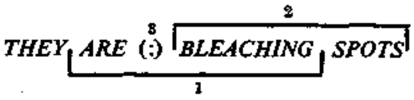
[From: Panel on Semantics and Syntactics,
Information Processing 1962: proceedings of IFIP Congress 62,
Munich 27 August – 1 September 1962 (Amsterdam: North-Holland, 1963)]

1. THE BOUNDARY BETWEEN SYNTAX AND SEMANTICS

A. G. OETTINGER *(USA)*

Semantics may be regarded, only half-facetiously, as that for which syntax alone cannot account. For example, the sentence of fig. 1, due to Knowlton, has three essentially different syntactic interpretations corresponding to four intended "meanings".



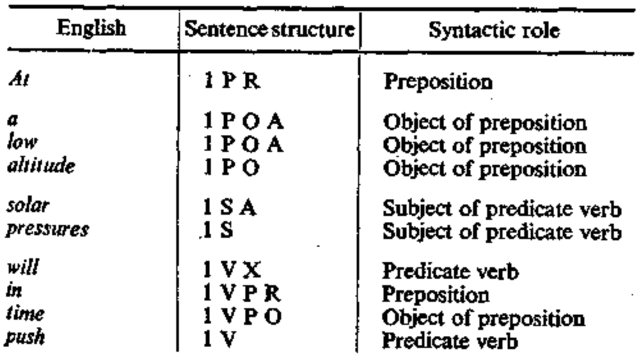
1) Those people are bleaching spots
2a) Those spots are bleaching
2b) Those spots result from bleaching
3) The facts are (:) bleaching spots

Fig. 1.

This example, while contrived, is characteristic of the situation obtaining when a syntactic analyzer, such as that described by Kuno & Oettinger elsewhere in these Proceedings, yields more than one syntactically plausible structure for a sentence.

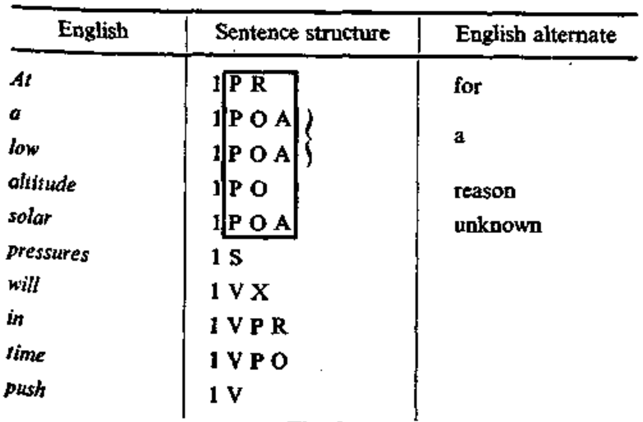
The alternatives shown in figs. 2, 3, 4 and 5 respectively, indicate only a few of the structural variants (detected by the program) within a single sentence from an ordinary technical text.

It is not known at present how to formalize the process of selecting among such alternatives. This seems to be a typical semantic problem, although when finally solved, it may be by reduction to essentially syntactic processes.



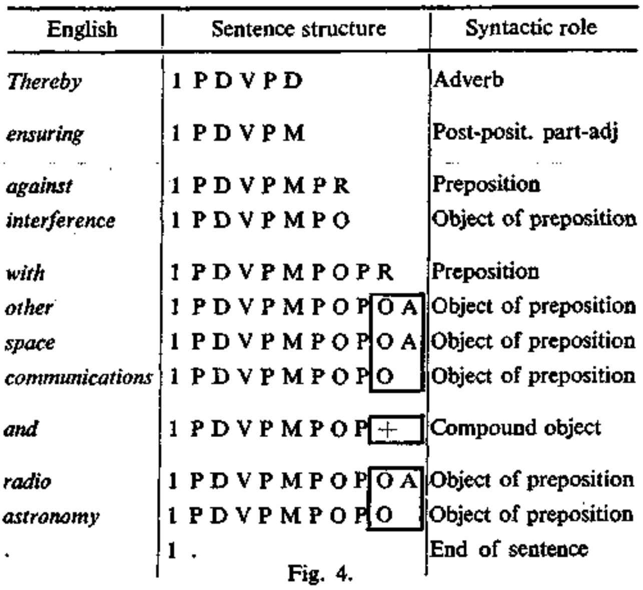
English	Sentence structure	Syntactic role
At	1 P R	Preposition
a	1 P O A	Object of preposition
low	1 P O A	Object of preposition
altitude	1 P O	Object of preposition
solar	1 S A	Subject of predicate verb
pressures	1 S	Subject of predicate verb
will	1 V X	Predicate verb
in	1 V P R	Preposition
time	1 V P O	Object of preposition
push	1 V	Predicate verb

Fig. 2.



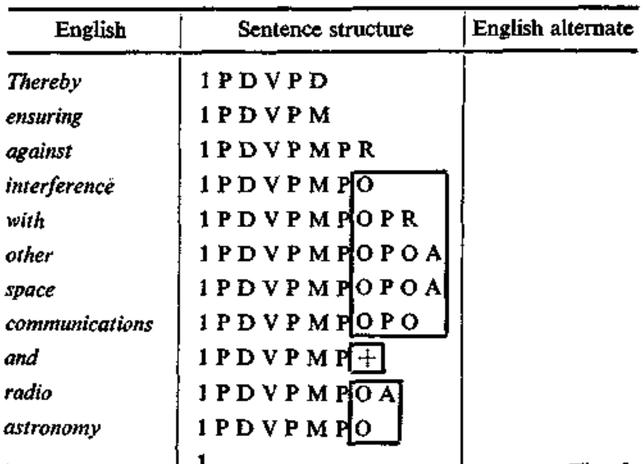
English	Sentence structure	English alternate
At	1 P R	for
a	1 P O A	a
low	1 P O A	
altitude	1 P O	reason
solar	1 P O A	unknown
pressures	1 S	
will	1 V X	
in	1 V P R	
time	1 V P O	
push	1 V	

Fig. 3.



English	Sentence structure	Syntactic role
Thereby	1 P D V P D	Adverb
ensuring	1 P D V P M	Post-posit. part-adj
against	1 P D V P M P R	Preposition
interference	1 P D V P M P O	Object of preposition
with	1 P D V P M P O P R	Preposition
other	1 P D V P M P O P O A	Object of preposition
space	1 P D V P M P O P O A	Object of preposition
communications	1 P D V P M P O P O	Object of preposition
and	1 P D V P M P O P +	Compound object
radio	1 P D V P M P O P O A	Object of preposition
astronomy	1 P D V P M P O P O	Object of preposition
.	1 .	End of sentence

Fig. 4.



English	Sentence structure	English alternate
Thereby	1 P D V P D	
ensuring	1 P D V P M	
against	1 P D V P M P R	
interference	1 P D V P M P O	
with	1 P D V P M P O P R	
other	1 P D V P M P O P O A	
space	1 P D V P M P O P O A	
communications	1 P D V P M P O P O	
and	1 P D V P M P +	
radio	1 P D V P M P O A	
astronomy	1 P D V P M P O	
.	1.	

Fig. 5.

2. A PROCEDURE FOR SYNTACTIC ANALYSIS

NAOMI SAGER *(USA)*

2.1 INTRODUCTION

For purposes of the present discussion, we summarize a procedure for syntactic analysis, previously reported in detail[1]), whose results incidentally bear on certain questions of ambiguity and meaning. The data for a given application of the procedure are sentences of a particular language. Given any set of words of the language, the procedure assigns each word to a word *category,* or a disjunction of different categories, which may be called a *representation of the word.* Each sentence, being a sequence of words, is representable by a sequence of categories corresponding to the successive words of the sentence. We call such a sequence a *sentence representation;* in the case where one or more words of the sentence are represented by a disjunction of categories we obtain a disjunction of sentence representations. The assignment of words to categories, which in broad outline accords with that of ordinary grammar (to nouns, verbs, etc.), is made in such a way that center and other strings (described below) can be defined as sequences of these categories; these strings are defined in such a way that the center strings represent the sentences of the language. Since the procedure uses formulations from the string theory of language[2]), a brief description of the theory will be helpful.

2.2 STRING THEORY

The string theory of language structure sets up, for each language separately, certain *elementary strings,* each string being a sequence of word categories. A string *X* will be said to be *inserted* in a string *Y* if it is adjoined to the right or left of string *Y* or of a symbol in *Y,* or replaces a symbol in *Y.* The strings are grouped into *types*, all strings of one type having the same insertion characteristic. The *rule of combination* for strings is that if a string of type *X* is inserted into a string of type *Y,* in accordance with the insertion characteristic for strings of type *X,* then the result is a string of type *Y.* The strings which are not inserted in other strings are called *center strings.* String theory then states that, for each sentence of the language, at least one representation of the sentence satisfies the conditions for being a center string.

2.3 STRING ANALYSIS

Given the string theory of a particular language (the elementary strings grouped into types and the detailed rules of combination for these types), we decompose a given sentence into its elementary strings; we display the elementary center string (for each decomposition there is only one), and show how each of the other elementary strings is positioned in accordance with the insertion characteristic of a type to which it belongs.

First, we observe that a given sentence must have a representation as a center string. Thus it must be possible to find an elementary center string, or to build up a derived center string, which is identical with one of the representations of the given sentence. Secondly, every word in the sentence corresponds to some category in this center string: the n^th^ sentence word corresponds either to the k^th^ category of an elementary string whose first k-1 categories have already occurred in this center string, or to the first category of an adjunction or replacement string permitted at the n^th^ position in the center string.

Given a sentence to analyse, the procedure then derives a suitable center string by generating at each successive position, from left to right, a list of all the grammatical extensions (over the range of one category) , of the center string as derived to date, and comparing this list with the representation of the corresponding sentence word. The result at the n^th^ position is successful when a category on the generated list is identical with (matches) a category assignment of the n^th^ sentence word. This identification simultaneously selects the category for the n^th^ position of the center string and associates the n^th^ sentence word with a category in an elementary string.

The starting condition for an application of the procedure is the requirement for a center string. Thereafter, when an initial category (head) of a string is matched, the procedure records a requirement for the matching of the remaining categories of the string (in order); these requirements, respecting the rule of combination, are nested. Thus, starting with the first position (*n*=l), the list generated at the n^th^ position consists of the current required category *X* (obtained from the nest of strings) and the heads of strings permitted at the n^th^ position in the derived center string. An application of the procedure is successful if it has obtained a match for each successive sentence word, and has satisfied all outstanding requirements when it reaches the end of the sentence. No match at some position means the particular analysis (attempted derivation) fails.

2.4 RESULTS

Table 1 shows the results obtained for a sample sentence.

First, we notice that the procedure decides to which category (tV or V) the word “consider” belongs, since in the string grammar used here, there is no string beginning with V which can adjoin to the right of the word “we”, and since V cannot be the second category of an elementary center string beginning with “we”. Similar situations occur frequently; the choice among a disjunction of categories is made in the normal course of the procedure.

Table 1
“We consider the following examples of ambiguity.”

	N	tV/V	T2	Ving/N	N	P	N
	N,	NN,…	T2	N,…			
1.	N	tV_N	{(T)	(Ving)	N	(P	N)}
2.	N	tV_{NN}	{[T2	Ving]	N	(P	N)}
3.	N	tV_{NN}	{(T)	N	N	(P	N)}
4.	N	tV_N	{(T)	(Ving)	N	(P	N)}

Second, we see that the procedure has constructed four center strings for the example sentence: in (1) and (4) the object of "consider" is N; in (2) and (3) it is NN; in (3), "following" is a N as in "his assembled following"; in (4) "following examples" is a composed noun, like "following devices" or "washing machines". Each different successful application of the procedure to a given sentence provides a different assignment of the words of the sentence to the categories of elementary strings — a different grammatical relation among the words. In all such cases the sentence is grammatically ambiguous, i.e. it can be understood in more than one (grammatical) way. Conversely, in all cases in which a sentence is grammatically ambiguous, there is more than one way of assigning its words to the categories of elementary strings.

Grammatical ambiguity differs from dictionary ambiguity; in the latter, the different meanings of a sentence are produced by different meanings of a word or words rather than by different grammatical assignments of the words. In grammatical ambiguity, there is a definite number of possible readings; the sentence can be understood explicitly as one or another reading; there are no intermediate possibilities, and no other readings can be inserted that would make sense to a speaker of the language. In dictionary ambiguity there is no such sharp distinction: one can often combine the meanings of the sentence by combining the meanings of the word(s) in question; one can often find intermediate meanings; one can invent additional meanings which would be acceptable as extensions, nonce forms, jokes, etc.

We have seen that string analysis (or an equivalent grammatical analysis) can characterize certain types of meaning difference; in addition it can characterize grammatical meaning in contrast to dictionary meaning.

2.5 REFERENCES

1) Sager, N.: *Procedure for Left-to-Right Recognition of Sentence Structure.* National Science Foundation, Transformations and Discourse Analysis Papers, No. 27. University of Pennsylvania. 1960.

2) Harris, Z. S.: *Computable Syntactic Analysis. Ibid.* No. 15. 1959. Revised in: String Analysis of Sentence Structure. Papers on Formal Linguistics, No. 1. (Mouton & Co., The Hague. 1962).

4. STATISTICAL SEMANTICS

L: B. DOYLE (*USA)*

I have twisted our theme slightly, from "Semantics and Syntactics" to "Semantics and Statistics." The parallel between these themes needs to be brought out. When a man reads a book, he knows word meanings primarily by his vocabulary, but he, as well as a machine, has to cope with problems of multiple meaning. He can do this, partly because of his familiarity with word-grouping habits in his language; we attempt to imitate his ability on our machines by means of *syntactic* analysis. He can also do it because he knows the topic; there is much evidence at hand today that this latter ability can be imitated on machines by means of *statistical* analysis.

Machine translation workers such as Oswald at UCLA have noted that the problem of multiple meaning becomes less when one restricts oneself to what can be handled with a micro-glossary. Suppose then, that one has an automatic method of partitioning entire libraries into specialized fields. One is then performing two useful functions automatically: categorizing the library, and alleviating ambiguity by reducing corpora to segments which can be put in correspondence with micro-glossaries.

A co-worker of mine, Borko[1]), used the statistical technique of factor analysis on the text of 600 psychological abstracts, to demonstrate that one can partition libraries into categories even within a single professional discipline. It is therefore easy to believe that various disciplines are statistically separable from each other.

I have made a small-scale investigation of statistical separation of homographs especially for this panel. For working material I used sample libraries of 100 documents each, in fields sufficiently diverse for one not to doubt their statistical separability. In particular, I used libraries of physics, European current events, and information retrieval. If one can separate these fields, will the natural consequence be separation of identical-word sets into homographic subsets? If so, to what extent?

Each document in the three libraries was represented by a list of its 12 most frequent words; to avoid undue labor in preparing the lists, documents were "simulated," their 12-word lists being derived from segments of documents. The 300 lists in the collection contained 3600 word tokens and about 1250 word types. The word types of interest in a study of homograph separation are those which occur in more than one library. In particular, I selected for study all the words which occurred on at least two lists in each of two libraries; there were 26 such words. Selection of these words was done with the help of a vocabulary inventory program written for the IBM 7090 by J. Olney and K. McConologue of the System Development Corporation.

I looked for three grades of homograph separation:

1) *Clean-cut,* in which all tokens in one library have meanings different from all tokens in another.

2) *Partial,* in which tokens of one homograph occur in two of the libraries, whereas one or more other homographs fall entirely within one library.

3) *Doubtful,* in which the number of homographs falling in two libraries is greater than one.

There is also a trivial case in which only one homograph is found for a given word. (This case is trivial only in being of small interest in this study — but in practical applications it is a case that we wish would occur more often.)

For the 26 word types, I found 14 instances of clean-cut separation, five of partial separation, and seven where only one homograph was present. There were no cases of doubtful separation. The clean-cut separations involved the words: “element”, “unit”, “program”, “force”, “frequency,” “pressure”, “power”, “nuclear”, “precision”, “reduction”, “space”, “satellite”, “operation”, and “system”.

A typical example of a clean-cut separation was the word “pressure,” which occurred twice in the current-events library meaning “influence” or “coercion”, while in the physics library it occurred seven times with the meaning “force per unit area”. As a second example, the word “frequency” occurred eleven times in the information retrieval library as “commonness of occurrence” or “number of times,” but occurred ten times as “number of times *per second*” in the physics library.*

The partial separations involved the words: “year,” “second,” “number”, “two,” and “machine”. A typical example was the word “second,” which referred to an interval of time only in the physics library, but which referred to ordinal position in both the physics and the information retrieval libraries. As a second example, the word “year” meant “a 365 day period starting on Jan. 1” in the current-events library, but meant “any interval of 365 days” in both the current-events and physics libraries.

It is hypothesized that as statistically separable libraries come closer together in subject matter there should be a smaller percentage of clean-cut separations and a larger percentage of partial and doubtful separations. To compensate for this, however, we should find a larger percentage of cases in which only one homograph occurs in both libraries.

Also, of course, given any two subjects we should find a decrease in the percentage of clean-cut separations as the libraries become larger. But this trend need not worry us, since in many cases the doubtful separations may be expected to be of a type in which, say, 100 tokens of homograph A and four tokens of homograph B fall in one library, while three A’s and 50 B’s fall in the other. For practical purposes this is an *almost clean-cut* separation.

4.1 REFERENCES

1) Borko, H.: *The Construction of an Empirically Based Mathematically Derived Classification System,* Proc. W. Joint Comp. Conf. San Francisco, California, (1962).

* Some linguists would object to some of the things I include as homographs. However, it is well to include shades of meaning as well as outright differences in an experiment such as this.

5. SYNTACTIC PROBLEMS IN SEMANTIC ANALYSIS *

M. E. SHERRY *(USA)*

5.1 APPROXIMATIONS TO SEMANTIC ASSOCIATION

During the past several years, much work has been reported on the determination of association strengths, or degree of relatedness, between pairs of English words in a document. Most commonly, the frequency of co-occurrence of two words within the document has been used for evaluating the strength of association of the word pair. A more sophisticated approach has considered the degree of textual proximity to determine this measurement. The works of Doyle[1]), Borko [2]), Stiles [3]), Maron and Kuhns [4]), Swanson [5]), and Luhn [6]) have been aimed at this problem.

The use of linear proximity seems to offer better criteria than mere co-occurrence of words. For example, an adjective modifying an immediately-following noun would be expected to be more strongly associated with it than with more distant nouns. These criteria appear to break down, however, as more complex sentences, particularly ones with modifying phrases and clauses, are considered. Take, for instance, the sentence, “The man who was walking with the boy ate the apple.” With linear ordering, “man” and “ate” seem weakly related to each other whereas “boy” and “ate2 seem strongly related.

But a syntactic analysis of this sentence reveals that “man” and not “boy” is strongly related to “ate”. “Man” is the subject and “ate” is the predicate of the sentence, whereas “boy” modifies “walking” which, in turn, modifies “man”. These relationships can be considered linkages so that “boy” is related to “ate” only via “walking” and “man”. By tracing a path through a diagram of such linkages, another measurement of proximity can be made.

This apparent difficulty has led us to study the use of syntactic proximity as opposed to linear proximity in the hope of achieving a better quantitative measure of the association strengths of the word pairs. Syntactic analysis, moreover, provides us with an extra criterion for evaluating association strength. In addition to seemingly more valid measures of linkage, relative weights can be assigned to different types of linkage, so that a subject-predicate relationship can have a different value than an adjective-modified noun relationship.

5.2 PROBLEMS IN SYNTACTIC ANALYSIS

To provide analyses for our experimental work, the Multiple-Path Syntactic Analyzer of Kuno and Oettinger[7]) is used. This program analyses the sentence,

* This work, done under a subcontract from Arthur D. Little, Inc., Cambridge, Mass., was sponsored by the Operations Applications Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force, under Contract AF 19 (628)-256.

producing the set of all possible syntactic analyses. The program is based upon the idea of predictive syntactic analysis first conceived by Rhodes[8]) and adapted by Sherry[9]).

Several fundamental problems have emerged during the course of our work, the most significant concerning the fineness of the grammar used in syntactic analysis. Very briefly, the syntactic analysis program works with a set of grammar rules and an associated set of word classes assigned to the words of a sentence. The program chooses the sets of word classes, one class per sentence word, that make up valid analyses. A coarse grammar-rule set results in analyses with little discrimination between different sentence types, and a narrow choice of values of association strengths. Furthermore, many sentence types cannot be successfully analysed. On the other hand, a very fine set of rules results in the additional successful analysis of less frequent sentence types and many different association linkages which can be assigned a greater variety of values. In the process, however, this fine set generates a multiplicity of acceptable analyses of a sentence, some differing only slightly from others.

If we agree that a reasonably fine grammar is required to permit at least one useful analysis of almost every sentence, provision must then be made for selecting the correct or desired analysis from those deemed acceptable. At this point there arises the question of whether one really needs so fine an analysis for further research as for the syntactic program itself. Indications point toward the necessity for the fine syntactic analysis, with a subsequent analysis to merge the results into more general sentence structures and groupings. This approach has obvious limitations since it does not even promise, let alone guarantee, a unique analysis per sentence.

The last and probably the most important question still to be answered is whether syntactic proximity really provides a significant improvement over linear proximity. As always, examples can be dreamed up to show the merits of a more sophisticated scheme. But whether or not the improvement is statistically significant is another matter. We believe the answer will not be obvious even after processing several texts, and most likely will itself involve considerable thought and experimentation.

5.3 REFERENCES

[1]) Doyle, L. B.: *Semantic Road Maps for Literature Searchers.* J. Ass. Comp. Mach. **8** (1961) 553-578.

[2]) Borko, H.: *The Construction of an Empirically Based, Mathematically Derived Classification System.* Proc. W. Joint Comp Conf. San Francisco (May 1962).

[3]) Stiles, H. E;: *The Association Factor in Information Retrieval.* J. Ass. Comp. Mach. **8** (1961) 271-279.

[4]) Maron, M. E. and J. L. Kuhns: *On Relevance, Probabilistic Indexing, and Information Retrieval, ibid* **7** (1960) 216-244.

[5]) Swanson, D.: *Searching Natural Language Text by Computer.* Science **132** (1960) 1099-1104.

[6]) Luhn, H. P.: *Potentialities of Auto-encoding Scientific Literature.* International Business Machines Corp. (1959).

[7]) Kuno, S. and A. G. Oettinger: *Multiple-Path Syntactic Analyzer.* These Proceedings.

[8]) Rhodes, I.: *A New Approach to the Mechanical Syntactic Analysis of Russian.* National Bureau of Standards Report No. 6295 (1959).

[9]) Sherry, M. E.: *Syntactic Analysis in Automatic Translation.* Doctoral Thesis, Harvard University (1960).