Sublanguage Grammers in Science Information Processing

This paper presents the results of an investigation into information structures in natural language science texts. A novel hypothesis was tested; namely, that the literature of a science subfield has characteristic restrictions on language usage which can be used to develop information formats for text sentences in the subfield. The formats provide a standard representation of the specific types of information found in sentences of subfield articles, though \textit{a priori} semantic categories are not used. The method of sublanguage grammars for obtaining information formats is described. Illustrations are drawn from a sublanguage grammar written for a subfield of pharmacology. Parts of the procedure are computerized or are being implemented.

Naomi Sager

New York University
Linguistic String Project
New York, New York 10012

- Introduction

It will be a major innovation in information systems when computer programs can detect the presence of particular information in natural language data bases purely on the basis of formal properties of the natural language texts themselves. It may then be possible to pinpoint, extract, compare and even tabulate specific information from texts in response to users’ queries. Such a capability depends on obtaining from the texts a regular, or formatted, representation of their contents. This seems at first like an impossible task. While the texts are physically a linear sequence of symbols, which lends itself readily to computer processing, the information carried by this sequence is vastly more complex. This disparity has discouraged investigators in the past from seeking a direct correspondence between the form and content of textual input.

Yet, the linearity of natural language texts is only a physical feature. There is a considerable amount of implicit structure in any stretch of connected discourse, and a great deal of such structure in texts within a well established scientific subdiscipline. The very fact that scientists within a particular subdiscipline understand each other, whereas their language often sounds to an outsider like a foreign tongue, indicates that there are restrictions on language usage within a science subcommunity (for purposes of communication within that subdiscipline) that carry specific meaning. While these facts are recognized in a general way, it heretofore has not seemed possible to characterize these restrictions in a systematic and useful way. This paper describes an investigation into the restrictions on language usage within a specific scientific subarea—in pharmacology—as an example of the kind of implicit information structuring which can be made explicit by the use of well-established, and in part computerized, language analysis techniques.

- Significance for Information Systems

The significance of this research for future developments in information systems lies in the result that an information format for the content of texts in a given subfield can be obtained. This format is a repeating pattern of term-classes and term-class relations in sentences of the text, where the term-classes and relation-classes are obtained by grouping into a single class the words or phrases which are most similar with regard to the other words or phrases they occur with in particular grammatical relations. The use of a method based on tabulated word-co-occurrences means that the analysis is not based on human judgments or prior

*This paper was originally presented as a talk at the Fall Joint Computer Conference, Anaheim, CA, Dec. 6, 1972. The work was supported by a Research Grant (RO1 LM 00720) from the National Library of Medicine, National Institutes of Health, Department of Health, Education, and Welfare.
semantic knowledge of the subfield. Nevertheless, the formats represent quite accurately the types of specific information in the text. Each “slot” of the format carries a particular type of information (e.g., in the pharmacology texts: presence of drug, drug action, physiological event—see below). When the format for a given subfield is established, text sentences in the subfield become instances of the format; and the particular slots of the format which are filled by portions of a sentence register the kinds of information the sentence contains. The formats are established by writing what is called a *sublanguage grammar* for the subfield. We can best illustrate the procedure by carrying through a specific example drawn from a pharmacology text.

### Sentence Analysis

To be specific, consider several sentences (shown in Fig. 1) from an article on the cellular effects of digitalis. These are the fourth and fifth sentences from a particular passage. We ask: What general methods can be applied to this linear string of words to obtain an organized record of the informationally important terms and their interconnections? The isolated words convey only a small part of the information; the basic unit of discourse is the sentence, in which the words appear in particular grammatical relations to each other. For example, in the simple sentence, *Digitalis reduces the concentration*, the verb *reduces* has the noun *digitalis* as its subject and concentration as its object. The fact that a verb requires a subject and object binds the words that satisfy this requirement into a tight grammatical unit centered on the verb, as illustrated in Fig. 2. These tight grammatical units are also the informational building blocks of the sentence.

It has been possible to obtain mechanically (i.e., by a computer program) a decomposition of the sentences of texts, where each sentence is broken into its elementary component word-strings, that is into tight grammatical units like that of SUBJECT-VERB-OBJECT illustrated above. Although the words in such a unit may be physically separated from each other in the sentence, the parsing program recognizes which words belong to an elementary unit and shows how the elementary units are grammatically interconnected to form the larger complex sentence. The program displays the analysis in an output parse, illustrated in Fig. 3. This is a slightly compressed view of the computer output, for the second text sentence of Fig. 1; the full parse printed by the computer names all the grammatical relations and assigns separate lines to prepositional phrases as well as to verb-containing units.

<table>
<thead>
<tr>
<th>PARSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>THIS RESULTS FROM THE SLOWING OF THE INFLUX OF POTASSIUM INTO THE CELL.</td>
</tr>
<tr>
<td>1. THIS RESULTS FROM 2,</td>
</tr>
<tr>
<td>2. THE SLOWING OF 3,</td>
</tr>
<tr>
<td>3. THE INFLUX OF POTASSIUM INTO THE CELL</td>
</tr>
</tbody>
</table>

Fig. 3

We note two things about this parse. First, the verbal center of a unit may “look like” a noun. This is not the case in line 1 where the verb *results* is an ordinary tensed verb. But in line 2, the verb *slowing* has an *ing* suffix, making it partially noun-like. And in line 3, the word *influx* has the form of a noun, though we recognize the grammatical relation of *influx* to *flow in*. Secondly, we note that the subject or object of a verb may itself be one of the elementary grammatical structures mentioned above. *Qua structure*, it occupies a line of its own in the output parse, but to indicate that the structure is the object of a particular verb, the line-number where the structure appears is written in the object position following the verb whose object it is. There can thus be a hierarchy of verbs, each one operating on the next one, provided the “next one” takes on the appropriate noun-like form. Thus, in line 1 of Fig. 3, results from operates on the verb *slowing* in line 2, which in turn operates on the verb in noun-form in line 3; in line 3, *influx* is a noun-form verb which has only concrete nouns (*potassium* and *cell*) linked to it. Schematically, the verb hierarchy in the parse has the form shown in the bottom right of Figure 4.

It turns out that the grammatical hierarchy of verbs has a direct informational correlate. In the parses of texts in a given field of science the lowest levels of the parse tree are satisfied by concrete nouns and verbs.
specific to that field of science, the object language, as it were, of that science, e.g., potassium, influx, cell, in Fig. 4. The highest levels of the parse tree subsume “intellectual” verbs, e.g., assume, study, show, demonstrate, etc., whose subject is human, and whose object

is a form of a sentence. We write $V_S$ for these verbs to indicate that their object is a sentence. $V_S$ verbs occur in all scientific writing, mainly in the passive, as illustrated in Fig. 4 by it was assumed that. The verbs between the highest and lowest levels, in the pharmacological texts at any rate, are mainly of two kinds: verbs $V_Q$, like slowing, increases, whose characteristic object is a quantity noun or verb and sentence-relating verbs $V_{SS}$, like results from, whose subject and object are both sentence-forms or their stand-ins, e.g., this in Fig. 4. Not every sentence contains words representing every level of the hierarchy.

The type of structure illustrated in Fig. 4 is already a step towards an information format for the sentence. The “higher-level” vocabulary, conveying human intervention, is separated from the science-specific vocabulary which concerns the objects of interest in the science. It is important to realize that this structuring is obtained entirely syntactically by a parsing program employing a general grammar of English and a lexicon in which the individual sentence words are coded only in respect to gross grammatical classes of English: Noun, Verb, Singular, Plural, etc., not with regard to their semantic properties, or special usage within a discipline.

In view of the success of the computer program in obtaining a crude informational segmentation of sentences, we may ask if the same methods can be carried further to obtain a more refined analysis. The parsing program uses a grammar of English with no special rules for science texts. But articles in a scientific journal are, after all, constrained more tightly than by the rules of English alone. The question which is raised is whether these constraints can be formulated analogously to grammatical rules, and whether such rules, if established, lead to a general statement of the regularities of specialized word-use in particular scientific disciplines. The answer in regard to the investigation of one such area was that it is indeed possible to state in a regular way what are well-formed sentences in a given discipline. The subfield grammar rules are more restrictive than English grammar rules. A sentence may be well formed in English and not well formed as a sentence in the subfield. For example, compare an English grammar rule with a subfield English grammar rule, illustrated in Fig. 5.

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. The subfield grammar provides a method for developing the special word sets and the particular relations among these sets, 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 those 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 subfield grammar. This result can best be illustrated by sketching the pharmacology subfield grammar which was obtained and indicating the methods which were used.
We start with the nouns and verbs which appear at the "bottom" level (lowest lines) of the computer parses, and collect into classes the nouns which are most similar with respect to the verbs they occur with, and similarly for the verbs with respect to the nouns they occur with. The nouns fall into largely disjoint sets (Fig. 6) which occur characteristically as the subject of particular verbs, or the objects of particular other verbs. Thus, the words *ion*, *sodium*, *potassium*, *electrolyte*, occur characteristically as the subject of flow in, flow out, move, accumulate, etc., as in the influx of potassium, the accumulation of intracellular Na, and as the object of transport, as in the transport of Na and K. We name this noun-set ION. It is interesting that in this subfield literature *sodium* and *sodium ion* are synonyms in the ION set, whereas they are clearly not so in other disciplines.

The words in the CELL-class, on the other hand, occur as the object, not the subject, of verbs like flow into, diffuse out of, as in K diffuses out of the cells, and as the subject of verbs like recover, is incubated, which do not take ION-words as their subjects. In detail, the distributions of the words in a class differ, and it is possible to split the classes into smaller and smaller units, reflecting the distinct properties of the real-world objects named by the words in a class. In practice, a cut-off point is reached in which the divisions reflect gross differences, e.g., between tissue words vs. drug words. While gross for the subfield, these divisions are detailed enough to characterize the major objects of interest in the subdiscipline.

In Fig. 7, we see a small sample of the lowest level verb classes, which are established on the basis of which noun-sets they occur with. Verbs like diffuse into and many other verbs of movement in this science (entry, inflow, outflow, accumulate) connect words in the ION class to words in the CELL or TISSUE class. We see this, for example, in such sentences as *K is actively accumulated and passively "leaks" out of the cell* whereas *Na is actively extruded [from the cell] and passively "leaks" in*, and also where the verb occurs in nominal form, as in the efflux of K in blood, ion movements in various tissues. The ION connecting verbs (replace, exchange with, compete with) occur typically in such sequences as cation exchange, the competition between Na and Ca or the competition of these ions for a receptor site, transcellular ion exchange processes, etc. Examples of the intransitive verbs which occur with the CELL or TISSUE class as subject are contract, beat, recover, as in the contraction of heart muscle. This class also includes some experimental verbs in the passive form, as in incubated cold-stored red cells. Fig. 7 shows only three of perhaps 15 "bottom level" verb classes (where "bottom level," again, means the lowest level in a parse diagram such as is shown in Fig. 3).
QUANTITY WORDS Q, VQ

TOXIC DOES OF DIGITALIS

CHANGES IN THE RATE OF MOVEMENT OF ELECTROLYTES INTO THE CELL

Fig. 8

Sometimes we find VQ operating directly on another V, without a Q in between, e.g., *the change in concentration, the slowing of the influx*. These operand verbs can be shown to contain a "hidden" quantity word in them, as can be seen by their often having dimensions associated with them, and by the possibility of inserting a quantity word without changing the meaning. Thus the slowing of the influx may have units associated with it, and can be acceptably paraphrased by the slowing of the rate of influx.

There are also verbs, like *equals, depends on, correlates with*, which connect two quantity words (Q, VQ, or V with hidden Q), as in *the amount of K-loss has correlated with the degree of augmentation of contractility, the rise in concentration depends on the number of beats*, there is a correlation between the activity of ATPase and the amount of pumping. There is even a verb (*parallel*) which occurs only as a connector between two quantity verbs: *The increase in tension parallels the increase in uptake*.

The richness of the grammar of quantity words in this subfield language reflects the critical importance of quantity in this science.

The next level brings in the sentence connecting verbs VSs, illustrated in Fig. 9. The sentences which appear on either side of VSs have their verbs in nominal form. Thus in Fig. 9, *could interfere with* is the connecting verb between *increasing extracellular Ca and the entry of K into the RBC* (red blood cell). Many of the sentence-connecting verbs sometimes occur with a particular noun-class NQ instead of a nominalized sentence as their subject, as in *Digitalis interferes with the flux, a digitalis induced augmentation of contractility*. NQ is the noun set which includes all the pharmacologically active agents.

At the level of sentence connecting verbs, we may have instead a verb VS (or a noun or adjective) which operated on a single sentence: *It is well known that digitalis can inhibit this transport*.

All of the nouns and verbs considered can have local modifiers on them, following ordinary English grammar, and the sentence may be connected by the subordinate or coordinate conjunctions of English. Certain adverbial modifiers of the sentence, those containing clinical or experimental vocabulary, have been called Ds.

- **Information Formats**

A subfield grammar provides a basis for structuring the information in each sentence and for mechanically processing the structured information. It has proved possible to construct a fixed format of the subfield grammatical operators and operands, which houses all the sentence decompositions obtained using the subfield grammar. Sentence segments fall into "slots" of the format depending on their subfield grammatical properties. Each slot has a fixed informational character, and each sentence carries the type of information of the slots which it fills. (The formatting has not been automated yet.)

The format for the elementary sentence is shown in Fig. 10. There is a slot for the subject N, the verb, and the verb object, which may have several parts. Slots for quantity words are also provided in accord with the analysis of quantifiers in the grammar. This particular formatted word sequence contains an ION-word as subject and a CELL-word as object. The verb is of the type which connects ION to CELL or TISSUE, and is shown still in its nominal form for which the grammatical constant *(has)* has been supplied; the object preposition is shown tied to *cell*, as it appears in the sentence.

![SENTENCE-CONNECTING VERBS, VSs](image)

![SE FORMAT](image)
In Fig. 11, we see a larger portion of the format. Each line contains a unary sentence composed of an elementary sentence \( S_E \) with its hierarchy of immediate operators, \( V_Q, N_O, D_S \). The formatted sentence is the first text sentence of Fig. 1: Toxic doses of digitalis consistently reduce the intracellular concentration of potassium in a wide variety of cells, including cardiac muscle cells. In 4.1, the elementary sentence is intracellular potassium concentration, which we saw formatted previously. Reduces with its local adverbial modifier consistently is a quantity verb \( V_Q \) operating on concentration. Digitalis with its quantifier toxic doses is an \( N_O \) operating on reduces. A regularization of the subfield grammar would factor reduces into the \( V_{SS}V_Q \) sequence causes lowering. In a wide variety of cells is an adverbial phrase which operates on the whole preceding sequency. Bridging 4.1 and 4.2 is the conjunctional verb including. 4.2 repeats 4.1 up to the \( D_S \) column because the part of the sentence following the conjunction contains an implicit repetition of the part preceding the conjunction, up to the contrasting adjunct. We have in fact a second sentence after the conjunction: Toxic doses of digitalis consistently reduce the intracellular potassium concentration in cardiac muscle cells.

![Fig. 11](image1)

Fig. 11 shows the format for the second text sentence of Fig. 1. This is holding the place of an entire unary sentence, as it should as a sentence pronoun, or "pro-sent." The sentence connecting verb is results from. The second unary sentence consists of an elementary sentence \( S_E \) with a \( V_Q \) operating on it. The \( S_E \) is again of the type which has an ION-word as subject and a CELL-word as object with a movement verb connecting them.

![Fig. 12](image2)

A summary of the format for the object language portion of the majority of sentences in this subfield is shown in Fig. 13. This grammatical structuring has a clear semantic interpretation relevant to the science. The deepest elementary sentences are non-quantitative statements in cell physiology and biochemistry, mainly movements in cells and extra-cellular space. The quantifiers and their verbs quantify these statements, and the immediately next set of operators relates the drugs to the cellular events. The conjunctions and sentence connecting verbs make it possible to state the relations between events while the higher level operator verbs with human subjects (not shown in the formats) express the scientist’s relation to the events.

![Fig. 13](image3)

### Discussion and Conclusion

The possibility of obtaining a representation of the contents of natural-language science texts without resorting to surrogates or a priori semantic categories is a surprising, if not startling, result. In undertaking this investigation we assumed from prior work with language that a correlation between word co-occurrence patterns and content could be established on some level in restricted subject matter areas. We ourselves surprised, however, at the detailed correlation of co-occurrence-based classes with semantic categories in the subfield. The possibility of formatting the information in texts, again without recourse to a priori categories, was an unanticipated result.

With regard to applications, it is still too early to say where this type of processing will be of most use in the information processing field. Its greatest asset is depth and precision, and the condition for its employ is a small, well-delineated subject matter area. One area of potential application is medical records, where certain parts of the records are customarily written or dictated in natural language, yet the standard expressions used and the repeating specific situations described suggest that a sublanguage structure may not be far beneath the surface. Some work in this area has been reported (5, 6).
There are many questions raised by these results. We would like to know how sublanguage grammars react to changes with time in the knowledge reported in the literature of the subfield (some preliminary work has been done on this within the pharmacology study). We would like to know how much of the sublanguage grammar is shared by neighboring subfields, and by distant subfields. We are also concerned with the computability and the cost of the methods. One thing is certain. Language is our major medium of reporting and storing information. It is likely that better understanding of the ways in which language carries information will lead to better systems for its storage, retrieval and utilization.

References