

Courant Computer Science Report #7

August 1975

Directions in Artificial Intelligence: Natural Language Processing

Ralph Grishman, Editor

Courant Institute of
Mathematical Sciences

Computer Science Department



New York University

Report No. NSO-7 prepared under Contract
No. N00014-67A-0467-0032
with the Office of Naval Research

COURANT COMPUTER SCIENCE PUBLICATIONS

Price

COURANT COMPUTER SCIENCE NOTES

- Programming Languages and Their Compilers,
J. Cocke & J. T. Schwartz, 2nd Revised Version,
April 1970, iii+767 pp. \$19.25
- On Programming: An Interim Report on the SETL Project.
Part I: Generalities.
Part II: The SETL Language and Examples of Its Use.
(Parts I and II are consolidated in this volume.)
J. T. Schwartz, Revised *June 1975, xii+675 pp.* \$17.25
- A SETLB Primer. H. Mullish & M. Goldstein, 1973, *v+201 pp.* 5.25
- Combinatorial Algorithms, E. G. Whitehead, Jr., 1973, *vi+104p.* 2.75

COURANT COMPUTER SCIENCE REPORTS

- No. 1 ASL: A Proposed Variant of SETL
Henry Warren, Jr., 1973, 326 pp.
- No. 2 A Metalanguage for Expressing Grammatical Restrictions
in Nodal Spans Parsing of Natural Language,
Jerry R. Hobbs, 1974, 266 pp.
- No. 3 Type Determination for Very High Level Languages
Aaron M. Tenenbaum, 1974, 171 pp.
- No. 4 A Comprehensive Survey of Parsing Algorithms for
Programming Languages, Phillip Owens, *Forthcoming.*
- No. 5 Investigations in the Theory of Descriptive Complexity,
William L. Gewirtz, 1974, 60 pp.
- No. 6 Operating System Specification Using Very High Level
Dictions, Peter Markstein, 1975, 152 pp.
- No. 7 Directions in Artificial Intelligence: Natural Language
Processing, Ed. Ralph Grishman, 1975, 107 pp.
- No. 8 A Survey of Syntactic Analysis Procedures for Natural
Language, Ralph Grishman, 1975, 94 pp.

A catalog of SETL Newsletters and other SETL-related material is also available. Courant Computer Science Reports are available upon request. Prepayment is required for Courant Computer Science Notes. Please address all communications to

COURANT INSTITUTE OF MATHEMATICAL SCIENCES
251 Mercer Street
New York, N. Y. 10012

COMPUTERIZED DISCOVERY OF SEMANTIC WORD CLASSES *
IN SCIENTIFIC FIELDS

Naomi Sager

Linguistic String Project, New York University

Abstract

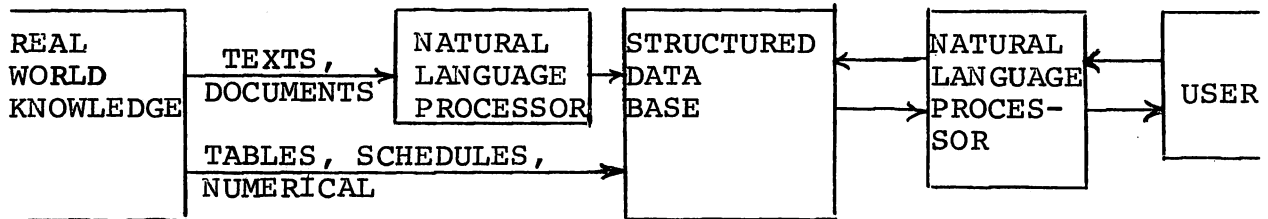
A procedure is described for automatically obtaining the semantic classes in a science subfield. This procedure is based on statistical coOccurrence data for words in particular relations in the text. The results of applying this procedure to a subfield of pharmacology are presented. Its use for structuring the information in natural language texts is discussed.

In placing our work in the context of AI research, it is helpful to distinguish two functions of language, language as a live medium of communication between human beings, or between a human being and a machine, and language as the major means of storing and transmitting the world's knowledge. Roughly this is the distinction between the spoken and the written word. Much of the research in artificial intelligence has been concerned with the former, either in the form of robotics or in question-answering systems, in which the natural language processor acts as an interface between the human user and a structured data base. The data base has not been in question; it has for the most part been entered into the system in tabular form. The artificial intelligence task has been to interpret and act upon the natural language input produced by the user. We have been concerned rather with language as a storage medium, with the problem of structuring a data base which is given in natural language.

These two different foci of research activity may be illustrated by the diagram in Figure 1. Most AI work in natural

* Research supported in part by NSF-OSIS Grant GN-39879

Figure 1



language processing has been located at the user end, whereas our activity has been located at the data or document source end. Our aim is to bring natural language data bases into such a form as to make them amenable to processing in user-directed systems for question-answering and data processing.

As an example, currently there is a great need in the medical community for computer programs that could process the information in patient records for evaluation of health care and for clinical research. This information is recorded in natural language, in dictated reports of clinic visits, in hospital discharge summaries, and the like. Although the contents of these records are limited and repetitive, their form is necessarily in natural language (except for certain parts, such as laboratory findings); e.g., no hospital administrator would dare to request physicians to substitute a multiple choice format for a dictated discharge summary. The question is thus posed as to whether there are methods for obtaining the equivalent of a predetermined format from the natural language material, in cases where it is either impossible or unacceptable to impose an a priori structuring.

Another area of application is the scientific literature. In this area, again, while there are review articles every few years in very small specialized areas of scientific knowledge,

no one has yet written a review article in the form of a table. Nor is it likely that that is going to happen very soon. Yet the exponential growth of the scientific and technological literature and the pressure to obtain and use available results quickly make it desirable to develop computer methods for processing the contents of the literature store, if this is at all possible. There are sufficient restrictions in the language that is used in these areas to suggest that natural language processing methods can be used to extract structures that would give us the equivalent of a structured data base. We have been working in this area, directing our attention to relatively small subject matter areas, though these are relatively large compared with the areas that have so far been treated by computer. We have worked for several years on texts in a subfield of pharmacology having to do with the cellular level action of digitalis, one of the main drugs used in cardiac therapy. The key to the initial success we have had in obtaining a structuring of natural language material is the fact that the pharmacologist in the case of the digitalis texts (and the physician in the case of medical records) is not writing in English seen as a whole language but in effect in a sublanguage of English. The limitations as to what can be sensibly said within the given field show up as regularities in the linguistic records of the material. One can then view the problem of finding appropriate data structures as one of first discovering the grammar of the sublanguage of the given scientific area.

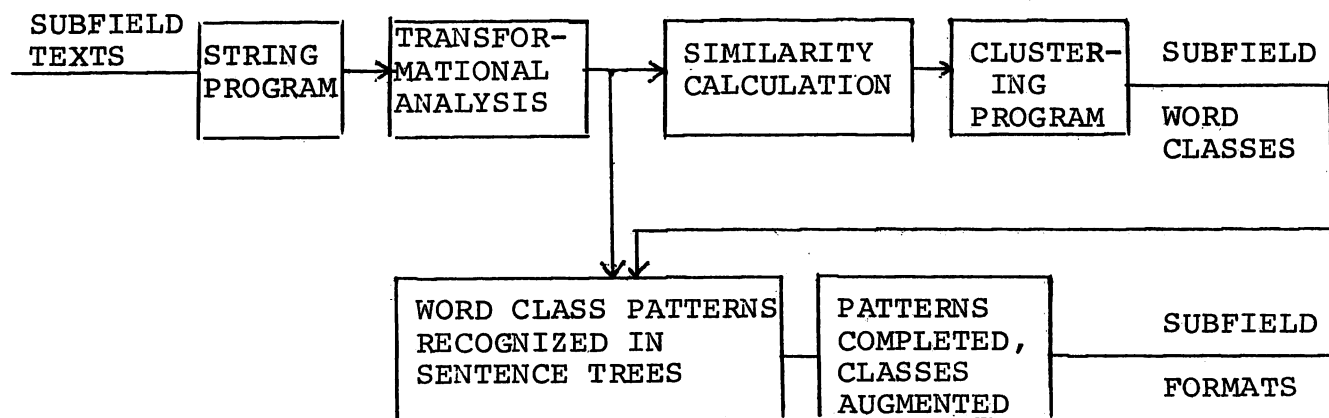
In contrast with the grammar of a whole language we expect the grammar of a sublanguage to have categories which correlate with semantic categories in the material; that instead of obtaining such general classes as noun, or noun singular, noun plural, etc. we would obtain subclasses on a grammatical basis which would represent the objects and relations of interest in the science. In the digitalis work that will be summarized presently it turned out that this is indeed the case. We used linguistic methods to obtain the grammar of the sublanguage,

and the resulting grammatical classes were found to correspond directly to the semantic classes of the field, as recognized by a pharmacologist working in the field.

One of the reasons that this task can be viewed as a grammar-writing problem is the fact that scientists in respect to their fields display linguistic behavior; that is, certain sentences are acceptable and certain sentences are not acceptable within their field, regardless of the truth or falsity of the sentences. It is just a question of whether they are possible sentences within the field. So, for example, "the influx of sodium" or "sodium flows into the cell" is an acceptable sentence in pharmacology. However, "the cell flows into sodium" is just as readily rejected by the pharmacologist as an educated speaker of English would reject an ungrammatical sentence in the English language. It is this kind of linguistic behavior which is captured in the form of a grammar of the science sublanguage.

How is such a grammar discovered? Observe in Figure 2 that on the input side we have subfield texts. There are two stages of syntactic analysis, followed by two steps of statistical analysis, the result being a set of word classes that correspond

Figure 2
Obtaining Information Structures from Subfield Texts



to the semantic classes in the subfield. Then we have two steps of pattern recognition, the last one of which is the human stage of post-editing and putting everything together. The result is what we call a subfield format, a kind of structuring that represents the types of information in successive sentences of the text. We will go over this step by step.

Consider first the two stages of syntactic analysis. These are done by the Linguistic String Parser [1,2,3] which performs a segmentation or a surface analysis of the sentence followed by a second stage of transformational analysis. The type of segmentation obtained using string analysis (illustrated in Figure 3) greatly simplifies the following stage of deep structure analysis. The output parse produces a great deal more grammatical information about every component in the sentence than is seen in Figure 3, but for present purposes the important thing is the segmentation that is obtained. In this simple sentence, "This results from the slowing of the influx of potassium into the cell", "this" appears with an indication that it is a referential to some previous item in the text; "results from" has as its object another segment, whose verb is "slowing" and whose object is "the influx of potassium into the cell". Notice that each segment of this nest of embedded structures contains a verb or verb-like element. In the first line "results from" is a tensed verb. In the second line "slowing" is a verb with an "ing" suffix, making it nounlike.

Figure 3
Parse

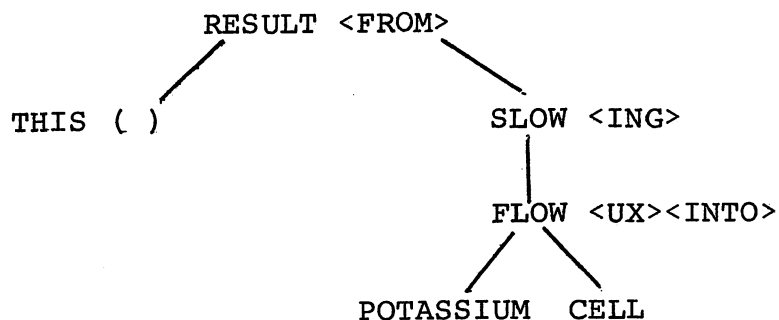
This results from the slowing of the influx of
potassium into the cell.

1.	THIS ()	RESULTS	FROM 2.
2.		THE SLOWING	OF 3.
3.		THE INFLUX	OF POTASSIUM
			INTO THE CELL

And in the third line we have "influx", which is morphologically a noun but is related to "flow into".

This hierarchy of verbs, which is obtained explicitly by transformational analysis, can be represented as a dependency tree, as shown in Figure 4. It turns out that in the parses there is already a semantic separation of what you might call the object language of the science from the meta-language. The lower parts of the parse tree contain the events and the objects of interest in the field, and the upper parts of the parse tree the human relations to these events, in such expressions as "it was assumed that", or "we observed that". Within the lower range there is a hierarchy of verbs. The lowest level here are verbs of qualitative events; then come certain quantitative verbs and then certain causal or sentence connecting verbs. This verb hierarchy can be made explicit by transformational analysis, based on the Harris theory of transformations. In a tree of this type each verb is seen to have certain arguments; at the lowest level the verb has nouns as its arguments, e.g. "flow into" with its arguments "potassium" (subject) and "cell" (object). This is operated upon by "slow", operated on in turn by "results from" which also has another argument of a sentential type.

Figure 4
Transformational Tree



This type of output from the first two stages of syntactic analysis leaves us a tree on which we can do statistical clustering operations to obtain the noun and verb classes. The trick here is to cluster by words that are similar in their grammatical position within the tree, i.e., have the same word as their operator or argument. For example, we read off from the tree in Figure 4 (and several others) certain word triples, as shown in Figure 5. At the top of Figure 5 there are two instances of "move" and one of "flow", all of them having "potassium" as their first argument, or subject. "Flow" and "move" are similar in several respects here. First of all, they have the same word as their first argument; that is, they have a similarity in the transformational trees by having the same word in the same argument position. Nouns also, e.g. "potassium" and "sodium", are found to be similar by the same type of criterion. That is, they have the same verb as their operator while they are in a given argument position.

Figure 5
SIMILARITY COEFFICIENTS

FLOW <UX><INTO>	POTASSIUM	CELL
MOVE <MENT><ACROSS>	POTASSIUM	MEMBRANE
MOVE <MENT><INTO>	POTASSIUM	CELL
MOVE similar to FLOW		
FLOW <UX><INTO>	POTASSIUM	CELL
FLOW <UX><OUT OF>	SODIUM	CELL
POTASSIUM similar to SODIUM		

We are now in a position to compare quantitatively the similarity of occurrences of words. We have seen qualitatively how words in particular grammatical classes are similar in a tree; now we will do it quantitatively. We will compare every pair of words in the entire corpus with respect to six possible grammatical relations. For example, in Figure 6, we see the data for "flow" and "move", taken from Figure 5. "Move" had two occurrences with "potassium" as subject. "Flow" has one such occurrence. The calculation is made on a very large matrix with six N entries for N words [4]. Each word is assigned a characteristic (normalized) vector based on tables of the type illustrated in Figure 6, showing the amount of "similar" occurrences the word has vis-a-vis all other words. We obtain a similarity coefficient, that is, a numerical measure of how similar each two words are in our text corpus, by simply doing a vector multiplication.

Figure 6
COMPUTING WORD SIMILARITIES

WORD PAIR FREQUENCIES

	POTASSIUM AS SUBJECT	POTASSIUM AS OBJECT	CELL AS OBJECT	
FLOW	1	0	1	ALL WORD PAIRS EACH PAIR IN 6 POSSIBLE GRAMMATICAL RELATIONS
MOVE	2	0	1	

NORMALIZED FLOW VECTOR \vec{F}
 NORMALIZED MOVE VECTOR \vec{M} (6n DIMENSIONS)

$$\text{SIMILARITY COEFFICIENT} = \vec{F} \cdot \vec{M}$$

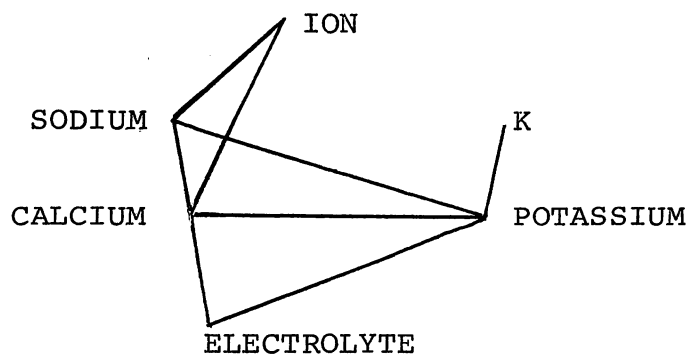
We thus obtain a number that represents the similarity of any two words in the corpus, as illustrated in Figure 7. The slashed squares contain similarity coefficients that are above threshold. The threshold here is .27. Every i, j square contains the calculated similarity between the two words in the i th row and j th column. We expect words that have a similarity coefficient greater than the threshold to form a cluster by a clustering algorithm, as illustrated in Figure 8.

Figure 7
SIMILARITY COEFFICIENTS

	K	POTASSIUM	SODIUM	ELECTROLYTE	ION	CALCIUM
K		<div>.457</div>	.151	.075	.110	.220
POTASSIUM			<div>.365</div>	<div>.348</div>	.147	<div>.514</div>
SODIUM				.148	<div>.449</div>	<div>.488</div>
ELECTROLYTE					.072	<div>.312</div>
ION						<div>.286</div>
CALCIUM						

Threshold = .27

Figure 8
WORDS WITH HIGH SIMILARITY COEFFICIENT FORM A CLUSTER



In Figure 8 we have replaced the slashed squares of Figure 7 by lines, showing that the words that have a similarity coefficient higher than the threshold form a cluster. And as you can see, the cluster of Figure 8 is of ion words but they are not all ion names; the cluster contains ion names plus classifications of ion names. They form a cluster because of their similar type of occurrence in the textual material. The clustering algorithm operates by adding one word at a time to a started cluster. A word is added to a cluster if the average of its similarity coefficient to all of the others in the cluster is above threshold. (Details are given in [4].)

I would like to show you the results of the experiment that we have done with the digitalis material. This experiment involved about 400 sentences from a much larger corpus, taken from 4 or 5 texts, not selected in any particular way, except that they were part of the set of digitalis texts that we had analyzed manually in an earlier study. We will look at some of the clustering output in detail, to demonstrate that the clusters are semantically coherent, so much so that we are able to name them. I should mention that the clustering program generates overlapping clusters and then we have a simple merging algorithm that merges those overlapping clusters that have a $2/3$ intersection. Figures 9 and 10 show the output from the merging algorithm, i.e. the merged clusters from a particular run. CG stands for "cardiac glycoside"; the words that fell into the merged cluster labeled CG are with one exception cardiac glycosides or drug classifiers. There is one interesting exception; the erythrophleum alkaloids are not glycosides, but they are a set of related compounds which in a particular text that we were analyzing happened to be compared with the glycosides and therefore they fell into the CG cluster. The output shows that there is a certain semantic coherence to the class generated by the clustering program. It doesn't mean that the machine is perfect and will make a perfect semantic class, but it does not put unrelated words into one

CLUSTERING OUTPUT

NOUN CLASSES

Run 11.13.74
T = 0.27

CG CLASS

CG	INHIBITOR
DIGITALIS	AGENT
OUABAIN	
DRUG	
STROPHANTHIDIN	
STROPHANTHIDIN 3 BROMOACETATE	
STROPHANTHIN	
CARDIOTONIC GLYCOSIDE	
COMPOUND	
INHIBITOR	
ERYTHROPHLEUM ALKALOID	

CATION CLASS

CALCIUM	ION
CA ⁺⁺	K ⁺
CA	
POTASSIUM	
K	
SODIUM	
NA ⁺	
ION	
ELECTROLYTE	
GLUCOSE	

ENZYME CLASS

NA ⁺ K ⁺ ATPase
ENZYME
ATPase

PROTEIN CLASS

FIBER	PROTEIN
PROTEIN	ACTOMYOSIN
CARDIAC	

SR CLASS

SR
SARCOPLASMIC RETICULUM

MUSCLE CLASS

MUSCLE	MUSCLE
HEART MUSCLE	VENTRICLE

CELL-TISSUE CLASS

CELL
MYOCARDIUM

ADP CLASS

ADP
E1

SPACE CLASS

SPACE
MILIEU

CLUSTERING OUTPUT

VERB CLASSES

RUN 11.13.74
T = 0.27

V_IC (MOVE)

MOVE
TURN OVER
INTRA
EXTRA
CONCENTRATE
FLOW

V_CI (CONTAIN)

CONTAIN
LOSE

V_M (EXCITE)

EXCITE
DEPOLARIZE

V_Q (INCREASE)

INCREASE	AUGMENT
DECREASE	INCREASE
CHANGE	

V_SS (CAUSE)

CAUSE	INFLUENCE	DEMONSTRATE	OPPOSE	EFFECT	DISSOCIATE
INDUCE	INHIBIT	SIMILAR	DIVERGE	PRODUCE	RELATE
PRODUCE	STIMULATE	CAUSE	SIMILAR		
LINK	AFFECT	RELATE			
INTERFERE	CONCENTRATE	DUE TO			
ACT	ACT	PRODUCE			
AFFECT	REVERSE	LINK			
RELATE	INDUCE				
TOXIC	TOXIC				
	PENETRATE				

V_S

(UNNAMED)

REPORT	TAKE	TRANSPORT	AUGMENT	REDUCE	DECREASE
OBSERVE	TREAT	EXCHANGE	IMPROVE	INFLUENCE	MEASURE

class, because it is based on distributional similarities of words. It should be noted that the word classes in Figures 9 and 10 are not selected from the data. They show the complete computer output for the run in question.

The verb classes shown in Figure 10 are interesting. Since the nouns have been clustered with respect to the verbs and the verbs with respect to the nouns we have labelled some of the verb classes in terms of their relational status to nouns. For example, V_{IC} are the verbs that related ion words to cell words. Notice that they are verbs of motion, or of arrested motion. "Moves", "turnover", "concentrate", "flow". ("Intra", "extra" are treated as verbs.) The V_{CI} are just the reverse; they relate cell words to ion words, e.g. "contain" in, "the cell contains sodium". Then we have a level of quantitative verbs: "increase", "decrease", "change", "augment", "increase". The parallel columns indicate that the clusters were not merged by the merging algorithm. Since the merging algorithm requires a 2/3 overlap there was no merging of pairs; the last step of putting these clusters together was done by simple alignment. Notice that the computer has distinguished six different kinds of sentence connecting clusters. We don't know as yet what this really represents. When we did a manual analysis of this material we put all of what seemed to be causal verbs, the relational verbs between events, into a single class, e.g. "cause", "induce", "produce", "link", "interfere", "act", ("act" is "act on", and "toxic" is "toxify" in this case), "influence", "inhibit", "stimulate", and so forth. All of these represent relations among events in the physiological world, in particular, among events that are initiated by the drug actions. And as you can see there are nevertheless differences recognized by the computer program, which would be very interesting to explore; the clustering algorithm saw the members of these subgroups as being more closely associated with each other than with the words in the other subgroups.

[Question] Is toxic a verb?

[Answer] No. But in the verb tree, predicates are treated

as verbs. In this case: Digitalis had a toxic effect. It toxified something and "toxic" was treated as a verb.

To be complete, we have to check what was missed by the algorithm as well as what was obtained. We did a control study by analyzing the same material manually, obtaining the semantic classes and formulas by observation and by doing a rough distributional analysis. We checked these classes with a pharmacologist. Then we compared the manually obtained classes with the computer output. It turned out that in each of the given categories the computer classes contained all of the high frequency occurring elements that were in the manual classes. For example, in Figure 11, the manually obtained cardiac glycoside class contains the same words that we got in the computer, plus some others, which have frequencies of occurrence that taper down toward singular occurrence. The computer was able to collect all of the similar words which had a high frequency and missed those which were of low frequency. In an application, this situation could be corrected by a larger corpus or augmenting the computer generated class (treated as a nucleus) manually.

A similar result is seen in Figure 12 with the cation class. The computer obtained 96 percent of the cation occurrences which were in the manual class, but missed a few of the others which could have fitted naturally into the class.

I would like to refer back to the diagram in Figure 2 for just a moment. We have obtained so far the two steps of syntactic analysis, and the two steps of statistical analysis which gave us the word classes. Now we want to go back into the sentence and obtain, with these classes, the patterns of information structures that we were after to begin with. It should be clear from the way the trees were built and the way the words were clustered from the trees that once we have the word classes and we plug them back into the sentence in the positions from which they were clustered we will get patterns of word class occurrence. By so doing, we will get either fragmentary or complete patterns

Figure 11

CG CLASS

<u>COMPUTER</u>	<u>MANUAL</u>	<u>NO. OF TEXTUAL OCCURRENCES</u>
CG	CG	156
DIGITALIS	DIGITALIS	118
OUABAIN	OUABAIN	70
DRUG	DRUG	15
STROPHANTHIDIN	STROPHANTHIDIN	5
STROPHANTHIN	STROPHANTHIN	4
STROPHANTHIDIN 3 BROMOACETATE	STROPHANTHIDIN 3 BROMOACETATE	4
CARDIOTONIC GLYCOSIDE	CARDIOTONIC GLYCOSIDE	3
COMPOUND	COMPOUND	7
INHIBITOR	INHIBITOR	5
	GLYCOSIDE	11
	AGENT	8
	DIGOXIN	7
	ACETYL STROPHANTHIDIN	7
	CARDIOACTIVE GLYCOSIDE	6
	DIGITALIS GLYCOSIDE	6
	DIGITOXIGENIN	3
	STROPHANTHOSIDE	2
	CARDIAC GLYCOSIDE	2
	DIGITOXIN	1
	STROPHANTHIN K	1
	DIGITALIS COMPOUND	1

88%

ERYTHROPHLEUM ALKALOID 6

442

RUN 11.13.74 T = 0.27

Figure 12

CATION CLASS		NO. OF TEXTUAL OCCURRENCES	
<u>COMPUTER</u>	<u>MANUAL</u>		
CALCIUM	CALCIUM	101	96%
POTASSIUM	POTASSIUM	90	
SODIUM	SODIUM	53	
CA ⁺⁺	CA ⁺⁺	48	
CA	CA	30	
K	K	29	
ELECTROLYTE	ELECTROLYTE	17	
ION	ION	15	
NA ⁺	NA ⁺	11	
GLUCOSE		7	
	K ⁺	6	
	CATION	3	
	MG ⁺⁺	3	
	MAGNESIUM	3	
	NA	3	
		<hr/>	
		412	

RUN 11.13.74 T = 0.27

depending on how much is present in each sentence. At this point some human intellectual work is necessary to put together fragments into more complete structures. We obtain formats which seem to account quite regularly for the repeated content in many text sentences, and as such have been called "information formats" [5,6].

In Figure 13, we show the information format for the sentence that we began with (in Figure 3). The word 'this' takes up a whole unary line. The conjunction here is the sentence connecting verb, "results from", leading to the next line, which contains 2 levels of embedding: the innermost part is an elementary sentence, "Potassium flows into the cell," operated upon by the verb "slows". This is really just the same sentence that was seen before, but it is a case now of a type of pattern which we recognize in very many sentences of the text. The format for this rather complicated field of the cellular action of digitalis has a very repetitive structure when seen in terms of the types of information contained in factual portions of text sentences.

Figure 13
Format of S_5

		S_E					CONJ
V_Q		N_1	V	P	N_2	D_s	
5.1	←-----{4,}-----→	THIS					RESULTS FROM
5.2	SLOWS	POTASSIUM	INFLUX	INTO THE CELL			

SENTENCE: THIS RESULTS FROM THE SLOWING OF THE INFLUX OF POTASSIUM
INTO THE CELL.

Information formats do not characterize the argument from sentence to sentence within a given text, In these procedures we are trying to find a typology for the kinds of information, the major classes of interest, and the types of relations that are of interest within a field. In this case (Figure 14) we find that the repeating pattern is one where there is some elementary qualitative sentence of either cell physiology or biochemistry at the deepest embedded level, a sentence which has a concrete verb, and concrete nouns as its subject and object. These elementary sentences are operated upon optionally by some quantitative material; and this in turn has some sequence of causal verbs operating on it, appearing one after the other with an initiating drug agent as the final subject. We tested this format by going through a number of articles and seeing if they could be formatted according to this pattern, and indeed the large majority of the sentences had this kind of pattern, with different detailed sentences within it, but all of them cases of the word classes in particular relations.

A few last words about the implementation of the components in the above procedures. The linguistic string parser has been operative in one programmed version or another since 1966 [see 1,2,3 cited above]. I would just mention those features of it

Figure 14

PARTIAL TEXT FORMAT FOR PHARMACOLOGICAL SUBFIELD

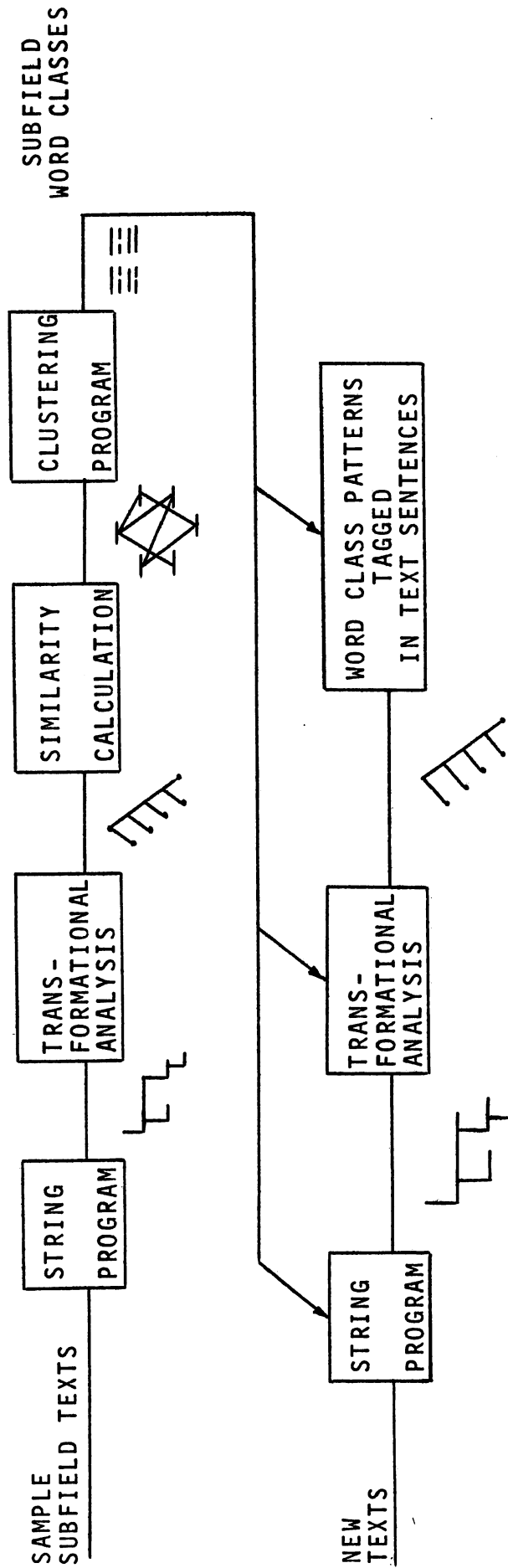
N_0 Q	$V_{ss} \dots V_{ss}$	V_Q	N_1 Q V Q (P) N_2 $P N_3$	D_s	CONJ
:					:
N_0 Q	$V_{ss} \dots V_{ss}$	V_Q	N_1 Q V Q (P) N_2 $P N_3$	D_s	.
DRUG AGENT	DRUG ACTION	QUANTI- TATIVE CHANGE	ELEMENTARY SENTENCE OF CELL PHYSIOLOGY, BIO- CHEMISTRY, ...	EXPERIMENTAL AND CLINICAL CONDITIONS	SENTENCE CONNEC- TIVE

which make possible its application to semantic problems in text analysis. First, it has a very large and detailed grammar of English, absolutely necessary for any real text analysis. Secondly it has an extensive treatment of coordinate conjunctions and comparatives, also without which you cannot do very much real text analysis. Thirdly it now utilizes a programming language which has been developed here for writing the necessary grammatical and semantic restrictions, and which also makes the stating of inverse transformations quite straightforward [7]. The implementation of the current parser, the programming language, and the transformational component have been done by Ralph Grishman. The transformations are being slowly entered into the system by Jerry Hobbs [8]. The treatment of coordinate conjunctions is the work of Carol Raze. The similarity calculation and clustering programs [ref. 4 cited above] were also implemented by Ralph Grishman, in cooperation with Lynette Hirschman, who also did the linguistic analysis.

In the study that I just described to you, the transformational component was not rich enough to handle the text sentences yet, so the trees for that stage were manually done. The trees were drawn by hand for the input sentences, using only such transformations as we are now implementing in the system.

Finally, how can these results be used? What I have been describing so far is a discovery procedure for data structures rather than an application of such structures in an actual fact retrieval question-answering situation. Nevertheless, we can sketch how some of the components would be used in an application, illustrated in Figure 15. Here we would be using the components shown in the top line to obtain the appropriate data structures, and using those shown in the bottom line to process new texts. The arrows into the string and transformation programs indicate a certain amount of semantic feedback, using subfield word classes to resolve syntactic ambiguity. Some of the problems that are difficult in parsing are solved when you have a sublanguage grammar. Once you have the detailed word classes you not only have the basis for stating information structures

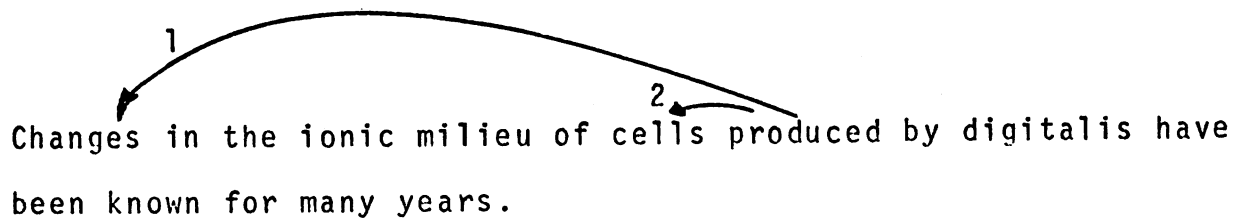
Figure 15
TEXT PROCESSING FOR RETRIEVAL



but you also have a semantic tool for resolving ambiguities that arise in parsing texts, as illustrated in Figure 16. In interpreting the sentence in Figure 16, it is clear here that changes in the ionic milieu of cells is what is produced by digitalis, rather than that the cells are produced by digitalis. But this ambiguity cannot be resolved without semantic constraints that are really quite sublanguage-specific (and are not universal semantic constraints in the language). In this case, in the digitalis sublanguage the verb "produced" never occurs in the sublanguage with "cells" as its argument. Therefore, even if we didn't know that "digitalis" was the subject of "produces" in this sentence we would know that there is no such format fragment as "produce cells". The second interpretation is far less likely than the first because it does not fit any of the formats that have been established for the sublanguage.

Figure 16

SYNTACTIC AMBIGUITIES RESOLVED BY
SUBLANGUAGE CONSTRAINTS



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

- | | |
|-------------------------------|-----|
| 1. Digitalis produces changes | YES |
| 2. Digitalis produces cells | NO |

References

- [1] Sager, N., The String Parser for Scientific Literature. In Natural Language Processing, R. Rustin, ed., Algorithmics Press, New York, 1973.
- [2] Grishman, R., The Implementation of the String Parser of English. In Natural Language Processing, R. Rustin, ed., Algorithmics Press, New York, 1973.
- [3] Grishman, R., Sager, N., Raze, C., and Bookchin, B., The Linguistic String Parser. Proceedings of the 1973 National Computer Conference, AFIPS Press, 1973, pp. 427-434.
- [4] Hirschman, L., Grishman, R., and Sager, N., Grammatically-based Automatic Word Class Formation. Information Storage and Retrieval, in press.
- [5] Sager, N., Syntactic Formatting of Scientific Information. Proceedings of the 1972 Fall Joint Computer Conference, AFIPS Conference Proceedings, Vol. 41, AFIPS Press, Montvale, New Jersey, 1972, pp. 791-800.
- [6] Sager, N., The Sublanguage Technique in Science Information Processing. Journal of the American Society for Information Science, Vol. 26, 1975, pp. 10-16.
- [7] Sager, N., and Grishman, Ralph, The Restriction Language for Computer Grammars of Natural Language. Communications of the ACM, in press.
- [8] Hobbs, J., and Grishman, R., The Automated Transformational Analysis of English Sentences: An Implementation. In ms.