Offprint

Sublanguage

Studies of Language in Restricted Semantic Domains

edited by
Richard Kittredge and John Lehrberger

Walter de Gruyter · Berlin · New York
1982

Foundations of Communication
Library Edition

Editor
Roland Posner

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

Contents

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:

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

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_e subtypes

In addition to the main S_e 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 $V_{V_qV_q}$ 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

nominalized form of the verb into a grammatical verb form. E.g., *intracellular sodium* in *Intracellular sodium is increased*, can be reconstructed into a sentence *sodium (is) intracellular*. In the formatted sentences, grammatical words which are supplied are enclosed in ( ).

U_1	C_1
U_2	C_2
	C_{N-1}
U_N	

Figure 1 – Sentence Format
U = Unary Sentence; C = Conjunction

The structure of a sentence in this grammar, after lost and replaced material has been reconstructed as far as possible, consists of a unary sentence U, or a sequence of U's connected by conjunctions C, where $C = V_{ss}$, V_{qq}, V_{VqVq}, C_{subord}, C_{coord}, *wh*, defined in the grammar above. This sequence, UCUC . . . CU, is the resultant of each C operating on a pair of sentences, where each of these operand sentences is itself a U or the resultant of a C operating on a pair of sentences. If there is more than one C in a sentence, parentheses are needed to indicate the hierarchical (non-associative) grouping of the operands of C's. In a formatted sentence, the U's and C's are arranged in columns, as in Figure 1. Grouping is shown by barred square brackets ⫞ ⫟. If no grouping is shown, it is understood that each C operates on the last preceding U.

N_o	V_{ss}	V_q	S_e	D_s

Figure 2 – Format for Unary Sentence
N_o: subject of V_{ss}, generally a pharmacological agent noun
V_{ss}: operator on V_q or S_e, generally causative verb
V_q: verb of quantity
S_e: elementary sentence of the sublanguage
D_s: adverb on the whole preceding structure to its left
The first two boxes and the last may be empty; the first and last boxes are repeatable

Each unary sentence U is an elementary sentence S_e, or an S_e with one or more unary operators on it. The unary operators are either N_0V_{ss} or V_q, as illustrated in Figure 2. Each S_e is one of the S_e subtypes described in the grammar above. The gross grammatical structure of S_e is illustrated in Figure 3. In all the above formats, certain particular word subsets which

appear in the structure may have particular sets of local modifiers operating on them. The grammatical form of these modifiers on nouns, is: adjectival phrases and possibly quantifiers Q; and on verbs: aspectual pre-verbs V_v, adverbial phrases D and quantifiers Q.

Tables I–III contain the formats for the first two paragraphs of the section *Effects on cellular potassium* in a review of Digitalis [8]. In the textual sentence preceding each format, the "scientist-level" portions are italicized; these can be separated grammatically from the more specifically science portions of the sentence, and have not been included in the formats. The format follows the pattern illustrated in Figures 1–3. The elementary sentence S_e appears between double lines, and where the verb V occurs in the sentence in nominalized form, that form has been retained in the format. The first preposition following a verb in S_e has been written along with the verb in the V column.

With regard to the individual sentences:

N_1 Q	V	Q	P N_2	P N_3

Figure 3 – Format for Elementary Sentence S_e
N_1, N_2, N_3: Nouns from the specifically pharmacological vocabulary
Q: Quantity word
V: Verb or word with verb root
Only N_1 and V are necessarily present in each S_e

In (1.): While *changes in cells produced by digitalis* is ambiguous in English as a whole, it is not ambiguous in the sublanguage since nouns in the class C (*cells*) do not occur as the subject of V_{ss} verbs (*produce*). In the sublanguage the word *changes* only operates on quantity words Q (e.g., *amount* , *rate*) or verbs which have an implicit Q on them. In the formats, therefore, *change* occupies the V_q position. In the format for sentence 1 this places *internal milieu of cells* in the S_e position, suggesting that it contains an implicit Q and V. This is supported by the fact that the paraphrase *changes in the amounts of X1, X2, . . . in cells* is an acceptable substitute for *internal milieu of cells* in all its textual occurrences in this sublanguage.

In (2.): The first part of sentence 2 contains lost repeated material (zeroing) which can be reconstructed because of the strong grammatical requirements on the superlative form: *Most prominent have been changes in . . .* is filled out to *Most prominent of these changes have been changes in . . . . These changes* is a classifier sequence replacing the full repetition of sentence 1, which is then shown in the format as the first (zeroed) unary sentence of 2.

In (2.3–2.5): The word which indicates that *changes* (along with *digitalis produces*) has been zeroed is the repeated *in* after *and*. In 2.2, the V in S_e is (*have*) *concentration in* (or: *concentrate to some amount in*), which in

Table I

Formatted Sentences 1–3 CONJ: conjunction; G *pharmacological agent* noun class; C: *cell* noun class; V_{IT}: verb with subject noun from $\bar{I}$ (*ion* super-class) and object noun from $\bar{T}$ (*tissue* super-class); other symbols are noted in Figure 1–3

1. Changes in the internal milieu of cells produced by digitalis *have been known for many years.*

	N_0	Q	V_{ss}	V_q D	N_1	V Q	N_2	D_s	CONJ
1.	G Digitalis		produces	changes	← internal	milieu of →	C cells		

2. Most prominent have been changes in the concentration of electrolytes and in electrolyte fluxes, that is, in the rate of movement of electrolytes in and out of the cell.

2.1	←----------	----------	----------	----------	---⟨[1]⟩---	----------	----------	(has) most prominent (case in)
2.2	G [Digitalis]	[produces]	changes in	I electrolytes	(have) concentration [in]	C [cells]		and
2.3	G [Digitalis]	[produces]	[changes] in	I electrolytes	fluxes P	C [cells]		, that is,
2.4	G [Digitalis]	[produces]	[changes] in	I electrolytes	movement in (at a) rate	C the cell		and
2.5	G [Digitalis]	[produces]	[changes in]	I [electrolytes]	[movement] out of [(at a) rate]	C [the cell]		

3. *Interest was initially focused on* changes in potassium; *more recently*, changes in calcium *have been recognized to be of great importance.*

3.1		changes in	I potassium	V_{IT}	$\bar{T}$		;
3.2		changes in	I calcium	V_{IT}	$\bar{T}$		

Table II
Formatted Sentences 4–6
Symbols as in Table 1

4. Toxic doses of digitalis consistently reduce the intracellular concentration of potassium in a wide variety of cells, including cardiac muscle cells.

	N_0	Q	V_{ss}	V_q D	N_1	V Q	N_2	D_s	CONJ
	G				I		C		
4.1	Digitalis			reduces	potassium	(has) con-	cellular	in a wide	including
		toxic		consistently		centration		variety	
		doses				intra-		of cells	
	G				I		C		
4.2	[Digitalis]			[reduces	[potassium]	(has) [con-	[cellular]	in cardiac	
		[toxic		consistently]		centration		muscle	
		doses]				intra-]		cells	

5. This results from the slowing of the influx of potassium into the cell.

5.1	←----------			----------	----------	----{4}----	----------	----------→	results from
					I		C		
5.2				slows	potassium	influx into	the cell		Concurrently

6. Concurrently, intracellular sodium and water are increased.

					I		C		
6.1				increases	sodium	(is) intra-	cellular		and [concurrently]
					I				
6.2				[increases]	water	[is intra-	cellular]		

Table III
Formatted Sentence 7
Symbols as in Table 1

7. *It is not certain whether* these linked changes in sodium and potassium are produced by a single effect or are separately mediated.

	N_0	Q	V_{ss}	V_q D	N_1	V Q	N_2	D_s	CONJ
7.1	single effect		produces ⫞	←		⟨6.1⟩	→		and
7.2				←		⟨5.2⟩	→		⫞ or
7.3	⫞N		mediates separately	←		[6.1]	→		[and]
7.4	N		[mediates separately]	←		[5.2]	→		⫞ wh
7.5				←		{⟨6.1⟩}	→		is linked to
7.6				←		{⟨5.2⟩}	→		

the sublanguage requires an object noun from the gross tissue-cell class $\bar{T}$. Similarly in 2.3, the V *fluxes* (with unspecified P) requires an object noun from $\bar{T}$. In the analyzed texts both of these Vs occurred almost exclusively with the noun *cell* as their object. The definitional connective *that is* between 2.2,3 and 2.4,5 supports substituting the word *cell* for $\bar{T}$.

In (3.): The sublanguage requirements on the noun class I (*potassium, sodium*) as the first noun in S_e, when S_e is operated on by V_q (*changes*), are that the verb be of the type V_{IT} or V_{II} and the second noun be of class $\bar{T}$ or $\bar{I}$. The continuity of this sentence with its surrounding sentences suggests that the verb is V_{IT} and the noun $\bar{T}$ (more specifically C:*cell*).

In (5.1): The pronoun *this* replaces the entire preceding sentence.

In (7.): *These linked changes in sodium and potassium* transforms into *These changes in sodium and potassium which are linked*. The portion up to *which are linked* is a classifier of the two preceding unary sentences, 6.1 and 5.2, pinpointed by the repetition of the words *sodium* and *potassium* in the classifier sequence. It is these two conjoined unary sentences which are operated on by *a single effect produces* in lines 7.1, 7.2, and again by *mediates separately* with unknown N subject, in lines 7.3, 7.4. The portion *which are linked* applies to both occurrences of 6.1 and 5.2 in 7.1–4. The *wh* in *which* is the connective and the *ich* part is a pronoun for the two sentences, as indicated by { }. The fact that the sentences were reconstructed by use of a classifier is indicated by the ⟨ ⟩ inside the { } in 7.5–6. Although this sentence seems empty, it is common in scientific writing for a sentence to consist of references to previous sentences with new operators and conjunctions operating on the pronouned sentences. The linearity of language makes it difficult to express complex interconnections between the events (sentences) except with the aid of such pronouned repetitions of the sentences.

The appearance of a word like *effect* in the column usually filled by a pharmacological agent noun G may herald the future occurrence of a new elementary sentence or a new set of conjoined elementary sentences (classified by the word *effect*) which will intervene between G and the present S_e. This appears to be one of the ways that new knowledge entering the subfield literature is reflected in the formats and the sublanguage grammar.

In fact, in the work described here, the first investigation, which covered digitalis articles up to about 1965, showed certain sets of words (including *mechanism*, *pump* and, differently, *ATPase*) appearing in the N_o or D_s column as an operator on S_e. In later articles, which were investigated later, these nouns appeared increasingly as subjects of new S_e subtypes listed above in the grammar, connected by conjunctions to the previously known S_e. The shift of these words from occurring as operators to occurring in (or as classifiers of) new S_e subtypes is the sublanguage representation of the advance of knowledge in the subfield.

Acknowledgements

This work was supported by Research Grants R01 LM 00720-01, -02, from the National Library of Medicine, National Institutes of Health, DHEW. Important parts of the sublanguage grammar are the work of James Munz, to whom many of the results and methods are due.

References

[1] F. W. LANCASTER: *Evaluation of the Medlars demand search.* National Library of Medicine, 1968.

[2] Proceedings of 1971 Annual Conference of the ACM, pp. 564–577.

[3] N. SAGER: *Syntactic analysis of natural language.* Advances in Computers, 8. F. Alt and M. Rubinoff (Eds.). Academic Press New York, 1967.

[4] N. SAGER: *The string parser for scientific literature.* Courant Computer Symposium 8–Natural Language Processing. R. Rustin (Ed.) Algorithmics Press, New York, 1973.

[5] String Program Reports Nos 1–5. Linguistic String Project, New York University, 1966–1969.

[6] D. HIŻ and A. K. JOSHI: *Transformational decomposition – A simple description of an algorithm for transformational analysis of English sentences.* 2eme Conference sur le Traitement Automatique des Langues. Grenoble, 1967.

[7] String Program Reports No 6. Linguistic String Project. New York University, 1970.

[8] A. F. LYON and A. C. DEGRAFF: *Reappraisal of digitalis, part I, Digitalis action at the cellular level.* Am Heart J. 72 4, pp. 414–418, 1961.

Editors' note: this article is reprinted from AFIPS Conference Proceedings 41, pp. 791–800, AFIPS Press, Montvale, N.J., 1972.