The Many Faces of Information Science

Edited by Edward C. Weiss

Published in 1977 in the United States of America by Westvew Press, Inc.

AAAS Selected Symposium 3

3

Information Structures in the Language of Science

Naomi Sager

INTRODUCTION

This paper presents results, and computer applications, of research into the relation between language structure and information, particularly as it appears in the language of science.

Information is not something separate from language. It is true we convey some meanings by extralinguistic means, but for all practical purposes, the way of storing and transmitting information, and probably of forming new information, is largely via language. To study the relation between information and language, our method has been to analyze scientific writing in a systematic way, using syntactic and statistical methods which can be applied with little change to written material from many sciences.

Using the regularities observed in the language material itself, we have developed computer programs for processing the information in natural language reports and articles. The programs convert the natural language text of the input documents into table-like structures (called information formats) by aligning words which carry the same type of information into a single column. The columns of the table are defined in such a way that the syntactic relations between the words of the sentence are preserved in the table. This way, no textual information is lost, and the original sentences, or paraphrases of them, are reconstructible from the table. At the same time, this mapping of the text sentences into formats makes the information in the text accessible for further computer processing and brings it into line with information presented in other forms, such as tables published in the literature.

SIMPLE PARSE DIAGRAM

GL 641 13,6,11 CALCIUM UPTAKE INTO LIVER MITOCHONDRIA APPEARS NOT TO BE AFFECTED BY CARDIAC GLYCOSIDES.

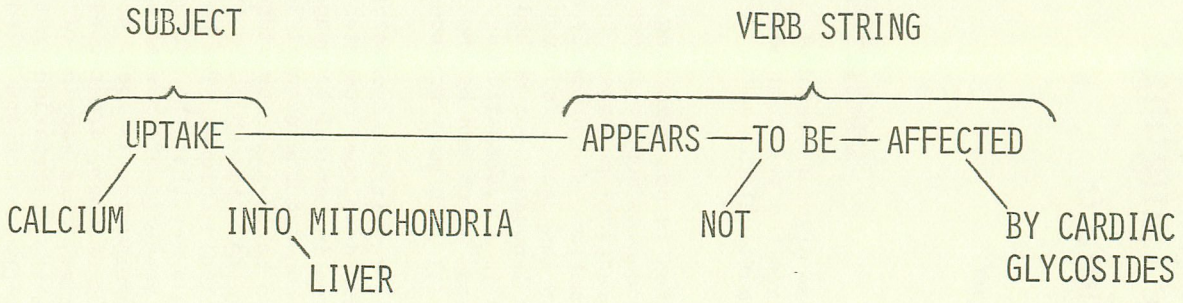

FIGURE 1

The main practical implication of these information formatting programs is that large files of technical documents on a given subject could be queried by computer for particular information, or summarized with respect to particular categories, without the necessity to code or alter the input natural language documents. A pilot experiment on radiology reports of cancer patients demonstrated that the computer system was able to transform the sentences of the English language reports into the appropriate tabular structures without loss of information, and to retrieve specific factual information from the computer-formatted reports. For each report the system was able to format the sentences of the report and from the resulting tables to answer such questions as: Was a test made (for given patient during given period)? Were the findings negative? Is there some question about the findings? When was the first metastasis reported? Where was it? and other questions (1).

It should be noted from the outset that these programs are not based upon semantic categories that are supplied beforehand by someone with knowledge of the given science. They are based on general properties of language structure and on the fact that words with similar informational standing in the science occur in similar positions vis a vis other words in the texts. We have demonstrated by means of a clustering program (to be described later) that the word classes of semantic value in a science subfield can be generated on the basis of the distributional similarity of the words in the subfield texts. For example, Ca^{++} and Na^{+} are found to be in the same class in some of our texts, not because they are known to be names of ions, but because they both occur as subjects of the same type of verb in the textual material.

Computerized Language Processing

It is only possible for the computer to convert the information in science documents from their natural language form into more regular forms by building upon the regularities which are in the language material itself. These regularities exist on two levels; one common to the language as a whole and one specific to the subject matter. The regularities which obtain for a whole language are summarized in its grammar. The first stage of computer processing, then, is to analyze each input sentence as an instance of a grammatical structure, specified in a computerized grammar provided to the program. The program which does this, a so-called "parser", segments each sentence into its major grammatical components (main clause, modifiers, etc.) and shows how the components are interconnected [Figure 1]. Parsing is important not only as a first step in breaking-up a large unwieldy

sentence into smaller more tractable units, but also because some of the grammatical relations recognized at this stage are themselves part of the information; to use a simple example, the relation of subject-verb-object in an assertion makes the difference between, e.g. the ion enters the cell and the cell enters the ion.

But while the parsing program provides a useful decomposition of the sentence into its grammatical components, the language provides for so very many different kinds of grammatical components that further regularization is necessary. Many grammatical forms are equivalent with regard to the substantive information they carry, for example, the active and passive forms. These equivalences can be utilized by the computer in order to reduce the number of alternative grammatical forms that have to be dealt with. The most common one among the equivalent forms is chosen as the base form and the program is equipped with procedures which transform occurrences of the equivalent forms into the base form. Very often the transformation fills out elliptical assertions into full assertions. This not only reduces the number of forms to be dealt with but regularizes the pattern of word occurrences which is important for informational alignment. As a simple example, the sentence Samples of Na_2SO_4 were irradiated and analyzed contains the segment and analyzed, consisting of and plus a participle. The computer eliminates this special form by expanding the segment to a full assertion, as though the sentence read: Samples of Na_2SO_4 were irradiated and samples of Na_2SO_4 were analyzed. From here it is a straightforward step to align corresponding parts of the assertions. In addition, in this example, the computer could reduce the two passive assertions to two active assertions with unspecified subjects: Someone irradiated samples of Na_2SO_4 and someone analyzed samples of Na_2SO_4. This would be useful if elsewhere in the texts active forms involving the same words occurred; but often we find that this is not the case. The so-called scientific passive serves as a better base form for much science material, especially laboratory procedures and measurements, where one wants to align both properties and procedures as predicates on the experimental material.

Sublanguage Grammars

The second type of regularity which appears in science writing is not common to the language as a whole, but is specific to the particular subfield of science from which the texts are drawn. One has to realize that journal articles and technical reports are communications between specialists who "talk the same language." Here the metaphor has literal

truth. Investigators or practitioners in a given field speak a language which is not identical to the over-all language, say English, even though they use English words and do not violate English grammar. They speak a sublanguage which differs from English in several ways. It does not use the full range of constructions permitted by the grammar of the whole language, and it is constrained by rules that do not apply in the language as a whole.

In the case of a whole language we know that there are grammatical rules because some word sequences are accepted by native speakers of the language as wellformed sentences while other sequences are rejected as ungrammatical. A similar situation exists in a community of individuals engaged in a specialized field of science. Certain statements will be accepted as possible within the discipline while others will be rejected as impossible or outlandish. I do not speak here of truth versus falsity or even accepted versus unconventional formulations, but of statements which run counter to common knowledge which is fundamental to the discipline. Thus, for example, the statement the ion enters the cell would be acceptable to a specialist working on cellular processes--it may or may not be true in a given case--whereas the cell enters the ion would be definitively rejected as being not merely false but unsayable in the science. This linguistic behavior on the part of the scientist indicates that rules analogous to the rules of grammar for the whole language are operating in the language of a science subfield.

It is by making explicit the regularities of language usage on the subfield level, that we are able to construct formats for housing the information in subfield texts. Just as a grammar provides syntactic formulas in terms of the classes Noun, Verb, etc., a sublanguage grammar provides analogous formulas in terms of those specific subclasses of Noun, Verb, etc. which are characteristic of the subfield, e.g. in cell biology, classes for ions, molecules, cell structures, verbs of motion, verbs of cause, etc. These subclasses are found by clustering (manually, or by computer) words with similar co-occurrence patterns vis a vis other words in the texts. The sublanguage formulas are thus summaries of syntactic regularities in texts constrained by a particular subject matter. They emerge as formats for the textual information because of the close relation on this level between form and meaning. Ions do only certain things; therefore ion-words occur as the subject of only certain verbs.

GL641 13.6.11 CALCIUM UPTAKE INTO LIVER MITOCHONDRIA APPEARS NOT TO BE AFFECTED BY CARDIAC GLYCOSIDES.

DRUG	V-CAUSE	ARG1	V-PHYS	ARG2	CONJ
CARDIAC GLYCOSIDES	AFFECT (APPEARS NOT TO)	CALCIUM	UPTAKE INTO	MITOCHONDRIA (LIVER)	.

FIGURE 2

INFORMATION FORMATS

To illustrate what is meant by an information format for science writing, consider the formatted sentence in Figure 2. This example, and several others in this paper, are taken from a study of journal articles in a subfield of pharmacology concerned with the mechanisms of action of digitalis and other cardiac glycosides. We analyzed manually, and with the aid of the computer, some 200 journal pages in this field, and found that sublanguage formulas covering the main factual results could be stated in the form of a sublanguage grammar, and that the sublanguage grammar could be used in procedures to map the text sentences into a limited number of format structures.

Fact Units

The formats obtained in the pharmacology study contained one case or another of the basic unit illustrated in the format in Figure 2 for the sentence Calcium uptake into liver mitochondria appears not to be affected by cardiac glycosides. In this sentence, as in most others in this material, there is an inner, or "bottom level," assertion (shown in the format between double bars) consisting of a verb with its subject and object. In the pharmacology texts, this assertion described an elementary physiological or biochemical event, which in the case of Figure 2 is the uptake of calcium into the mitochondria of the liver. The inner assertion here is an instance of a formula that recurred over and over in these texts, $N_{ION}\ V_{MOVE}\ N_{CELL}$, in which a noun in the ion class is connected to a noun in the cell or cell substructure class by a verb in a class which expresses movement, though the class is defined by its syntactic position connecting the above two noun classes. Examples of other elementary assertions encountered in this literature were those covering ion interactions, enzyme activity, tissue contraction or contractility, protein behavior and ions binding to molecules.

Operating on the elementary assertion, very often, was a noun-verb pair, shown in Figure 2 to the left of the double bars, consisting of a drug word and a verb of roughly causal character (affect, influence, etc.) possibly negated or quantified, as in this sentence. The causal pair is said to "operate on" the elementary assertion because the latter appears as the object of the causal verb. Notice that we had to perform the regularizing transformation passive → active in order to reveal that uptake is the object of affect, since it appears in the sentence as the subject of the passive construction appears not to be affected. Also, while uptake appears in the sentence as a noun, it is in fact a

GL 641 2.2.1 MORE DETAILED STUDIES OF THE AFFECTS OF CARDIAC GLYCOSIDES ON SODIUM AND POTASSIUM MOVEMENTS IN RED CELLS HAVE BEEN MADE BY KAHN AND ACHESON (99), SOLOMON ET AL (168) AND GLYNN (67).

HUMAN	V-STUDY	DRUG	V-CAUSE	ARG1	V-PHYS	ARG2	CONJ
K AND A (99) S ET AL (168) AND G (67)	HAVE MADE MORE DETAILED STUDIES OF	{CARDIAC GLYCOSIDES	AFFECT	SODIUM	MOVE IN	RED CELLS	AND
				POTASSIUM	[MOVE IN]	[RED CELLS]}	.

FIGURE 3

nominal form of the verb take up, so it is mapped into the verb column. A transformation hunts for the arguments of the verb among the adjuncts of the nominal form (uptake) and maps them into the verb-argument slots. As illustrated in the format of this simple sentence, the major fact type in this pharmacology material was composed of an elementary assertion drawn from a prior science (cell physiology, biochemistry), with the pharmacological agent entering only on a higher grammatical level, as an operator on the elementary assertion.

Fact vs. "Meta-fact"

The somewhat longer sentence formatted in Figure 3 utilizes format columns that were not shown in Figure 2 because they were empty there. Notice the two new columns on the left, labelled HUMAN and V-STUDY. Factual assertions involving only the concrete objects of investigation in the science and their interrelations (the two inner sections of the format) are syntactically separable by the computer from the words describing the scientists relation to the facts. Verbs like study, present, discuss, assume, report, which have exclusively human subject nouns and carry the connotation of the scientists' intellectual activity appear as higher level operators on the operator-structure already built up from the words in the "object language" of the science.

Notice also in Figure 3 that there is a conjunction column CONJ on the right which contains words that connect one line (or several grouped lines) of the format to another line or lines. This is a major departure from tables for quantitative data. Here the conjunction is and, but in other cases the conjunction may have the form of a verb or a phrase (e.g. is associated with, is the basis for). Apart from grammatical conjunctions,only words which have the syntactic property of operating on a pair of (nominalized) sentences are accepted in the CONJ column. The words in the CONJ column are much the same in different subfields, whereas the words in the innermost columns are highly specific to the field.

A last point to notice in Figure 3 is the presence of reconstructed word occurrences, shown in square brackets. The conjunction and is responsible for ellipsis in this case. Sodium and potassium movements in red cells can be expanded to sodium movements in red cells and potassium movements in red cells on the basis of general grammatical properties of the conjunction and. The expansion of the phrase into two assertions does not imply that the events are independent of each other; only that the connection between them is not more explicit here than their conjoining by and.

LA 721 1.1.5 THE POSSIBILITY THAT ADMINISTRATION OF DIGITALIS, THROUGH ITS INHIBITION OF THE NA+ - K+ COUPLED SYSTEM, PRODUCES AN INCREASE IN NA+ - CA++ COUPLED TRANSPORT AND THEREBY AN INCREASE OF INFLUX OF CA++ TO THE MYOFILAMENTS IS DISCUSSED AND IS PRESENTED AS A POSSIBLE BASIS FOR THE MECHANISM OF DIGITALIS ACTION.

HUMAN	V-STUDY	DRUG	V-CAUSE	V-QUANT	ARG1	V-PHYS	ARG2	CONJ
[AUTHOR]	DISCUSSES {1 2}	DIGITALIS (ADMINISTRATION OF)	PRODUCES POSSIBLY	INCREASE	NA+ - CA++ COUPLED	TRANSPORT		AND THEREBY
		[DIGITALIS (ADMINISTRATION OF)]	[PRODUCES]	INCREASE	CA++	INFLUX TO	MYOFILAMENTS } 2	THROUGH
		[DIGITALIS] = ITS	INHIBITION		NA+ - K+ COUPLED SYSTEM		} 1	AND
[AUTHOR]	PRESENTS	← [{ }] 1 1 →						AS BASIS FOR (POSSIBLE)
		DIGITALIS	ACTION MECHANISM					

FIGURE 4

Data Structures vs. Argument

A third formatted sentence, shown in Figure 4, is sufficiently complex so that it illustrates in itself some of the regularizing effect that formatting achieves for a whole text. When the sentence is read without reference to the format, it is not at all apparent that there is so much repetition of similar elements. As the format shows, the sentence consists of 4 interconnected factual units of the same basic type. The texture, and the intellectual content, comes from interrelations among similar data structures, in the use of conjunctions at different levels of grouping, in the introduction of qualifying modifiers and higher level operators, and in the use of reference, either explicitly via pronouns or implicitly via ellipsis. These features belong to the argument or reasoning in the text, which can be separated from the factual units mapped into the inner portions of the format lines. Turning first to the individual fact units in Figure 4, the inner portion of the first line, stripped of its qualifiers, says that digitalis produces an increase in Na^+ - Ca^{++} coupled transport. In this unit, Na^+ - Ca^{++} coupled transport is an instance of the formula N_{ION} V_{MOVE} N_{CELL} seen previously, even though the cell-word is not present here. In oft-repeated material, the subject or object of the verb is frequently dropped, or sometimes the verb if it is unique to the stated subject or object. This is the case in the third line, where transport is suppressed but easily reconstructed because of the subject, Na^+ - K^+ coupled system.

Notice in Figure 4 the presence of a new column V-QUANT between the innermost assertion and the columns DRUG, V-CAUSE. The V-QUANT column was not shown in preceding figures because no words like increase, decrease, etc. were present in the sentences. In line 3, the V-QUANT column appears to be empty. But in effect the word inhibition covers both the V-CAUSE and V-QUANT columns, since elsewhere we find in similar contexts that inhibit and cause a decrease in are used interchangeably.

The format in Figure 4 introduces the use of pronouns and other devices of reference. In the third line, the antecedent of the pronoun its, namely digitalis, has been reconstructed as the subject of inhibit. This follows the pattern throughout, that the class of pharmacological agents occurs as the subject of verbs in the V-CAUSE class. (Syntactically, in this sentence, the entire phrase administration of digitalis may be the subject of inhibit; but it matters little to the representation of the information in the sentence, since in this sublanguage, digitalis and the administration of digitalis are used interchangeably as

ON THE DAY OF ADMISSION SWELLING WAS NOTED OF THE LEFT TIBIA AND FOOT AND WAS ASSOCIATED WITH TENDERNESS. SHE WAS SEEN THE NIGHT BEFORE ADMISSION IN THE EMERGENCY ROOM BECAUSE OF A TEMPERATURE OF 105, AND NO CAUSE NOTED. THERE WAS NO EVIDENCE OF UPPER RESPIRATORY INFECTION.

			TREATMENT			PATIENT STATE					TIME				
	CONJ	PATIENT	INST	V-TREAT	T-P CON	BODY-PART	BODY-MEAS	QUANT	SIGN/SYMPT	EVIDENCE	P	Q	UNIT	P	REF. PT.
1						LEFT TIBIA			SWELLING	WAS NOTED	ON		THE DAY	OF	ADMISSION
2	AND					[LEFT] FOOT			[SWELLING]	[WAS NOTED]	[ON		THE DAY	OF	ADMISSION]
3	AND	← 1-2 →													
4	WAS ASSOCIATED WITH								TENDERNESS						
5	·	SHE	EMERGENCY ROOM	WAS SEEN	BECAUSE OF		TEMPERATURE	105					THE NIGHT	BE-FORE	ADMISSION
6	·					UPPER RESPIRATORY			INFECTION	THERE WAS NO EVIDENCE					

FIGURE 5

subjects of V-CAUSE verbs.)

The fourth format line has the interesting property that the whole object language, or factual, portion of the format is empty of physically occurring words. The three preceding format lines, seen as a unit, are repeated implicitly as the first operand (subject) of the binary relation is a basis for, where the second operand (object) is the mechanism of digitalis action. Reasoning in science writing is characterized by devices of this sort. A single assertion becomes a nominalized sentence within another sentence; a sequence of interconnected sentences becomes an element of a later sentence by implicit repetition or by pronominal reference (this, this process, etc.). In this way it becomes possible for complicated interrelations to be expressed in the physically linear medium of language.

PROPERTIES OF SCIENCE INFORMATION

From a study of information formats in different subfields, one gets a picture of how scientific information is carried by language both in respect to the unique informational characteristics of each science and in respect to the general properties of information viewed over science as a whole.

First, the information formats of a particular science reflect the properties of information in that particular science in contrast with other sciences. The pharmacology formats, for example, displayed a characteristic predicational hierarchy in which the pharmacological agent occurred on a higher grammatical level as an operator on an inner sentence from cell physiology or biochemistry, reflecting the role of the drug as an outside element that affects on-going processes. Quantity and quantity-change were important in the pharmacology formats (not all quantity columns are shown in the example formats, e.g. dosage) reflecting the importance of quantity relations in this science. This type of format contrasts with one that was obtained for a corpus of clinical reporting, where both the columns and the relations among the columns were quite different. An example of the clinical format is shown in Figure 5 in a simplified version and without further explanation simply to illustrate a different information structure. In the clinical format, time columns are essential to the information, whereas they were almost entirely absent in the pharmacology formats. In the clinical formats, there is very little predicational hierarchy and virtually no argument, both of which were present in the pharmacology formats. The structure of the clinical information, displayed in the formats, is an interplay

between columns containing treatment words and columns containing words that describe the patient's state; successive rows are linked primarily through time sequence, with the conjunction columns playing a secondary role.

While the formats for different subfields differ, as they should, to capture the specific character of information in each field, they have certain properties in common that hold for all of science writing. To mention just a few:

(1) Statements about science facts are separated by the grammar itself from the science facts proper; the role of the human investigator is syntactically separable from the report of factual events.

(2) The report of a complex event has a structure composed of a hierarchy of different types of operators, the ultimate "bottom level" operand being the carrier of the most elementary objects and events. When a given science draws upon a prior science, the material from the prior science appears as the operand of material from the given science.

(3) Argument is carried by connectives between the data structures built up in this hierarchical fashion. Whole units are carried forward by the telescoping of an operator-hierarchy into a single noun phrase or a substitute "pro-word," or by the controlled dropping of words permitted by the grammar.

(4) What is universal is the amount of repetition and regularity that is found in all science writing, once stylistic variations and equivalent grammatical forms are eliminated. Every individual piece of writing contains some repetition (or it would not be connected discourse). Across a single specialized discipline, the same items repeat in different combinations and with variations, as though all the texts were part of a single extended discourse. Although the texts each bring in some new feature they are sufficiently similar as to fit into an overall formulaic characterization. These formulas, or information formats, are then a powerful tool for organizing information on a given topic, when that information comes from diverse sources and in diverse forms.

The common features in information formats, noted in (1)-(4) above, suggest that a generalized matrix for all factual writing in science is a possibility. Such a matrix could provide guidelines for the development of new types of data structures in computerized information systems, which in the future should be able to handle information from the natural language part of texts as well as citation information, numerical data, and other forms of information. At present, it appears that this matrix could be tabular in form (a numerical table would

be a special case), but should allow for more than two dimensions, and in particular would have the following special features:

(a) Columns or groups of columns could be specified to have a hierarchical operator-operand relation between them (expressing levels of predication if present).
(b) A distinguished column or columns would carry connectives between successive rows or groups of rows (expressing conjunction in language).
(c) Provision would be made for pointers from an element in one row to an element of another row, or to an entire row or group of rows (accommodating pronouns and other referentials).

Even at this early stage of work on a generalized matrix, we have some information with regard to the types of columns that can be expected, for example, EVENT, QUANTITY, TIME, CONDITION, CONNECTIVE, though not every science will have all of these. The EVENT columns will always be present; every science has some elementary assertions concerning the primary objects of investigation. While the content of the sub columns of an EVENT unit varies from science to science, the existence of "bottom level" elements and relations is universal.

In view of the importance of quantitative data in many sciences, columns for QUANTITY and QUANTITY-CHANGE will undoubtedly often be needed. Fortunately, quantity words have distinctive linguistic properties which enable them to be recognized by the computer and treated specially. The same holds true for time words and the mapping of time words into TIME columns.

Many factual statements in science contain a statement of the conditions under which the statement applies or under which the reported observation was made. These conditions usually occur syntactically as adverbial modifiers, and hence can be sent into CONDITION columns, even if the content is not repetitive enough to have special columns for it in the matrix. It will be extremely interesting to see how the selection of specific columns varies from field to field as the matrix is further tested.

The CONNECTIVE column is universal. Through it are channeled the links between individual facts, which links are important both for retrieval and for further characterization of the information content. A wealth of data about scientific information lies in the contents of the CONNECTIVE column. The procedures we have developed isolate the factual units and align them. What is left, if it is not "meta-fact" (the scientist's own relation to the facts that goes into its own

columns) is connective material. The connectives are not just grammatical conjunctions, but also verbs and other expressions which carry causal and other relations. A study of this material would yield a great deal of practical insight into the "grammar of science." Since we know that on the one hand the connectives of logic are not sufficient to carry scientific reasoning and on the other hand the full power of natural language is too rich, we could home in, by empirical study, on just what types of connectives are used, and possibly develop sets of synonym classes which would represent the basic connectives in most frequent use. Such a result would have wide implications both for practical information processing and for the philosophy of science.

COMPUTER PROGRAMS

A battery of computer programs have been developed to do the language analysis and information formatting described above. The basic tool, without which the long, complicated sentences of scientific writing could not be machine-analyzed, is a large computerized grammar of English (2). This grammar, which required a reorganization of language data into a computable system, was developed over a period of years with support from the National Science Foundation. The grammar is applied to sentences by a parsing program which has gone through several implemented versions (3, 4). The latest version (5) runs on the CDC 6600, and includes a special programming language (6) and its compiler as part of the system, as well as a component for executing transformational procedures (7). Among the implemented transformations, one worthy of special note because it overcomes a special complexity of language is the procedure for expanding conjunctional sequences to their full explicit form (8).

Mention has been made of the fact that the sublanguage word classes are defined on distributional rather than semantic grounds. A clustering program (9) was written to do just that: to calculate similarity coefficients for all word pairs based on the frequency with which they both occurred in the same grammatical relation to the same other words, and to form classes of those words whose calculated similarity coefficients were higher than a given threshold. Details of this algorithm are given in the paper cited. It was found that the word classes formed in this way correlated well with classes that had been defined semantically for the same material.

The information formats for a given subfield are developed by people, not by machine, even though a number of the components for doing the job are at hand: the programs for

sentence analysis and transformation, and the program for sublanguage word-class generation. Nevertheless, it is a task of some complexity to abstract the major patterns or formulas of subfield information.

The major use of the computer programs thus far is in subfield applications. Once the information format is clearly specified and the word classes are defined, formatting transformations are written which map the output of the parsing and regularizing programs into the columns of the implemented format. To date, we have had the opportunity of carrying out the complete process on a sample of English-language radiology reports (1, 10) and currently on clinical records (hospital discharge summaries). In principle, and I believe in practice, the methods should apply in any subfield where the subject matter is relatively circumscribed and words are used in a relatively constant way. An initial "tooling up" is of course required for each subfield application, in order to develop the specific formats and subfield word dictionary required by the formatting programs.

In addition, it should be stressed that each subfield application is not a programming task that begins afresh. On the contrary, both the computer tools and the type of information structures produced are quite general. With regard to tools, it must be clear that if the grammar of English used by the computer did not have a broad coverage of the language, it would not be possible for the system to analyze sentences from many different kinds of texts, which it has been shown able to do. In addition, a large amount of grammatical detail is required in order to obtain the correct parse (or a small number of parses if the sentence is actually ambiguous). The computer grammar scores high in this regard; in applications, our experience has been that the first parse formats correctly in over 85% of the sentences.

With regard to information structures, the general matrix for science writing described above provides the framework for the implementations of formats for specific subfields. Each application is approached as a problem in tailoring the general programs for use on a particular subject matter. Also, particular features that repeat are generalized. The implemented framework approaches more and more a general program for formatting the information in scientific documents.

FUTURE INFORMATION SYSTEMS

Computer programs for processing natural language are reaching the stage of application at a time when changes in

the technology of information production and dissemination are making computerized natural language data bases widely available. In addition to the well-established library-oriented data bases, there are now many large special-purpose computer stores in the form of files and records, whose magnitude is such that there is a need for computer programs to access and summarize the different kinds of information in the documents. In this setting, a computer capability for processing natural language takes on practical importance.

The fact that large computerized stores of information in natural language form are being created in publishing and in record-keeping is in itself creating a demand for techniques which can process data in natural language form. This is true not only in the area of official scientific publication, where a change in printing technology makes such data bases possible, but also in the area of file management, where institutions in medicine, industry, and government find it convenient to put large stores of natural language records into computer-readable form, primarily for convenience of access and storage, but also, hopefully, to do retrieval and routine processing automatically.

An example of this process on a small scale in an institutional framework is the case of a hospital which computerizes its patient files for quick back-up to the written charts and for transactional purposes; then, finding itself with this large natural language data base, seeks computer techniques for processing the contents of the documents to obtain the summaries and other information required for health care evaluation. We are bound to see a pressure of this sort arising wherever natural language files are computerized. Once information is available in computer readable form, users inevitably want the computer to process it.

Viewing the future of information services broadly, F.W. Lancaster recently made these projections (11):

> The pattern is more or less inevitable: more data bases in natural language form because "publication" itself will be electronic; more searching of data bases directly by scientists because these files will be readily accessible through terminals in offices and homes; more need for a natural language search approach because the person who is not an information specialist will not want to learn the idiosyncracies of a conventional controlled vocabulary and, even if he were willing to master one controlled vocabulary, the range of data bases that will be readily accessible to him virtually precludes

the conventional controlled vocabulary approach.

> Those concerned with the design of information systems should now be concentrating on functional requirements for the user-oriented, natural language systems of the future.

There are several different ways in which natural language processing techniques may contribute to information systems in the future. For one, natural language might be the preferred medium of communication of the user with the system, as suggested by the above quotation and by others in the information field (12). When the techniques of dialogue analysis and interpretation are further advanced, programs should be able to sort requests according to the type of response which would be appropriate and guide the user to find the desired information in the computer store. Successful experiments with this type of system have been made (13). Such programs could serve as a front end to existing document retrieval systems as well as to systems which include other information services.

Natural language processing techniques could also be important in locating and retrieving specific information, i.e. in "fact retrieval." In testing whether particular documents, or parts of documents, in the data base contained the information requested, advanced techniques such as formatting might be applied to both the document and the request. This would test whether there was a match of fact pattern, as opposed to just an overlap in vocabulary. To obtain the proper passages to test in this way, the technique of answer-passage retrieval (14) might be appropriate.

In restricted subject areas, an extension of some of the more sophisticated programs being developed could perform some of the data processing tasks now performed by research and clerical assistants, e.g. sorting, filing, retrieving, screening and summarizing information in natural language from various source documents, according to given categories. These functions are, of course, a tall order for computers. However, the basis for such programs is being laid in the work described in this paper and in (15). This research, in conjunction with other basic studies in information science such as those reported in this symposium, are looking ahead to a future information technology which combines advances in computer capabilities with knowledge about the nature of information, to solve the problem of the information explosion. A new information technology could turn what is now a burden--too much accumulated information--into a resource, by providing scientists and technologists with direct access,

via user-oriented computer systems, to the large and increasing stores of knowledge.

ACKNOWLEDGEMENTS

The research reported in this paper was supported in part by the National Science Foundation under grant no. SIS 75-22945 of the Division of Science Information, and in part by research grant 1-R01-LM-02616 from the National Library of Medicine, National Institutes of Health, DHEW.

REFERENCES

1. Hirschman, L., and R. Grishman, Fact Retrieval from Natural Language Medical Records. Submitted for publication in Medinfo 1977.
2. Sager, N., A Computer String Grammar of English. String Program Reports (S.P.R.) No. 4, Linguistic String Project, New York University, 1968.
3. Raze, C., The FAP Program for String Decomposition of Scientific Texts. S.P.R. No. 2, Linguistic String Project, New York University, 1967.
4. Grishman, R., The Implementation of the String Parser of English. In Natural Language Processing, R. Rustin, ed., Algorithmics Press, New York, 1973.
5. Grishman, R., N. Sager, C. Raze, and B. Bookchin, The Linguistic String Parser. Proceedings of the 1973 Computer Conference, 427-434, AFIPS Press, Montvale, N.J. 1973.
6. Sager, N. and R. Grishman, The Restriction Language for Computer Grammars of Natural Language. Communications of the ACM, vol. 18, 390-400, 1975.
7. Hobbs, J. and R. Grishman, The Automatic Transformational Analysis of English Sentences: An Implementation. International Journal of Computer Mathematics, in press.
8. Raze, C., The Parsing and Transformational Expansion of Coordinate Conjunction Strings. S.P.R. No. 11, Linguistic String Project, New York University, 1976.
9. Hirschman, L., R. Grishman and N. Sager, Grammatically-based Automatic Word Class Formation. Information Processing and Management, vol. 11, 39-57, 1975.
10. Hirschman, L., Grishman, R., Sager, N., From Text to Structured Information: Automatic Processing of Medical Reports, Proceedings of the 1976 National Computer Conference, AFIPS Press, Montvale, N.J., 1976.
11. Lancaster, F.W. The Relevance of Linguistics to Information, Proceedings of the 1976 FID/LD Workshop on Linguistics and Information Science, in press.
12. Panel: Can Present Methods for Library and Information Retrieval Service Survive?, Proceedings of the 1971

Annual Conference of the ACM, 564-567.

13. Hillman, Donald J., Customized User Services Via Interactions with LEADERMART, Information Storage and Retrieval 9, 587-596, 1973.
14. O'Connor, John, Retrieval of Answer-Sentences and Answer Figures from Papers by Text-Searching, Information Processing and Management, 11, 155-164, 1975.
15. Sager, N., Evaluation of Automated Natural Language Processing in the Further Development of Science Information Retrieval, Final Report to the Division of Science Information of the National Science Foundation; S.P.R. No. 10, Linguistic String Project, New York University, 1976.