was made by Naomi Sager at New York University within the framework of a project in information retrieval. During the late 1960's Sager and her colleagues at the Linguistic String Project implemented Harris's string analysis for the automatic parsing of scientific texts. By carrying out a distributional analysis on a large corpus of articles concerning the pharmacology of cardiac glycocides, the NYU researchers were able to set up the word classes and subclasses which play a special role in the most important information-bearing sentences of that sublanguage. For example, a noun set containing *ion*, *calcium*, Na^+, *sodium*, etc. could be identified and separated in the sublanguage syntax and semantics from another noun set containing *heart*, *tissue*, *membrane*, etc. The "acceptable" sentences of the cardiac glycocide sublanguage could be precisely delimited by means of these classes, whereas no such precision could be brought to bear on the notion of "acceptable sentence" for English as a whole. This meant that when the string parser of general English gave multiple analyses for sublanguage sentences, it was usually possible to filter out unwanted ones automatically by the addition of sublanguage-dependent restrictions on lexical selection stated in terms of the distributional classes. Sager's work, described in chapter 1 of this book, gave important support for the treatment of these restrictions in a specialized *sublanguage grammar*.

More recently the NYU investigators have explored the possibility of using a precise sublanguage description to "drive" a program which automatically converts sentences of the text into a structured information format. The characteristic syntactic patterning of a sublanguage makes it possible to determine a mapping from each important sentence type, defined in terms of the distributional classes, to an underlying relational representation. This format can then be used for various information-processing operations. Several types of hospital records have been automatically converted into a data base using this format. In chapter 2 Hirschman and Sager give a detailed account of the procedure. Construction of a grammar and dictionary is made possible by applying transformational decomposition and distributional analysis to a representative corpus of sublanguage sentences. During automatic processing each

sentence is given a gross syntactic parse using a general string parser, then the lexical restrictions of the sublanguage are applied to filter out uninterpretable syntactic parses; finally, the parsed sentences, represented in terms of the sublanguage word classes, are mapped into the structured information format. The authors provide examples of parsed text and its formatted representation and give a full discussion of the difficulties in each stage, some of which remain to be solved if wide-scale implementation of practical formatting is to be achieved.

Sublanguage grammars have also been applied in the field of automatic translation. Researchers on the University of Montreal's project TAUM (Traduction Automatique Université de Montréal) have written sublanguage grammars for analyzing English texts and for generating the corresponding French texts. After some initial experiments with unrestricted language, TAUM researchers adopted a more modest goal: namely, the translation of specialized texts. In 1976 they produced a system called METEO for translating weather reports from English to French, which is now in daily use. The parsing grammar is not simply a subset of the rules of grammar for standard English, but contains rules that are not found in standard grammars (see chapter 3 of this book). The experience at TAUM demonstrated the feasibility of machine translation of texts limited to a well defined field – contingent on a thorough study of an extensive corpus to find out *how the language is actually used* in that field. Following the success of METEO, TAUM centered its attention on aviation maintenance manuals. The vocabulary is considerably larger than that needed for METEO and the grammar is much more complex (though not as widely deviant from a grammar of standard English as in the case of METEO).

The sublanguage approach to automatic analysis of texts appears to be yielding results. The commercial success of machine translation in the foreseeable future likely depends on the possibility of writing sublanguage grammars for texts in particular fields. Of course, it is to be expected that in many cases texts from several fields may share a common syntax in spite of major lexical and semantic differences.

In chapter 3 Lehrberger describes the properties of a particular sublanguage on the basis of an extensive corpus, showing how certain syntactic properties as well as lexicosemantic restrictions result from restrictions on subject matter and the purpose of the text. He presents arguments in support of the view that automatic translation is practicable for certain sublanguages, even if not for the language as a whole. Writing a formal grammar for a sublanguage, no mean task in itself, is not of the same order of difficulty as writing one for the whole language or for the "standard language". But Lehrberger argues that a description of the whole language would have to include a description of its sublanguages (which overlap to various degrees) and relations between them, and that the standard language may be useful in establishing such relations since sentences of a sublanguage

can be paraphrased in the standard language. Finally, he suggests further areas for study such as phonological characteristics identified with certain sublanguages, growth of sublanguages with scientific and cultural changes, and the effects of sublanguages on usage in other parts of the language.

The initial successes in applying sublanguage analysis to problems of information retrieval and automated translation have raised some general questions about the extent to which different sublanguages may have different sentence and text structures, and the ways in which different languages exhibit these sublanguage peculiarities. How are the principles of sentence and text structuring which each language has at its disposal exploited differentially according to sublanguage? How are the special properties of sublanguages relatable to the language as a whole? In chapter 4 Kittredge explores some of the more salient aspects of structural variation in several sublanguages of English and French. English sublanguages vary widely with respect to their inventories of syntactic structures and the frequency of these structures. On the level of text structure there is also wide variation in the type and frequency of use of these structures. There is generally greater comparability between parallel sublanguages of English and French, particularly in technical areas where texts have a well-defined purpose. Although it is difficult to assess the possibility of stylistic borrowing between technical subcultures, Kittredge feels that the structural similarities are due more to common purpose than to common sublanguage semantics. He also considers the questions of homogeneity and boundaries of sublanguages, showing that certain sublanguages have tightly structured "cores" which are embedded in a looser matrix whose lexical restrictions are closer to those found in the general language. This distinction between core and matrix levels may be an extension or generalization of the distinction between science and meta-science components noticed in early work on sublanguage. This perspective on sublanguage variation and homogeneity has implications for the design of new techniques of automatic analysis and synthesis of texts which will be well-formed in a given language, and gives a firmer foundation for optimism in formalizing and mechanizing translation in restricted technical subfields.

During the past decade, a number of researchers in artificial intelligence (AI) have investigated the problems of natural language understanding, usually within restricted semantic domains. Since AI workers are primarily interested in knowledge structures and reasoning, the tendency within this paradigm has been to illustrate new approaches to knowledge representation and manipulation by using constructed examples which contain only the kinds of problems that the novel system or approach is designed to handle. More recently, however, AI groups such as the Speech Understanding Project at Stanford Research Institute have carried out empirical studies of actual language use in controlled domain-dependent situations. Barbara Grosz' article in chapter 5 is devoted to the analysis of

task-oriented dialogues, based on a corpus of conversations between mechanics who must exchange their complementary knowledge by linguistic means in order to achieve a common goal – the assembly of an air compressor. The major concerns here include the structuring of a dialogue into sub-dialogues, the relation of ellipsis to focused information, and the relation between the speaker's level of expertise and his choice of linguistic forms. The set of dialogues in Grosz' corpus does in fact represent a certain sublanguage. But she is more interested in the process by which complex knowledge is *transferred* through specialized language usage than in how a set of utterances reflects the propositional content of that knowledge. It is more a study of how the pragmatic parameters of the communication setting influence global discourse structure than of the microstructural aspects of lexical selection, semantic subclasses of words, and domain-dependent syntactic patterning. This view of sublanguage structure is therefore complementary to that of the other authors in this volume.

Certain written sublanguages where precision is important have sometimes been subject to attempts to standardize or otherwise "improve" the conventions of linguistic expression. This is particularly the case where texts from a specialized domain acquire the status of public documents. Norms have been created for standardizing the composition of technical manuals and weather reports, although these often concern such minor features as abbreviations, punctuation and layout conventions.

In the case of legal documents, such as insurance policies, civil codes and other written instruments of the law to which the broad public has access, there has been a recent movement towards the use of "plain language". The paper by Charrow, Crandall & Charrow (chapter 6) indicates some of the dimensions of legal language which have impeded comprehension on the part of nonspecialized users. A number of factors, including social ones, have made it difficult to modernize or otherwise simplify some of the most marked features of legal sublanguages. The sociolinguistic questions involved in standardizing professional sublanguages are bound to become more and more important as the movement towards "democratizing" specialized areas of knowledge is faced with the growing subspecialization and professionalization of the increasingly complex subject matter.

The notion of sublanguage was developed independently in the Soviet Union beginning in the late 1960's. That development is discussed in chapter 7 by W. Moskovich who played an active role in the study of the structure of patent descriptions. Moskovich suggests some general properties that serve to differentiate sublanguages from natural languages in their entirety and describes the analysis of the sublanguages of organic chemistry, patent formulas and weather reports. He lists many other sublanguage studies carried out in the Soviet Union and emphasizes the need for a taxonomy of sublanguages, including the investigation of hierarchical relations among them.

In chapter 8 H. Hiż discusses the connections between different sublanguages of the same language. He also stresses the distinction between expressions that belong to a sublanguage proper and those which can be considered metalinguistic in a particular context; for example, in a book on arithmetic some of the expressions are *in* arithmetic while others are *about* it. From this point of view, just because a discourse is restricted to a particular field F, the entire discourse cannot be said to be within the sublanguage of F. It would follow that a study of the overall structure of texts in a given field would not alone reveal the structure of the sublanguage of that field. Of course, one might also claim that, without isolating metalinguistic expressions, if texts in a given field have characteristics which distinguish them from other texts, then these are the characteristics of the sublanguage of that field. The question is to some extent a terminological one, depending on how the notion of sublanguage is defined.

Closely related to the notion of sublanguage is that of register; the article by Zwicky and Zwicky (chapter 9) has been included here to help clarify the use of the latter term. The Zwickys consider register as one of four dimensions of linguistic variation: dialect, style, register and linguistic routine ("routines" include verse forms, secret languages, riddles, etc.). Registers are described as varieties of language "associated with specific contexts and specific functions of language in those contexts". The authors point out that although the style/register distinction is a very fine one and the term *register* has been used too loosely by some writers, there is, nevertheless, a useful concept of register, distinct from style; in fact, a given register may employ different styles. A register involves "an association between a set of linguistic features, the contexts in which these forms appear, and the uses to which the forms are put in these contexts". Some clear examples given in the article are baby talk, newspaper headlines and recipes. Recipes, in addition to forming a register, also constitute a sublanguage; neither baby talk nor newspaper headlines can be considered as sublanguages however, since there are no corresponding well-defined semantic domains. Sublanguage constitutes another dimension of linguistic variation in addition to register, style, dialect and routine.

The problem of distinguishing various sublanguages from each other and from the language as a whole can be studied on the purely semantic level of propositional content as well as on the level of linguistic structure. One important approach to this problem has grown out of the attempt by philosophers of language to characterize both the coherence and the content of texts in terms of sets of propositions associated with each sentence.

The paper by Bellert and Weingartner (chapter 10) presents a general theoretical framework for distinguishing different types of texts, based on the kinds of additional sentences (propositions) necessary for their interpretation. Certain fundamental varieties of scientific text can be distinguished according to whether the auxiliary sentences required for their

interpretation consist only of logical sentences, or also include hypotheses, laws or theories. A text which is a statement of a scientific theory or an introduction to a field of knowledge also has a distinguishable set of auxiliary sentences. "Everyday language" texts have still different characteristic auxiliary sets and can be subdivided depending on their use of such items as indexical signs.

The final article by Zellig Harris, already commented on at the beginning of this introduction, gives some perspective on the variety of papers in the collection. His pioneering work on sublanguage is increasingly relevant in the light of recent developments in automatic processing of texts in various fields and studies of specialized use of language. The succinct characterization of sublanguage found in Harris' theoretical work of the 1960's has stood up remarkably well in the course of detailed work on actual cases.

The term *sublanguage*, as used in this collection, is relatively new and the systematic study of sublanguages is still in its infancy. We believe that the notion of sublanguage deserves consideration along with the more familiar notions of dialect, style, register and standard language. We recognize the need to sharpen the definition and hope that the studies presented here will lead to a critical analysis of the concept as well as further development of the consequences for linguistics and various language-related fields.

Montreal, May 1980 R. K. & J. L.

Chapter 1

Syntactic formatting of science information

Naomi Sager

Introduction

It has been increasingly recognized that science information systems have need of natural language processing. F. W. Lancaster, author of the National Library of Medicine Study of the performance of the MEDLARS system, [1] spoke of this at the 1971 annual conference of the ACM, in the panel "Can Present Methods for Library and Information Retrieval Service Survive?"[2] He noted that "there is a definite trend away from large carefully controlled vocabularies and toward natural language processing, or at least machine-aided indexing," and quoted Klingbiel's remarks to the effect that "highly structured controlled vocabularies are obsolete for indexing and retrieval" and that "the natural language of scientific prose is fully adequate for these purposes."

In the direction of more flexible, user-oriented systems, the question has also been raised as to whether computer methods can be developed for accessing the information in scientific articles directly, without the mediation of a librarian or systems expert between the user and the stored information. Professor J. Belzer, chairman of the above panel, raised this question: "Our so-called information retrieval systems are in fact not information retrieval systems. They are bibliography producing systems, and we store documents and not information. . . ." "Were the system able to supply him (the user) with the information he wanted, it would not be necessary for him to read the entire document." In light of these remarks, we ask: Is it indeed possible for a mechanical system to identify the portions of a text which contain specific information? Can the information in sentences of the natural language text be organized on the basis of computer processing of the text so that each sentence becomes a case of a regular pattern which is both linguistic and informational, i.e., a format?

That the answer to this question is "yes," is suggested by the results of a recent research into the specialized use of language in scientific subfields. The discourse in a science subfield has a more restricted grammar and far less ambiguity than has the language as a whole. We have found that the research papers in a given science subfield display such regularities of occurrence over and above those of the language as a whole that it is possible to write a grammar of the language used in the subfield, and that this specialized grammar closely reflects the informational structure of discourse in the subfield. We use the term sublanguage for that part of the whole language which can be described by such a specialized grammar.

The sublanguage grammar provides a method for developing the particular word classes (the special-word sets) and the relations among these classes which are of special significance in a given science subfield, i.e., which are the linguistic carriers of the specific knowledge in the subfield. Yet these categories and relations are not determined *a priori* for the subfield. Rather, they are the interpretation of the formal grammatical categories and relations of the sublanguage grammar. Thus, in the pharmacological sublanguage which was investigated, the two noun subclasses I (containing, e.g., *ion*, K^+) and G (containing, e.g., *drug*, *digitalis*, *glycosides*), which in the subfield have the significance "ions" and "pharmacological agents," respectively, and play crucially different roles in the physiological mechanisms being described, are obtained as separate classes because they occur with different classes of verbs: e.g., I as the object of such verbs as *transport*, G as the subject of such verbs as *inhibit*. It then turns out that the sublanguage word classes, which are established on the grounds of what other grammatical classes they occur with (as subject, object, etc.), are the linguistic counterparts of the real-word objects, events, and relations which are studied and described in the given subfield.

A sublanguage grammar leads to a grammatical format for sentences in the sublanguage in which the words in each "slot" of the format are found to correspond to a particular kind of information in the subfield. For the pharmacological subfield whose grammar is summarized below, there are grammatical slots corresponding to: biochemical or physiological events, quantitative relations, drug actions, connections between science facts, and experimental and epistemic relations of the scientist to the objects and facts of the science. As with the sublanguage grammar itself, the words of a sentence are not assigned to the slots of the sentence format on the basis of their semantic properties, but on the basis of their subclass standing vis-a-vis other grammatical word classes in the sentence. A description of the formats for the pharmacology sublanguage and examples of formatted sentences are given following the summary of the sublanguage grammar, below.

Sublanguage Grammar

The following is a sketch of the sublanguage grammar for the pharmacological subfield dealing specifically with the cellular level actions of the cardiac glycosides (digitalis).

Location of the science vocabulary in the sentence structure

For purposes of this work, the structure of a sentence can be represented by a string decomposition obtained mechanically by a computer program, [3], [4], [5], or by a transformational decomposition, [6] or a transformational lattice [7]. In the latter two types of analysis, each sentence of the sub-

language is decomposed into one or more elementary sentences S_e with a succession of (partially ordered) operators which operate on the S_e or on the S_e with operators on them. For example, in the sentence *It is clear that toxic doses of digitalis are regularly associated with a loss of myocardial K*, a simple version of this analysis is shown by grouping the words of the sentence into levels corresponding to S_e and the successive operators:

$$\text{It is clear that} \left(\begin{array}{l} \text{toxic doses of} \\ \text{digitalis is regularly} \\ \text{associated with} \end{array} \left(\begin{array}{l} \text{a loss of} \\ \text{myocardial K} \end{array} \right) \right).$$

When sentences from articles in the science subfield are decomposed by any one of the above methods, it is found that the vocabulary which is characteristic of the subfield (called here the science-specific vocabulary) occurs in a distinguished portion of the decomposition, i.e., in nodes corresponding to S_e and the immediate operators on S_e (the "bottom" nodes of the lattice or string decomposition), while the more general science vocabulary is at the intermediate nodes of the lattice or string decomposition. The top nodes are occupied by epistemic vocabulary presenting the scientist's relation to the science facts [4].

Form of S_e

When we consider the science-specific verbs in the bottom-most nodes of the sentence decomposition, i.e., the verbs in S_e, we find that the subject of these verbs is a science-specific noun, and the object (if the verb is transitive) is also a science-specific noun, or several, interspersed with prepositions (e.g., *the cell loses potassium, ions flow into the cell*). Letting N and V stand respectively for the science-specific nouns and verbs in S_e, and P for a preposition selected by the given verb, a formula for the elementary sentence is:

$$S_e = N_1VP_1N_2P_2N_3$$

where a given verb may have only a portion of the $P_1N_2P_2N_3$ sequence as its object, or in some cases a longer sequence.

In the sublanguage many of the science-specific verbs have only one or two object possibilities, fewer than in their use in English as a whole. In some cases a prepositional phrase would be an object of a verb in the sublanguage whereas in English as a whole it would be considered an adjunct, e.g., *exchange* (*across membrane*). This fact reduces the ambiguity in the sentence analysis, and simplifies the work of obtaining a sentence analysis by computer.

N sets in S_e

A compact description of the main types of elementary sentences is obtained if we collect the science-specific nouns into (almost entirely) disjoint sets, chief of which are:

G (pharmacological agent)	e.g., glycosides, digitalis, digoxin, ouabain, erythrophleum alkaloids
I (ion)	e.g., K^+, Na^+, Ca^{++}, potassium, sodium, calcium
T (tissue)	e.g., muscle, strips of ventricle, vesicles, epithelium, fibers
C (cell)	e.g., cell, red cell
M (membrane)	e.g., membrane
H (heart)	e.g., heart, atrium, myocardium
O (other organs)	e.g., kidney
F (fluid)	e.g., fluid, medium solution, suspension

Certain nouns in these sets are pure synonyms in the sublanguage, completely interchangeable under whatever verb they occur with: *sodium, sodium ions, Na* (the first two are of course not synonyms in other areas of science writing).

Certain words are classifiers of particular sets (e.g., *ion* for K, Na, Ca, Cl), with such word-sequences as *these ions* being synonyms for particular ones of these in a particular textual occurrence. There are also verbs which are used as classifiers of certain sets of verbs (e.g., *act*).

Certain nouns occur as fragment-names of other nouns. A noun N_1 occurs as a fragment-name of N_2 if there exists a possible sublanguage sentence "N_1 is a part of N_2," and if in the given occurrence, N_1 occurs as the subject/object of a verb which elsewhere has N_2 as its subject/object. For example, in *The glycoside inhibits the Michaelis component of influx, the Michaelis component* is a fragment-name of *influx*.

In considering the combinations of nouns and verbs occurring in the texts, we note that while each of the above noun sets appears uniquely as the subject or object of certain verbs, there are also verbs which take their subject or object from particular unions (marked /) of these sets. There are also verbs which take their subject or object only from particular subsets of these sets (e.g., only sodium and potassium in I, or only Ca).

V-sets of S_e, and main S_e subtypes

Verb subclasses can be set up on the basis of verb occurrences in particular environments composed of the above noun sets. The environments are cases of the S_e formula. Some of the main environments for classing S_e verbs

are listed below, followed by a sample list of verbs in each class. The statement of the subject and object noun classes with which a given verb on the list occurs is limited to the occurrences of that verb in the sentences of the articles which were analyzed. The verb classes are largely disjoint, but a given verb may be in more than one class. Verbs whose active form would have a human subject and a science-specific noun as object are stated in the passive form.

T/H__:	contract, relax; is isolated.
T/C__:	is washed, is cooled, is cold-stored, is warmed, is incubated, is fresh.
T__:	is fractionated, is prepared, shortens.
C__:	rest, swell, recover.
H__:	beats, fails, is quiescent, is stimulated, survives, responds inotropically, functions, works, (has) activity.
M__:	is permeable, is leaky. (These could be obtained from I__M, below).
I__I:	replace, exchange with (across membrane).
I__C/T:	move (in)to, enter, flow in/out, occupy (site in), is stored in, is sequestered in, concentrate in, accumulate in/at, distribute in, constitute composition of.
I/G__C/T:	diffuse into, are in, leave, localizes in, is removed from.
I__M:	permeate.
G__M:	penetrate.
C__I:	regain, expel, is loaded with.
C/T__I:	extrude, eliminate, is depleted of, leaks, are deprived of, gain.
H/C/T__I/G:	lose, take up.
O/T__I:	excretes, turn over, release.
G__T:	is absorbed into, is located in (region), reaches, combines with, is injected into; poisons, inactivates.
T__G:	gets rid of, responds to, resists, is exposed to, is treated with.
F__F:	equilibrate.
T__F:	is suspended in, is surrounded by, is bathed in.

X__X (for any set X): is (ouabain is a glycoside).

Grouping the main S_e subtypes

If we consider the above list we note that there are only a few types of subject/object pairs for these verbs. To obtain a more compact representation, we define an inclusive tissue class $\bar{T}$ = T/C/H/M/O, and an inclusive class $\bar{I}$ = I/G. In terms of these super classes the main environments above can be summarized as follows, defining the verb classes V_T, V_{II}, V_{IT}:

$\bar{T}V_{T}$
$\bar{I}V_{II}\bar{I}$
$\bar{I}V_{IT}\bar{T}$

The additional type $\bar{T}V_{TI}\bar{I}$ can be included in the above types by taking the verbs in the passive.

While the grouping of S_e subtypes into supertypes is a convenient reduction of a large amount of data, the individual subtypes within one supertype may behave differently under further operators. This is the case with $\bar{I}V_{IT}\bar{T}$ where $IV_{IT}C$ (*ions leave cell*) occurs under such operator sequences as *digitalis inhibits* (see below) whereas $GV_{IT}C$ does not.

It is found that the verb classes defined in this way are very nearly disjoint. The noun super-classes above are disjoint collections of the virtually disjoint noun subclasses established above.

Furthermore, if we consider the verbs in the list, we find that with the exception of those noted below, most of the verbs refer to movement or the result of movement: moving in or through (*flow into, transport*), staying in place (*occupy, sequester*), being in a place by virtue of having moved (*concentrate, accumulate, distribute*), favor moving or staying (*select, resist*). Many of these verbs are indeed synonymous in respect to these elementary sentences, and the others could all be replaced in these elementary sentences by synonymous word sequences, a base verb *move* with particular prepositions and quantifiers (e.g., *permeate*: move through; *gain*: move in to a greater degree than move out, etc.). The verbs which do not relate to movement are mainly the intransitive and laboratory verbs at the beginning of the list, and certain particular verbs, such as *poison, inactivate, destroy* and *respond to* and *equilibrate* in the latter part of the list. This main set of elementary sentences of the subfield is thus composed of a single verb *move* with directional and quantitative modifiers which connect $\bar{I}$ to $\bar{T}$, (and $\bar{I}$ to $\bar{I}$ in respect to $\bar{T}$, e.g. *exchange*).

Other S_c subtypes

In addition to the main S_c subtype (covering ion transport phenomena) which is described in some detail above, there are several further S_e subtypes which are important in the subfield:

- S_e whose main nouns are *contractile proteins, actin, myosin* and characteristic verbs are *slide along, fold along* (*the sliding and perhaps folding of actin molecules*).
- S_e whose main nouns are *ATPase, ATP*. One such S_e has *ATPase* as subject and *ATP* as object, with *hydrolyze* as a characteristic verb. Most frequently the S_e verb occurring with *ATPase* is *act,* which is a classifier verb for more specific S_e verbs.

- S_e whose characteristic verbs are *carry*, *transport* (across membrane) with I (e.g., *sodium*) as characteristic object, and with *mechanism, substance, pump,* as frequent subjects, when the subject is given explicitly.

In addition to the above, in some articles or parts of articles, there are elementary sentences whose vocabulary is drawn (in part) from noun classes not mentioned above. Examples of these are: *The curve flattens toward the x-axis, cardiac glycosides possess unsaturated rings, the potential is negative.* These elementary sentences are found to be sentences of other, related, sciences and techniques on which our particular subscience draws.

Local modifiers of N and V; and wh-connectives

Certain additional words operate on the words of the S_e sentences. The operators on the nouns may appear as adjectives, prepositional phrases and other modifiers. The operators on the verb may be adverbs, prepositional phrases and other modifiers. The noun modifiers can be reconstructed into separate sentences connected by a relative pronoun (*that, which,* etc., indicated by *wh*) to the given sentence, and the verb modifiers into separate sentences connected to the given sentence by a bisentential verb V_{ss}. Below, in proposing a format for the content of each sentence, we will suggest that instead of transforming all modifiers out of the sentence, as one does for language as a whole, we consider if there are any word sets in modifier position which in this sublanguage are especially dependent on their host words, or which never have an explicit conjunctional relation to it; these, we suggest, might best be left in modifier slots next to their host word in the format.

Aspectuals, V_v

Certain verbs V_v (not science-specific verbs treated above) operate on verbs as more or less aspectual modifiers. In English, they occur either in pre-verb or post-sentence position, and most can be transformed from one to the other: *He commenced speaking, His speaking commenced.* In this sublanguage, only a few are used, and all are aspectual in meaning (including the negative), and apparently all can occupy the pre-verb position: *not, fail to, appear to, tend to, be engaged in, undergo, persist, continue, remain, become, commence, start.* E.g., *the force starts to increase, the steroids undergo interconversion, depolarization persists* (persist in depolarizing). Several of these are synonyms of each other in the sublanguage.

Quantifiers, Q

Certain verbs (e.g., *flow, transport, lose, gain, accumulate*) can have a modifying quantifier Q: *in an amount, at a rate*; or when the verb is nominalized: *amount of, rate of*. This holds for certain adjectives and nominalized adjectives (e.g., *toxicity, activity*) and even nouns (*force*). Some can even be considered to contain a quantifier (e.g., *concentration* is synonymous in this sublanguage with *amount of concentration*). Quantifiers can also be considered to be modifiers or predicates of certain nouns: *amount of digitalis, digitalis is present in a certain amount.*

There are certain other verbs (different from any listed in preceding sections) which operate on these Q. Of these, there is a subset V_q whose members have Q as their subject, and there is a subset V_{qq} whose members have Q as their subject and Q as their object. An example of V_q is *decrease* in *the size of the overshoot decreases*; an example of V_{qq} is *equals* in *the amount of alcohol in . . . , equals the amount of alcohol in . . .* , and *the chloride ratio equals the potassium ratio.* A quantifier Q occurring with V_q or V_{qq} is often omitted (zeroed), since its original presence can be reconstructed from the grammatical requirements of the V_q or V_{qq}. Thus, in addition to: *raise the internal sodium concentration,* we find also: *raise the internal sodium.*

The chief verbs here are:

V_q: *decrease, reduce, fall, increase, rise, change, run down, level off, stand still.*

V_{qq}: *equal, differ from, range from__to__, be twice, vary with, correspond to, depend on, determine, reach.* Certain V_{qq} appear also with a human subject with the two Q's in the object: *compare, correlate* (an amount with an amount); *determine, calculate* (an amount from an amount).

There is also a $V_{V_qV_q}$, i.e., a verb having V_q both as subject and as object: *parallel* (*the increase in tension parallels the increase in uptake*). That a verb should require such a hierarchy of object-types is unique in the sublanguage, and not common, if it exists at all, in the language as a whole.

The V_q and V_{qq} can operate not only on Q but also on V_q: *the rise* (*in amount*) *depends on . . .* , where *depend on* is a V_{qq} operating on *rise* and *rise* is a V_q operating on Q *amount*. There are also purely causative verbs whose objects are Q or V_q: *double, accelerate, minimize, depress.*

We see that a complex structure of quantifiers and quantifying verbs operates in this sublanguage. As in the case of the verbs reducible to *move*, above, many of the quantifying verbs here are synonyms, or are replaceable by a few base verbs with modifiers on them.

Verbs connecting two sentences, V_{ss}

There are certain verbs, not included in any of the preceding sets, which have nominalized sentences both as subject and object. These verbs are the bisentential V_{ss}. A particular property of these verbs is that if their first nominalized sentence is *presence of X* or *action of X*, where X = the noun subclass G (rarely, I), the words *presence of* or *action of* are omittable, yielding X as the apparent subject of the V_{ss}: *glycosides inhibit . . . .* These V_{ss} are: *affect, is concerned in, bring about, cause, produce, confer, make, generate, induce, initiate, trigger, promote, stimulate, prolong, protect from, restore, control, interfere with, inhibit, limit, delay, antagonize, depose, reverse, block, arrest, abolish, obstruct, prevent, switch off.*

Instead of considering *glycosides inhibit sodium efflux* as reduced from *action of glycosides inhibits sodium efflux* (an $SV_{ss}S$ construction), we can consider the G noun, when it appears as subject of V_{ss}, to constitute a special N-class N_o. Then *glycosides inhibit sodium efflux* would be a case of an $N_oV_{ss}S$ construction. We use the latter analysis in the format, below. Here, too, it is clear that there are many synonyms with respect to the use of these verbs in the sublanguage, so that the vocabulary could be reduced.

There are a few other sentence-connecting verbs which may be called conjunctional V_{ss}. Here, G does not occur as possible subject: *involve, accompany, relate to, lead to, depend on, be based on.* Similar to these are certain passive forms: *be linked to, be coupled to, be related to,* which in the active form have a human subject.

Subordinate conjunctions, C_{subord}

There are a number of subordinate conjunctions between sentences: *if S then S, S when S,* etc. There are also certain prepositions used conjunctionally between nominalized sentences: *No contracture occurs on depolarization, Recovery does not occur in the absence of oxygen.*

Coordinate conjunctions, C_{coord}

There are conjunctions between S, or between identically classed words: *and, or.*

Sentence grouping (non-associativity of connectives)

All the sentence connectors, including *wh,* can operate on each other, i.e., an $SV_{ss}S$ or an SCS can serve as subject or object of a sentence connector. When there is more than one connective, the grouping of sentences is semantically non-associative, but sequences of SCS, where $C = C_{coord}$, are associative.

Epistemic operators

Finally, there are many verbs with epistemic meaning, whose subject is human and whose object is a sentence: *believe*, *publish*. The human subject is often omitted when the sentence is nominalized in the passive.

Summary

Grammar. This sublanguage had a definite grammatical structure consisting of:

(1) a set of elementary sentences, formed out of a few sets of subscience-specific nouns and verbs; and occasional other elementary sentences of a few other subscience vocabularies.
(2) aspectual operators on verbs.
(3) (omittable) quantifiers Q on certain verbs or nouns, with quantifying verbs V_q and V_{qq} operating on the Q or on V_q; and a verb V_{VqVq} operating on two V_q's.
(4) the noun-modifying *wh*-connective.
(5) sets of sentence-connecting verbs V_{ss} and conjunctions C, which can operate on each other.

Vocabulary reduction. In each word set, various words are used synonymously or can be replaced by a common base word with differentiating modifiers. Hence the vocabulary in each word set can be greatly reduced, at least for the purposes of a standardized informational representation.

Semantic interpretation. The particular word sets (especially after their vocabulary has been reduced) and the way they operate on each other reflect quite closely the structure of information in the science. E.g., a main S_e subtype is *I/G move in T/C*; and the main appearance of glycosides is not in the elementary sentence, but as subject of the causative operator verb on the S_e. Also, the complexity of the quantity words reflects the importance of quantitative relations in this subfield.

Sublanguage Sentence Format

A sublanguage grammar provides a basis for structuring the information in each sentence and for mechanically processing the structured information.

A parse of a sentence, whether carried out by hand or by a computer program, is a decomposition of the sentence into parts which are segmented, and related one to the other, in terms of the grammar used. When the grammar includes, in addition to the grammatical requirements and transformations of the language as a whole, also the special word subsets and restricted combinations of the given science sublanguage, the sentence seg-

ments and their relations are found to fit the informational categories and relations of the subscience. It is possible to construct a fixed format of the grammatical operators and operands which houses all the sentence outputs obtained using the sublanguage grammar, so that the grammatical decomposition (parse) of each sentence locates the sentence-segments in particular slots of the format. Each of the slots has a fixed informational character, and each sentence carries the type of information of the slots which it fills, in their relation to neighboring slots in the format.

Aside from the sublanguage grammar, it is known that in language in general there are certain grammatical processes which lead to the loss of words in a sentence or to the replacement of words by informationally less explicit ones. The reverse process of supplying the lost or more specific words is especially important in formatting sentences. The main such processes are:

(1) Loss of repeated words (called "zeroing"), especially after a conjunction. E.g., *changes in the concentration of electrolytes and in electrolyte fluxes* can be filled out to include the zeroed word *changes* after *and*, to yield *changes in the concentration of electrolytes and* [*changes*] *in electrolyte fluxes*. In the formatted sentences, below, zeroed words which have been reconstructed are enclosed in [ ].

(2) Replacement of a repeated word or sentence by a pronoun, e.g., *its* in *the inotropic action of digitalis cannot be attributed to its effect on potassium metabolism*, and *This* in *This results from a slowing of the influx*. A so-called bound pronoun occurs in words like *which*, which can be analyzed as a conjunction *wh* followed by a pronoun *ich* standing either for a preceeding noun or sentence. In the formatted sentence, material which has been reconstructed in place of a pronoun is enclosed in { }.

(3) Replacement of a repeated word or sentence by a classifier of the word or sentence, usually as part of a sequence containing *the*, *this*, *these*, etc., e.g., *the drug* replacing a second occurrence of *digitalis* in the same sentence, or *these effects* replacing the repetition of a preceding sentence. The combination of a pronominal element (e.g., *these*) with a classifier word or phrase eases the task of identifying the antecedent of the pronoun. In the formatted sentences, material which has been reconstructed on the basis of classifier sequences is enclosed in ⟨ ⟩.

(4) Grammatical constants. When a sentence occurs as the subject or object of an operator verb, the sentence may be nominalized, e.g., *an influx of potassium into the cell* following the operator verb *results from*, nominalized from *potassium flows into the cell*. In reconstructing the sentence which had been nominalized it is sometimes necessary to supply an informationally neutral word in order to make the