[Proceedings of the *International Conference on Scientific Information*, November 16-21, 1958;
pp.917-935]

The Feasibility of Machine Searching of English Texts

VICTOR H. YNGVE

ABSTRACT Similarities between the literature search problem and mechanical translation raise the hope mat an ultimate goal of directly searching texts written in English may be attainable. The role of a grammar in such a device is discussed. As a first step in reaching the ultimate goal, a very simple language is examined. This language, English Dialect A, has sentences consisting of single English terms. The hierarchical ordering of the terms is expressed in the grammar instead of in the terms. This provides certain advantages when used in a search program.

Literature searching and information retrieval problems in general and the Patent Office problem in particular are of interest because their solution would be of great practical value. They also raise questions of theoretical interest, the solution of which would advance greatly our understanding of human language and perhaps even of human knowledge. These problems concern the nature of the various artificial languages proposed for retrieval purposes and ultimately the nature of English, since it is from a background of English that these artificial languages are devised and since they are used for encoding information originally expressed in English.

The process of encoding information for retrieval purposes has similarities to other language translation processes which are being investigated intensively in the field of mechanical translation (MT). In this field, work is being done with the object of finding out how such natural languages as German, Russian, French, and English can be translated into one another automatically by machine.

The object of this series of reports is to see if some of the insights and techniques being discovered in the field of MT might be applicable to search and retrieval problems and to see if the insights developed by combining the two points of view might shed some light on the more basic linguistic problems.

VICTOR H. YNGVE Massachusetts Institute of Technology, Cambridge, Massachusetts, and supported in part by the National Bureau of Standards.

The ultimate system

It is perhaps appropriate before embarking on a research program that we state an ultimate goal which is high enough to be a serious challenge and which can serve as a guiding principle during the course of the research. For the purposes of this series of reports we will take as our ultimate goal the development of a system in which the documents to be searched and the questions to be asked are expressed in English and all necessary operations are fully automatic. If we reach this goal we can say that the machine literature searching problem has been solved in principle. The word "English" will be used here to represent one of the natural languages. The ultimate solution could also be expressed in terms of languages other than English. The additional question of whether the proposed solution would be economically justified in any given situation can be answered more easily after we know how to search English text.

Recent developments in linguistics and in mechanical translation would lead one to expect that this goal may actually not be too remote to consider. It is the belief of some in the field of MT that it will eventually be possible to design routines for translating mechanically from one language to another without human intervention. Since accurate translating must leave the meaning unchanged but expressed in a different language and a literature searching operation must search for a particular set of meanings, there are many similarities between the two problems. However, since a searching system using English as the basic language has been explicitly rejected as too difficult and visionary if not actually impossible by some and merely overlooked or ignored by others, it might not be amiss to recount some of the obvious advantages of such a system before plunging into the not inconsiderable difficulties standing in the way.

The first advantage of searching English texts directly is that there would be no need for manually encoding the tremendous bulk of the patent literature, to say nothing of the other pertinent literature. It is of course assumed that a print reader can be developed which could serve as an input mechanism operating directly from the printed patent documents. The special problems associated with pictures and diagrams will not be discussed here. A second advantage associated with the elimination of the manual encoding step is the elimination of abstracting. The whole file would be available for search. Abstracting has the disadvantage that it must inevitably leave out some details as being unessential. These details then will not be encoded and can not be retrieved even if they are wanted at some future date. A third advantage is that the question posed by the examiner is already posed in English and he will not have to take the time to express it in a special machine language or wait for someone to do this for him. All communication between the machine and the patent examiner in this ultimate system would be in English.

The difficulties are not the language, but our understanding of it. The difficulties are not that English has no uniform or logical rule for the naming of things, not the ambiguousness of English words, not the wide diversity of phrasing and sentence structure which might be used in the same situation, not the arbitrariness of the conventions of language. English is in fact almost ideally suited to the search task. It is the language in which the patents are written and in which the questions are asked. It is the basis of the present patent classification system. It is used extensively in the present search procedure. Even all the decisions of the examiners as well as of the courts are based on how the patent document reads (in English), and on how the law reads (in English), and how it has been interpreted (in English). For all these purposes English serves us well. There are no difficulties with English, we use it effectively every day. The trouble is that we do not yet understand enough about the rules of our language to be able to instruct a machine to use it.

The role of grammar

A language is a system of symbols and the rules for combining them which can be used for communication. The grammar of a language consists of a list of the symbols and a statement of the rules. Ultimately, a grammar will have to be contained in the machine if the documents and questions are available to it only in English. In MT we are finding out how to put grammars in a machine, but so far in information searching, little has been done along these lines. It may be of interest to see what the grammars of some of the machine languages in current use look like. Many of these languages consist essentially of descriptors. The grammar for a language like this is a list of the descriptors, a very simple grammar indeed, and the languages are very simple but apparently quite effective if properly used.

Most of the machine languages that I have seen proposed for search purposes seem to represent an attempt to find a happy compromise between two conflicting requirements. One of these requirements is that the language be simple enough to be used directly with searching devices which can carry out only simple operations like matching and elementary logical operations, for example, "indicate a match if you find descriptor A and descriptor B, but not descriptor C associated with the same document." The other requirement is that the language be rich enough to express all the information through which it is desired to search, specifically information that is expressed in the document in English. Machine languages have been designed to attempt to reach an ideal compromise—to effect an impedance match between the English of the document and the binary decisions of the machine. One of the ingenious devices used is the method of showing explicitly in the code the inclusion relations between the terms. For example, the code for animal could be contained in the code for mammal and this in turn contained in the code for horse. This device is also used in the UDC. It has the advantage of reducing the grammar to a simple list and the search procedure to a simple match of a whole code word or part of one.

But the fact that a horse is a mammal and that a mammal is an animal is after all not a fact about the real world but a relation between symbols in our language. We just happen to have adopted the convention that certain animals are classed together and given a special term. The terminology is convenient but completely arbitrary. It is a part of the language, a fact of English grammar. If this fact is needed for search purposes, it should be stored in the machine as a rule of grammar, not furnished to the search mechanism each time explicitly in the code for horse. The problem is matching information originally expressed in English to the binary search criteria and binary search operations. It is asking too much to require that all possible answers to search questions be carried explicitly in the encoding language when they can more easily be carried implicitly in the language and brought into explicit form when needed by machine manipulation with the aid of a stored grammar.

A step-by-step approach

As important as it is to set a high ultimate goal, it is equally important to find a succession of short term goals, each of which can be quickly reached, each taking us one step nearer to the ultimate goal.

Many things will have to be discovered about English grammar before we will be able to search patents directly. For example, we will have to discover the various mechanisms the language uses to keep the reader informed as to whether the thing under discussion is the same thing that was mentioned before or something new. We will have to discover how a text can be unambiguous to the reader although nearly every individual word used is ambiguous in isolation. We will have to find out what is the connection between the subsumed-included relations familiar to the Patent Office and the linguistic categories familiar to those who have been working on the structural analysis of English. We will have to find formal connections between widely divergent ways of saying essentially the same thing. In addition there is much that we will have to learn about searching. If we had today a complete grammar of English which was capable of rendering explicit all the relations and distinctions implicit in the document, I doubt that we would know how to utilize it effectively in a machine search situation. We would be embarrassed by the very wealth of the information available. Much more must be learned about search situations.

As a first step in our approach to the ultimate goal, I suggest that we work with a very simple grammar which we will call English Dialect A. We will explore this dialect and its relations to the search problem carefully and learn from it what we can. Then we will devise an English Dialect B which will be more like English, and so on. These dialects will be chosen in such a way that we can reach an understanding of the search problem and of the linguistic situation in a relatively short time. At the completion of each step the dialect will be available for mechanization so that at any point a machine may be used to assist us in the research. Pilot or experimental systems can be set up at any point. Ultimately, we will be able to handle English as it is written; practical results applicable to particular search problems may appear from time to time along the way.

English Dialect A

English Dialect A is a language in which each sentence is one word. We will call the words terms, and list them in the grammar of the language. The terms are related only by means of a hierarchical system which is also expressed in the grammar.

The terms of English Dialect A are taken intact from English. They may be single words or phrases in English, but they are treated as single terms in Dialect A. The relation of Dialect A to English is that the meanings of the terms and their hierarchical relationships will be as near as possible to their meanings and relationships in English. Examples of terms in Dialect A are

COLLIE
CHESAPEAKE RETRIEVER
SHETLAND SHEEP DOG
HUNTING DOG
DOG FROM THE KENNELS OF JOHN SMITH

The rules of the grammar express the hierarchical system. They are formalized by writing pairs of terms in relations such as

ANIMAL	=	MAMMAL
MAMMAL	=	DOG
DOG	=	COLLIE

These rules are interpreted to mean that the term on the left represents a genus of which the term on the right is a species. It is clear that rules of this kind can completely specify a hierarchical system. It is also clear that the three rules above imply

ANIMAL = DOG
ANIMAL = COLLIE
MAMMAL = COLLIE

so that it is unnecessary to write such rules in the grammar.

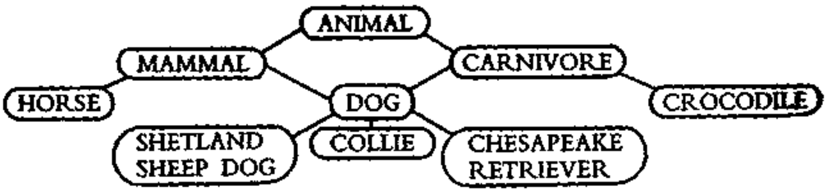
All the relations giving the species of one genus are collected together and called subrules of one rule. Since the left sides of these subrules are all the same. there is no ambiguity in omitting the left side for each subrule except the first one.

DOG = COLLIE
= CHESAPEAKE RETRIEVER
= SHETLAND SHEEP DOG

It is also possible to have the same term subsumed under two or more genera. One may have for example

MAMMAL = DOG
CARNIVORE = DOG

The grammar thus does not exhibit a simple tree structure.



ANIMAL = MAMMAL
= CARNIVORE
MAMMAL = DOG
= HORSE
CARNIVORE = DOG
= CROCODILE
DOG = COLLIE
= CHESAPEAKE RETRIEVER
= SHETLAND SHEEP DOG

A grammar can be interpreted as a computer program for deducing from a given term all the terms subsumed under it. The computer program would discover that subsumed under the term MAMMAL were to be found the terms HORSE, DOG, SHETLAND SHEEP DOG, COLLIE, and CHESAPEAKE RETRIEVER.

Such a program would merely have to search among the terms on the left for MAMMAL, find that under MAMMAL are HORSE and DOG. The program would then search in turn for HORSE and DOG on the left hand side to find what is subsumed under them, and continue the process until the resulting terms could no longer be found on the left.

The grammar can also be interpreted the other way around, as a program for deducing from a given term all the terms under which it is subsumed. For this purpose, it is convenient to rewrite the grammar with the left and right hand sides of the rules reversed and the subrules reordered so that terms again appear only once on the left.

```
SHETLAND SHEEP DOG      =  DOG
CHESAPEAKE RETRIEVER    =  DOG
COLLIE                  =  DOG
DOG                     =  MAMMAL
                        =  CARNIVORE
HORSE                   =  MAMMAL
CROCODILE               =  CARNIVORE
MAMMAL                  =  ANIMAL
CARNIVORE               =  ANIMAL
```

We will call this latter a recognition grammar and the former a construction grammar.

A machine with a program of the above type could be used for literature searching in the following simple way: The documents would be represented by descriptors selected from among the terms of English Dialect A. An effort would be made to use the most specific terms possible for each document. The question would also be a term selected from those of the language, and would generally be a generic term. The program could operate on the descriptors of the documents using a recognition grammar, or it could operate on the question using a construction grammar, or both. In any case, additional terms would be generated for the document, for the question, or for both. The machine would then proceed to test for an exact match between a document term and a question term.

The next question to explore is what to do about ambiguous terms. For example,

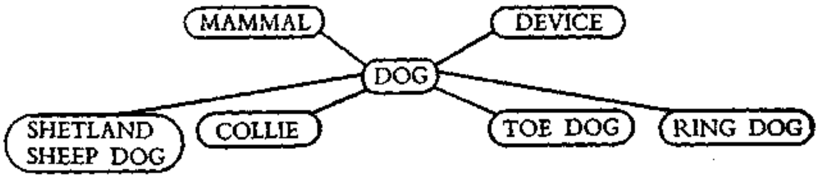
```
MAMMAL    =  DOG
DEVICE    =  DOG
```

This situation is not to be confused with the previous problem

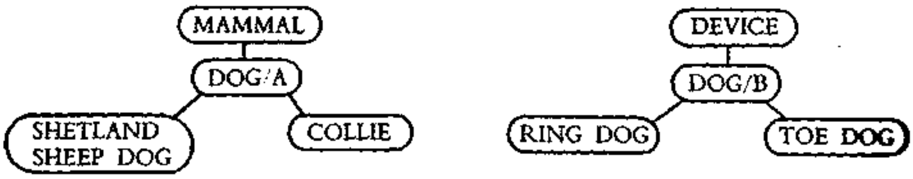
```
MAMMAL     =  DOG
CARNIVORE  =  DOG
```

where DOG is really unambiguous. It has subsumed under it COLLIE, etc., but not TOE DOG, RING DOG, or CHAIN RAFTING DOG, which will have to be subsumed under the DOG that is a DEVICE.

English Dialect A has no way of resolving the ambiguity of ambiguous terms, but English does, by the use of context, and dialects to be developed later will. But English Dialect A has a way of dealing with a problem arising from the use of ambiguous terms, that is, the problem of being able to deduce correctly that a COLLIE is a MAMMAL and that a TOE DOG is a DEVICE even though the language contains the ambiguous term DOG.



There are several ways of dealing with this problem. Perhaps the one best adapted to our purposes is a subscript notation. Two new terms are introduced into the grammar to resolve the ambiguity of DOG as far as the internal workings of the grammar are concerned.



ANIMAL	=	MAMMAL
	=	CARNIVORE
MAMMAL	=	DOG/A
	=	HORSE
DEVICE	=	DOG/B
	=	CLEVIS
CARNIVORE	=	DOG/A
	=	CROCODILE
DOG/A	=	COLLIE
	=	CHESAPEAKE RETRIEVER
	=	SHETLAND SHEEP DOG
DOG/B	=	TOE DOG
	=	RING DOG
	=	CHAIN RAFTING DOG

For the internal workings of the program, DOG/A and DOG/B are different terms, but for the purposes of comparison with descriptors or questions, the subscripts are ignored and DOG is ambiguous. It can only cause trouble now if it is actually used as a descriptor for a document or as a question. The trouble will take the form of selecting extra unwanted documents.

Our approach is not to try to invent unambiguous terms and require that the encoder and the questioner use them. Instead, our effort is to provide for the encoder and questioner a language that is as close to English as possible, so as to improve the match between man and machine. We arrange our grammar so that ambiguous terms, if used, will cause the minimum of trouble. The possibility of their incorporation gives us a language more like English and thus more natural to use. The ambiguity of a term is really not a property of the term, though we speak of it that way. It is the property of the grammar of the language to which the term belongs. A term may be ambiguous according to one grammar, but not according to another. Ultimately, we hope to have the grammar for a dialect of English that will effectively be able to handle the patents in their original language.

English Dialect A will be of interest linguistically if it is elaborated to include many more terms. Some of the formal devices that English uses to indicate class inclusion will become obvious through a comparison of the linguistic forms that English uses for generic and specific terms. English Dialect A may also be of some utility for immediate application to a certain class of information search problems.

ADDENDUM: THE SECOND STEP

Review of Dialect A

English Dialect A was linguistically very elementary. It consisted of sentences that were composed of single terms. The terms were taken intact from English. They were noun phrases consisting of a noun head and one or more modifiers. The terms were arranged in hierarchical fashion by means of rules in a grammar which expressed the inclusion relations. The hierarchical structure was not a simple tree because a given term could be subsumed under more than one other term. Documents were to be encoded into terms of Dialect A and the questions were also to be posed as terms in Dialect A. But instead of searching by a simple match only, the machine would first generate all other relevant questions on the basis of the grammar and search for answers to them all. Alternatively, the machine could supply the documents with extra terms according to the grammar. Search was then to proceed on the basis of an exact match. It was conceived that any term used as a search question would retrieve any document described by that term or any term lower down in the hierarchy.

English Dialect A had important advantages over many of the other methods of encoding for search. The judgment of relevance was done by the machine on the basis of a stored grammar, not on the basis of the document codes alone.

The system could be used to find things from a completely new and different point of view merely by changing the grammar. No reencoding of the file would be necessary for this. The system could be brought up to date easily in the face of changes in the interests of the questioners. Older systems that attempted to incorporate the hierarchy directly in the codes had the disadvantage that the classification system could not be changed without reencoding, but in Dialect A, the classification system could be completely overhauled with no change at all in the encoding of the documents. In addition, the codes of Dialect A could be much more compact than codes carrying hierarchical information explicitly.

English Dialect A also had some serious shortcomings. It would have to contain a very large number of terms to be of much use, there being no facility to combine terms into more complicated expressions. No relations could be expressed between terms. There was also the problem that in Dialect A one always searched for a species, given a genus while, in fact, one sometimes wants to search for the genus given the species.

English Dialect B

English Dialect B bears some resemblance to English Dialect A. It has all of the advantages of Dialect A and some additional ones. Some of the shortcomings of Dialect A are eliminated in Dialect B. The main difference between the two dialects is that Dialect B has sentences consisting of two parts, a modifier and a term. The terms of Dialect B are the same as the terms of Dialect A, that is, they are nouns and expressions with noun heads and certain types of modifiers. By introducing a two part sentence, the number of possible sentences is not limited to the number of terms as it was in Dialect A, but approaches the product of the number of terms and modifiers. It is assumed in Dialect B that each modifier can be used with each term, an assumption that is not entirely warranted in practice.

A preliminary investigation was made of the requirements of a search and retrieval language, and of the kinds of simple dialects that seemed to hold promise of being useful. This investigation suggested that an understanding of certain special pre-noun modifiers would be a good second step in our understanding of how English serves as a retrieval language. Consequently, pre-noun modifiers were examined in some detail. Traditionally, pre-noun modifiers have been divided into two groups: the descriptive adjectives, such as *large, small, red,* and *old,* and the limiting adjectives such as *some, the,* and *these.* Of these two groups of adjectives, the descriptive adjectives have already been incorporated in the terms of Dialect A. It was felt that an investigation of their role in the retrieval situation could best be postponed until more was understood of the limiting adjectives.

It appears that the primary function of limiting adjectives is referential. They serve to refer to something (which is not directly named) in terms of some other thing or category (which is directly named). For example, the phrase *this dog* points out a certain definite object not named, but referred to in terms of a definite category designated by the term *dog.* These modifiers thus seem closely bound with the specification of sets and subsets of named items. Generally, the limiting adjective specifies the set or subset, and the descriptive adjectives and noun (a term in Dialect A) serve to name the set from which the subset has been taken. Of course, many if not all the descriptive adjectives can also serve to specify subsets. For example, *some tart apples* can be conceived of as a subset of *some apples,* but on the other hand, the set *some apples* is not guaranteed to contain such a subset. It seems best, therefore, to regard the descriptive adjectives as purely descriptive in line with the traditional view and treat them together with the noun as terms in a terminological hierarchy. In other words, the term *tart apple* is subsumed under the term *apple,* in the sense that all tart apples are also properly described as apples.

Limiting adjectives can readily be divided into two groups, those that render the noun phrases incorporating them self-contained so that their meanings are clear without reference to the immediate context, and those that require or imply reference to the context. *A, some, many* are in the first group. *This, those, the* are in the second group. Since our new dialect allows only one noun phrase to a sentence, it seems appropriate at this point to limit ourselves to the limiting adjectives that render the noun phrase, that is the sentence in our new dialect, self-contained.

Noun phrases in English can be classified into three mutually exclusive categories, those that are plural, those that are singular, and those that are uncountable. We have the contrast between *apples, an apple, apple.* Singular and plural are familiar enough. The uncountable category includes the so-called mass nouns (water), category names (sulphur), proper names (John), and so on. *Washington* in the sentence: “He lives in Washington.” is a proper name and therefore uncountable, although, of course, it can also be used in the singular and in the plural: “There are several Washingtons in the United States; the Washington that I mean is a state.” Almost all nouns can be used in all three categories. Some nouns undergo a definite meaning change when changing from uncountable to singular or plural: Carbon is an element, but a secretary makes a carbon of a letter. This phenomenon is ignored in English Dialect B. It should be investigated further in the future.

In the above examples, the three categories, plural, singular, and uncountable, are distinguished respectively by the plural *s*, the indefinite article *a* or *an*, and neither article nor plural *s*. We will thus take as three of the limiting adjectives in Dialect B, the following:

Ø—S
A —Ø -
Ø—Ø

where we indicate zero by Ø to distinguish it from the letter O, and write A to stand for both A and AN. When these are combined with the term APPLE, we get

APPLES
AN APPLE
APPLE

We are now ready to specify in detail the structure of English Dialect B. The sentences consist of two parts, a term from English Dialect A and a modifier selected from the following list. To make a sentence, the terms are inserted in place of the X’s in the list.

Ø X Ø
A X Ø
Ø X S
SOME X Ø
SOME X S
ALL X Ø
ALL X S
ANY X Ø
ANY X S
EVERY X Ø
MANY X S
ONE X Ø
THREE X S

as well as all the other numbers.

In order to be able to use Dialect B in a retrieval situation, we must know more of its grammar. Specifically, we must know how questions and answers are related in the language. For a question, we assume the following form:

Does this document show . . . ?

and for encoded information, we assume the following form:

This document shows....

In the place of the three dots, we are at liberty to substitute any appropriate sentence from the dialect. In Dialect A, a document was found if it was described by a term lying lower down in the hierarchy than (subsumed under) the term used as the question. We now ask how this must be modified by the addition of the limiting adjectives. In order to obtain some information on this point, several questionnaires containing sample questions and document codes in Dialect B were circulated to Patent Office personnel. The results of these questionnaires have been carefully analyzed and have been partially incorporated in this dialect. Further investigations of this general nature may serve to modify the dialect in some details, but the overall structure will likely remain unchanged.

The first thing to be noted about the retrieval order is the behavior of expressions in the plural, singular, or uncountable. It turns out that a question in the uncountable should retrieve a document code in the singular and plural as well as in the uncountable. This is understandable because AN APPLE as well as APPLES contain the substance APPLE. This relation seems to be generally true for a large number of nouns. Also, the singular should retrieve the plural as well as the singular, since if one has APPLES, one also has AN APPLE in harmony with a broad interpretation of the meaning of the singular. We thus have in Fig. 1 the hierarchy of these sentences in Dialect B.

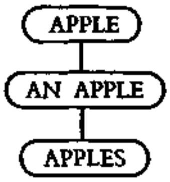


FIGURE 1. Retrieval diagram for plural, singular, and uncountable.

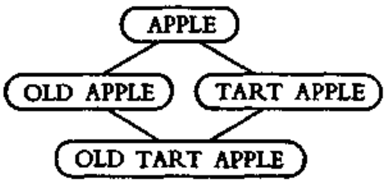


FIGURE 2. Retrieval diagram for some terms from Dialect A.

The next thing to investigate is what happens when there is a sentence with one term in the question and a sentence with another term, related to it hierarchically, in the descriptor. We have in Fig. 2 a hierarchy of terms of the type investigated in Dialect A. When each of these four terms is combined with the three modifiers, we have the result in Fig. 3. The correct diagram for

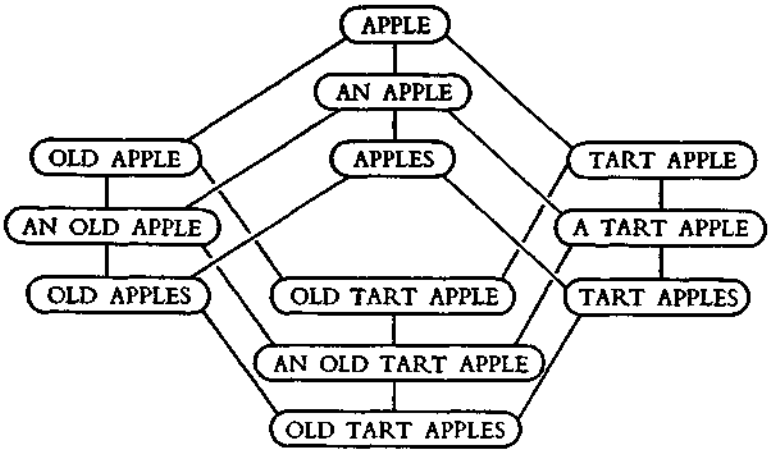


FIGURE 3. Retrieval diagram resulting from the combination of the diagrams c Fig.1 and Fig.2.

retrieval purposes is obtained by combining the first two diagrams in a way that one will recognize as the direct product. In the diagram of Fig. 3, any one of the sentences may be used in a question. When so used, it should retrieve descriptors matching itself or any other sentence lower down in the diagram.

Investigation of the numerals as limiting adjectives reveals that in Patent. Office practice they support at least two distinct meanings. THREE APPLES includes three or more apples, or just three apples. For our purposes, we can add subscripts on the numerals in the grammar as we did in Dialect A, and we can also add two new more precise modifiers and equate them to the subscripted numerals.

THREE/1 ≡ THREE OR MORE
THREE/2 ≡ JUST THREE

In line with the above considerations, it appears that the numeral plurals should be handled as shown in Fig. 4. When we take the direct product with the simple hierarchical diagram of terms given in Fig. 2, the result is rather obvious, but involved.

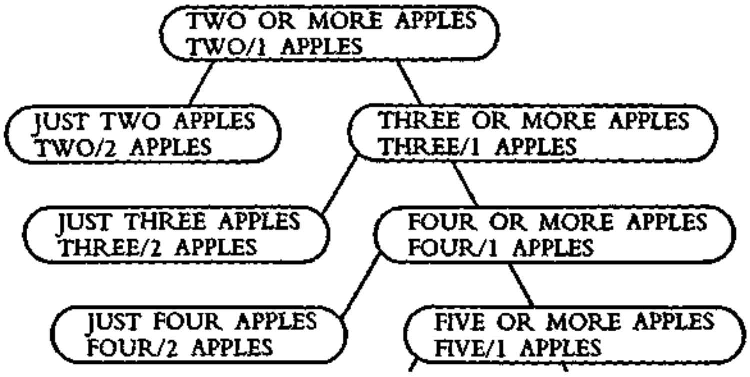


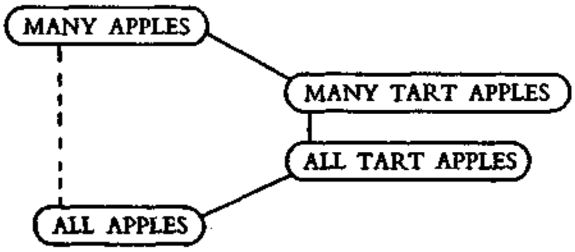
FIGURE 4. Retrieval diagram for sentences with numeral modifiers.

The results with the modifiers SOME and MANY show that for retrieval purposes we should set up the following equivalences:

Ø X Ø ≡ SOME X Ø
A X Ø ≡ ONE/1 X Ø ≡ ONE OR MORE X S
Ø X S ≡ TWO/1 X S ≡ TWO OR MORE X S
SOME X S ≡ THREE/1 X S ≡ THREE OR MORE X S
MANY X S ≡ FOUR/1 X S ≡ FOUR OR MORE X S

[1] Calvin N. Mooers, “A Mathematical Theory of Language Symbols in Retrieval,” page 1327.

The results with ALL X S and EVERY X Ø show that they should be considered equivalent in Dialect B for retrieval purposes. Furthermore, it is clear that they should be at the bottom end of the series of numeral adjectives discussed above, assuming that *all* is a great many. However, when one tries to take the direct product and draw a diagram of the retrieval hierarchy, one meets with a surprise. The question ALL TART APPLES retrieves the descriptor ALL APPLES. This is the reverse of all the other inclusions between TART APPLES and APPLES, and is shown in Fig. 5. It is clear that in the case of ALL, we cannot make use of the direct product as we did in the other cases.



Note: We assume that there are many apples.

FIGURE 5. Reversal of inclusion relations between sentences with ALL.

ALL is not the only modifier that exhibits this reversal of the hierarchical inclusion relations among the terms. ANY is another that behaves in this fashion, as in Fig. 6. There is another peculiarity with the word *any*. Of its several meanings, one can be used only in questions (and also in negative statements, but these are outside of Dialect B). In spoken English, the different uses of the word *any* are partly separated by stress:

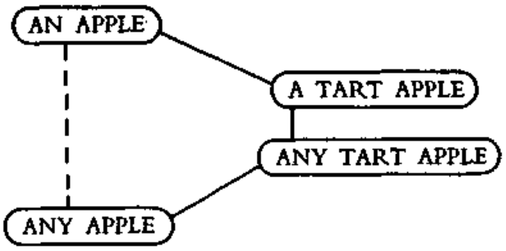


FIGURE 6. Reversal of inclusion relations between sentences with ANY.

Does this document show any$_1$ apples?
No, this document doesn't show any$_1$ apples.
Yes, this document shows an apple.
Does this document show *any*$_2$ apples?

No, this document doesn’t show *any*$_2$ apples, it shows some *tart* apples. Yes, this document shows *any*$_2$ apples.

In its first use, *any* with the singular is used to question the uncountable category of mass nouns, whereas with the plural it is used to question the singular and plural (count nouns). In its second use, *any* is pronounced with stress and carries a meaning related to *any kind of, any species of,* or *any whatsoever,* again with the word *any* carrying stress. This is the use of ANY in Fig. 6.

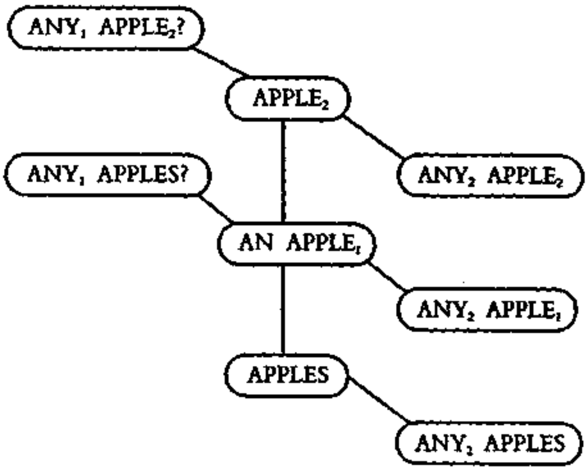


FIGURE 7. A retrieval diagram involving two meanings of ANY.

In Fig. 8, the phrases are from English rather than Dialect A, and various words are underlined to indicate where stress is to be placed when reading aloud. *Apple* has been subscripted to indicate the mass-count distinction in Figs. 7 and 8: subscript 1 for count singular, subscript 2 for mass.

Further investigation would be required to determine the degree of generality of these results. There may be classes of nouns that behave differently from *apple*. It is also not yet completely clear that the relation between APPLE and TART APPLE should be treated in the same way for search purposes as the different kind of relation between HALOGEN and CHLORINE.

The computer program for the grammar of Dialect B should be able to derive from the given question all the other possible implied questions that are needed for matching with the descriptors in the document, or it should derive possible questions from the document descriptors, or some combination of the two. Let us investigate only the former, deriving all descriptors that should be retrieved by the question.

We could, as we did in Dialect A, arrange the grammar in such a way that by moving down through the structure by a series of rules, one could come eventually to all of the points covered by the question. A better way, however, is not to work with the direct product, but to operate with two structures, one for each part of the sentence, the modifier and the term. This factorization of the product into modifier and term results in a great simplification in the program. One has a structure for the modifier in which one moves downward in order to find the points for which to search. At each point there is the possibility of choosing any of a series of terms from the hierarchy of terms. It is essential that the program go through the rules for the modifiers first before going to the hierarchy of terms because in some cases one has to move up through the hierarchy of terms, and in other cases one has to move down in the hierarchy. We will indicate this by giving the structure for the modifiers, and instead of using X to indicate where the various terms from the hierarchy should be placed in turn, we use A to represent a term and all those above it, i.e., more general, and V to represent a term and all those below it, i.e., more specific. A points up, V points down. We can then reserve the symbol X for those terms for which no other term in the hierarchy can be substituted. This has been done in Fig. 9. Modifiers involving ANY/1 can be used only in questions. All other modifiers can be used either for questions or for encoded document descriptors.

The grammar will be contained in the machine as a series of rules, much like the rules described for Dialect A. A possible method of search is then as follows. Locate the modifier from the question in the modifier structure and make a list of it and all modifiers below it in the structure. For each modifier in the list, make another list of it with all the terms (from the hierarchy of terms) that are either above or below the term in the question, as required, and then search the file for an occurrence of one of these derived descriptors.

There is, of course, the trouble that the list of modifiers is an infinite one, containing as it does all of the numbers. This and other problems affecting the speed of search can be handled by a slightly more sophisticated routine in a rather straightforward manner. Many of the techniques of search involving rules of progression, ordering, and screens, worked out for the HAYSTAQ system can be used with advantage.

Assessing where we are

Let us examine Dialect B in the light of a clear and concise statement of objectives. The first objective that is stressed is that a search is concerned much

[2] B. E. Lanham, J. Leibowitz, H. R. Koller, and H. Pfeffer, Organization of Chemical Disclosures for Mechanized Retrieval, *Patent Office Research and Development Report No. 5,* U. S. Patent Office, June 14, 1957.

[3] Don D. Andrews and Simon M. Newman, Activities and Objectives of the Office of Research and Development in the U. S. Patent Office, *J. of the Patent Office Soc., 40,* [2], 79-85 (1958).

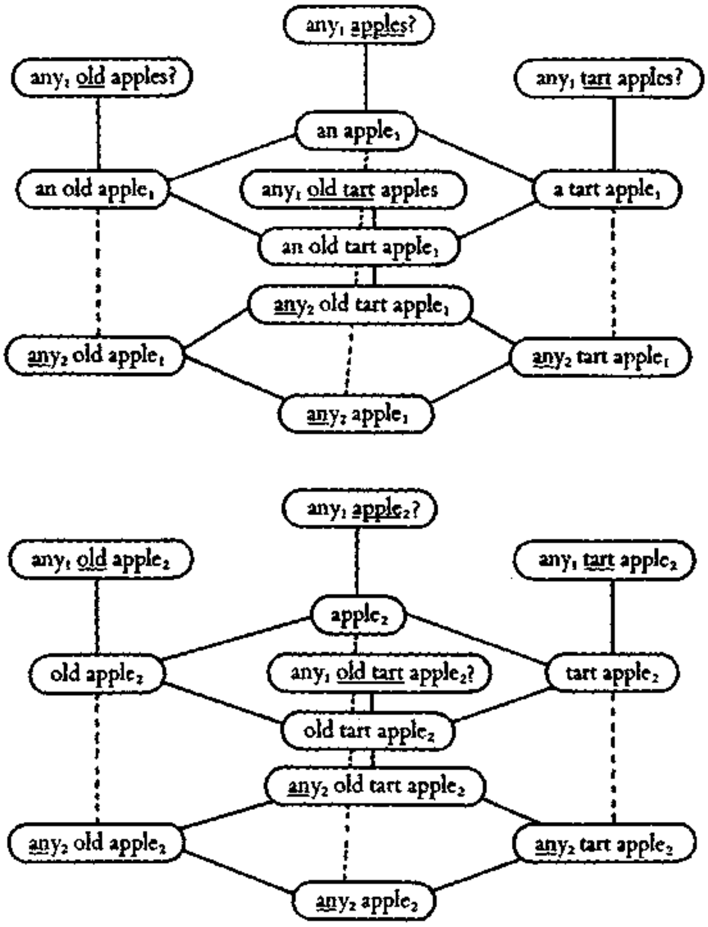


FIGURE 8. Retrieval diagrams involving the modifier ANY, and various hierarchical relations between the terms.

more with interrelations between two or more elements than with the number of items or the detail with which they are found. This objective has not yet been met. It will be approached in later dialects. It seems necessary that we concern ourselves in the earlier dialects with how to describe and search for elements or items within the terminological structure of English before we tackle the problem of searching for them in combination or for their interrelations. It is worth mentioning, however, that we can already search in Dialect B for A CONNECTION, A MANUFACTURING PROCESS, A SUPPORT. THREE INTERMESHED GEARS, etc. We do have the disadvantage at present that we have to enter as terms, as in Dialect A, MANUFACTURING PROCESS, INTERMESHED GEARS, etc. We are not yet able to build these expressions up from their elements. This will come later, at the time when it seems to fit logically into a dialect and can be added easily.

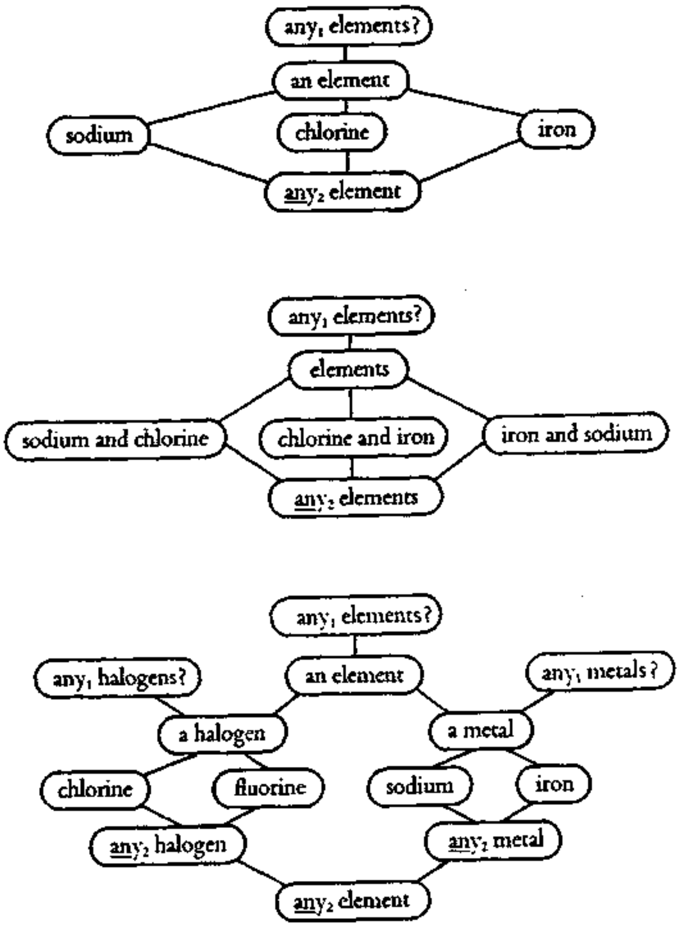


FIGURE 8. *(Continued)*

Even Dialect A had the capabilities of meeting the second objective, that every statement of a technical article should be retrievable in any frame of reference. The statements are searched for directly in their entirety. When new documents are added to the system it is only necessary to provide them with descriptors. In the ultimate system, the English of the document itself will be the descriptor for the document. The logic of relevance is determined on the basis of the grammar in the machine, and not on the encoded form of the descriptors. For this reason, it is very easy to incorporate the developing experience of the users of the system. In order to expand or change the logic of inclusion that the system operates with, it is only necessary to modify the grammar in the machine. One does not have to alter the codes associated with the documents.

The third desired feature of patent application searching was the ability to retrieve a species when a genus is requested, and also, where applicable, retrieve

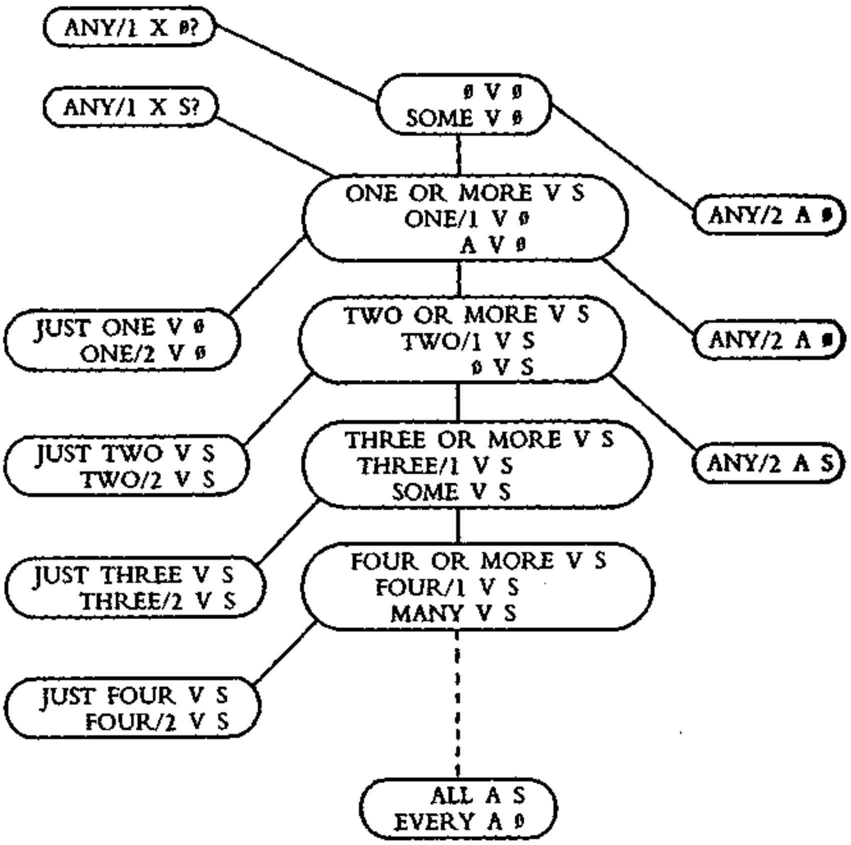


FIGURE 9. The final retrieval diagram for sentences in Dialect B.

a genus when a species is requested. This requirement was not met in Dialect A, but it has now been met in Dialect B with its inclusion of ANY, EACH, and ALL.

The fourth and fifth features, ability to recognize alternative items and the ability to handle implicit or explicit absence of features have not yet been included.

Probably the most important problem remaining is the first one, representing interrelations. Our procedure in deciding what to add to a dialect **to** get another dialect is to try to add what seems to be most needed. If it cannot be added because the dialect must have something else first, then we try to **add** the prerequisite. The final decision can be considered as a compromise between what is most needed and what can be added most easily at that stage.

Remaining problems

Many questions have been raised by this investigation that should be answered. Some have already been mentioned; some others are listed here. Every question that is answered satisfactorily will take us just that much closer to our ultimate goal of being able to search English texts directly.

1. What other modifiers in English share with the modifiers of Dialect B the property of being clear without reference to the immediate context?
2. Which of these behave in an identical fashion for search purposes to the ones treated here?
3. What is the behavior of the others in the retrieval situation, and how can they be incorporated in a dialect?
4. Are there classes of terms that would require a different hierarchy of modifiers?
5. Can the meaning changes of some terms when used with different modifiers be systematized?
6. What would have to be done to introduce the descriptive adjectives into a dialect separately from the nouns so as to reduce greatly the number of items that would have to be stored in the grammar?
7. Are the inclusion relations between the same noun with different descriptive adjectives independent of the choice of noun?
8. Are the inclusion relations of different nouns with the same descriptive adjectives independent of the choice of adjective?