IV Information Analysis and Dictionary Construction

G. Salton and M. E. Lesk

1. Introduction

At the base of any information system must always be a system of information analysis, used to decide what a given information item, or a given search request is all about. In a conventional library system, this analysis may be performed by a human agent who uses established classification schedules to decide what category, or categories, will most reasonably fit a given item. In certain other well known indexing systems, keywords or index terms may be manually assigned to documents and search requests, to be used for the identification of information content.

Regardless of what type of analysis is performed, and in particular regardless of whether the analysis is done manually or automatically, it is necessary to start with a set of carefully prepared instructions specifying the allowable steps, and setting forth in detail the meanings and implications of choosing one or another of the permissible alternatives. These instructions often take the form of dictionaries of various types, listing the allowable information identifiers, and giving for each a definition which regularizes and controls its use. As will be seen, such dictionaries may take a variety of forms, including almost always so-called "see" references which provide links for entries to be replaced by other preferred terms, and "see also" references which designate cross-references applicable to the dictionary items. Negative dictionaries may also exist, containing terms or categories which should not be used for purposes of information identification.

In view of the importance of the initial information analysis and classification — all later search and retrieval operations are of course of no avail in the absence of a careful and consistent determination of information content — it is appropriate to examine in detail the problems connected with the generation and use of dictionaries. Accordingly, the present study specifies the form of a variety of dictionaries which have been found useful in information analysis, and examines some of the principles of dictionary construction. Emphasis is placed on those dictionaries which can be used for natural language analysis, since many of the information items and of the search requests to be stored may be expected to be expressed by words or word strings in the natural language. Performance characteristics are given, based on search results obtained with various dictionaries, and several methods are suggested for the construction of dictionaries by semi-automatic means.

2. Language Analysis

Consider the problem of taking a document or search request in the natural language, and of attempting to use some automatic procedure to generate content identifications for the input texts. Such a task immediately raises many difficulties brought about by the complexity of the language, and by the irregularities which govern the syntactic and semantic structure. The following principal problems must be dealt with [1]:

1) words which carry out syntactic functions but which do not contribute directly to the specification of information content

must often be eliminated (but some words, such as "can" may occur both as significant and non-significant words);

2) many distinct words may be used to supply the same or related meanings; such synonymous words or expressions must be recognized if an accurate content analysis of documents and search requests is to be undertaken;

3) many words can be used in several different senses depending on the context (for example, a word like "base" may variously represent military bases, lamp bases, bases in baseball, and so on); it is important to identify such homographs, and if possible to recognize the proper meaning in a given context;

4) many types of syntactic equivalences occur in the language, where completely different constructions are used to represent the same general idea; as an extension of the overall synonym problem, it is important to recognize at least the principal types of syntactic paraphrasing;

5) the use of indirect references is prevalent in the natural language, where pronouns, collective names, and other particles are used to refer to entities presumably known by the context; the identification of the proper antecedents of such pronouns is difficult, particularly for cases where many different words can operate as antecedents;

6) relations may exist between words which are not explicitly contained in the text, but which can be deduced from the context, or from other texts previously analyzed; the identification of such relations requires deductive capabilities of considerable power;

7) the meaning of many words may change with time, or contrariwise, new words may be created to refer to entities previously referred to in different terms (for example, the unit of time previously known as "millimicrosecond" is now generally known as "nanosecond").

If the natural language is used as primary input to an information system, any content analysis system will have to include methods for consistent language normalization. One of the most effective ways for providing such a normalization is by means of suitably constructed dictionaries. The following types of dictionaries appear to be of interest in this connection:

1) a negative dictionary containing terms whose use is proscribed for content analysis purposes;

2) a thesaurus, or synonym dictionary, which specifies for each dictionary entry, one or more synonym categories, or concept classes; ambiguous entries are then replaced by many concepts and many different words (synonyms) may map into the same concept category; a thesaurus is then used to perform a many-to-many mapping from word entries to concept classes;

3) a phrase dictionary may be used to specify the most frequently used word or concept combinations (called phrases); such a phrase dictionary can often increase the effectiveness of a content analysis by assigning for content identification a relatively unambiguous phrase, instead of two or more ambiguous components (for example, the terms "program" and "language" are more ambiguous, standing alone, than the phrase "programming language");

4) a hierarchical (tree-like) arrangement of terms or concepts, similar to a standard library classification schedule, which makes it possible, given a certain dictionary entry to find more general concepts by going up in the hierarchy, or more specific ones by going down (for example, from a concept such as "syntax", one can obtain the more general "language", or the more specific "punctuation").

Dictionaries do not, of course, completely eliminate language ambiguities, but they can serve to reduce the effects of many irregularities by using appropriate dictionary mapping algorithms. For example, a correspondence between a word and a single concept may receive a higher weight than one between a word and a multiplicity of concepts, since the former presumably implies a unique meaning for that word while the latter implies ambiguity.

Even if almost all terms used in a given context are inherently ambiguous, the juxtaposition of many multiple mappings can often identify the appropriate concept classes with reasonable accuracy. The relevant categories will normally be reinforced, since they apply to many terms, while the extraneous categories will be randomly distributed.

Consider, for example, the set of terms: "base", "bat", "glove", "hit". Each term is ambiguous, and a given multiple thesaurus mapping may specify the correspondences shown in Table I. In that table, three categories are shown for the word "base", and two categories for each of the other terms. Despite the apparent ambiguities, a document identified by the four original terms can nevertheless be assigned to the "baseball" class with reasonable expectation of success, since the other categories occur more or less at random for the given terms, whereas the "baseball" class is always present.

The principal advantages of synonym and phrase dictionaries for purposes of content identification may then be summarized as follows:

1) they permit a consistent assignment of concept classes to items of information thereby replacing either keywords and index terms assigned to documents and search requests, or the words occurring in them;

2) they can often be used to resolve ambiguities by looking at the pattern of occurrence of the concepts;

3) they can serve for the analysis of many different subject fields and for different types of usage, since it is possible to adapt the dictionary to the particular search environment.

Concept Classes / Original Terms	Lamps	Games Baseball	Animals	Military Usage	Clothing
base	✓	✓		✓	
bat		✓	✓		
glove		✓			✓
hit		✓		✓	

Sample Thesaurus Mapping

Table I

On the negative side, dictionaries are often difficult to construct, particularly if the environment within which they are expected to operate is subject to change; furthermore most dictionaries are useless unless their mode of usage is consistent for all operations. Obviously if a dictionary is used in one way for information classification and in another for information searching, an effective result cannot be guaranteed.

Various thesaurus types are examined in more detail in the next few paragraphs.

3. Dictionary Construction

A) The Synonym Dictionary (Thesaurus)

As previously explained, a thesaurus is a grouping of words, or word stems, into certain subject categories, hereafter called concept classes. A typical example is shown in Fig. 1, where the concept classes are represented by three-digit numbers, and the individual entries are shown under each concept number. In Fig. 2, a similar thesaurus arrangement is shown in alphabetical order of the words included. The concept numbers appear in the middle column of Fig. 2 (concept numbers over 32,000 are attached to "common" words which are not accepted as information identifiers); the last column consists of one or more three-digit syntax codes attached to the words to be used for purposes of syntactic analysis.

When constructing a thesaurus to be used for vocabulary normalization, one immediately faces three types of problems: first what words should one include in the thesaurus; secondly, what type of synonym categories should one use (that is, should one aim for broad, inclusive concept classes, or should the classes be narrow and specific); finally, where

408 CISLCCATION
JUNCTICN
MINORITY-CARRIER
N-P-N
P-N-P
POINT-CONTACT
RECOMBINE
TRANSITION
UNIJUNCTION

409 BLAST-COOLED
HEAT-FLOW
HEAT-TRANSFER

410 ANNEAL
STRAIN

411 COERCIVE
CEMAGNETIZE
FLUX-LEAKAGE
HYSTERESIS
INDUCT
INSENSITIVE
MAGNETORESISTANCE
SQUARE-LOOP
THRESHOLD

412 LONGITUDINAL
TRANSVERSE

413 CAPACITANCE
IMPEDANCE-MATCHING
IMPEDANCE
INDUCTANCE
MUTUAL-IMPEDANCE
MUTUAL-INDUCTANCE
MUTUAL
NEGATIVE-RESISTANCE
POSITIVE-GAP
REACTANCE
RESIST
SELF-IMPEDANCE
SELF-INDUCTANCE
SELF

414 ANTENNA
KLYSTRON
PULSES-PER-BEAM
RECEIVER
SIGNAL-TO-RECEIVER
TRANSMITTER
WAVEGUIDE

415 CRYOGENIC
CRYOTRON
PERSISTENT-CURRENT
SUPERCONDUCT
SUPER-CONDUCT

416 RELAY

THESAURUS EXCERPT IN CONCEPT NUMBER ORDER

Fig. 1

	CONCEPT NUMBERS	SYNTAX CODES
BLOCK	663	070043040
BLUEPRINT	58	070043
BOMARC	324	070
BOMBARD	424 0343	043
BOMBER	346	070
BOND	105	070043
BOOKKEEPING	34	070
BOOLEAN	20	001
BORROW	28	043
BOTH	32178	008080012
BOUND	523 0105	070043134135
BOUNDARY	524	070
BRAIN	404 0235	070
BRANCH	48 0042	070042
BRANCHPOINT	23	070
BREAK	380	043040070
BREAKDOWN	689	070
BREAKPOINT	23	070
BRIDGE	105 0458 0048	070043
BRIEF	32232	001043071
BRITISH	437	001071
BROAD-BAND	312	001071
BROKE	380	134104
BROKEN	380	135105
BUFFER	24	070043
BUG	69	070
BUILD	80	043
BUILT	80	134135
BULK	558	070
BURNOUT	69	070
BUS	61	070
BUSINESS	472	070
BUT	32027	091012
BY	32020	074013
BYTE	31	070
C-1100	155	070
CALCULATE	605	043040
CALCULATOR	237	070
CALCULUS	506	070
CALL	32283	070043045040
CAMBRIDGE	444	070
CAN	32118	009
CANCEL	385	043
CANNED	182	134135
CANNING	182	136137071001
CANNOT	32102	009
CANONICAL	706	001
CANS	182	133
CAPABILITY	32269	070
CAPABLE	32269	001071
CAPACITANCE	413	070
CAPACITOR-DIODE	228	071001
CAPIT	340 0213	043
CARD	27	070
CARE	32186	070040
CARGO	331	070
CARRIER	316 0061	070
CARRY	28	070043040

THESAURUS EXCERPT IN ALPHABETIC ORDER

Fig. 2

should each word appear in the thesaurus structure (that is, given a word, what are to be its assigned concept classes).

Consider first the words to be included. There is usually not much question about the fact that common function words (such as "and", "or", "but") should not appear in the synonym dictionary, since these words out of context provide no indication of subject matter. A significant problem does, however, arise in connection with very frequent words. These may be non-technical words in the general vocabulary such as "discuss" and "make"; or they may be technical words which, in their particular environment, are in effect reasonably common. For example, in a collection dealing with computer science, such words as "machine", "computer", or "automatic" are in effect common words with reasonably high frequency. If such frequent words are included in a synonym dictionary, most documents will exhibit occurrences of these words, and therefore significant matching coefficients may be obtained between documents and requests, even though the technical texts may be really quite dissimilar (except for the fact that they may deal with computers); if on the other hand these words are excluded, it then becomes possible that one or another document cannot be retrieved when in fact it is pertinent. Obviously some compromise must be made as usual, between one's interest in retrieving everything even remotely useful (that is, between the necessity of obtaining high "recall"), and the need not to obtain too much extraneous material (the need for high "precision").

A similar problem arises in connection with very low frequency words. If, for example, a term such as "Morse Code" is excluded from the dictionary, then the very few documents dealing with this type of code may not be retrievable. On the other hand, if "Morse Code" appears in a thesaurus category together with many other types of coding systems, then a request for "Morse Code" could also produce many other documents dealing with coding systems, but not with the specific system wanted.

Once the words to be included in the dictionary are chosen, the second main problem which arises is the one dealing with the type of synonym categories to be created. It is clear that if very broad and somewhat fuzzy categories are wanted, such that a given category includes both somewhat specific terms and also somewhat broader ones, then the resulting dictionary will in general interpret a question in a reasonably broad sense, and as a result the recall, that is the proportion of relevant documents retrieved, will likely be rather high. At the same time the precision may be low, since it must be expected that much irrelevant material will also be produced in the process. If on the other hand the categories are very specific, the chance of picking up irrelevancies is much smaller and therefore the precision is increased; the recall may suffer, however, since relevant matter is likely to be missed at the same time. In either case, that is whether the categories used are broad or specific, problems will arise if words with very different frequency characteristics are included in the same category. Obviously the effectiveness of the specific terms is much smaller, if these terms are in fact considered equivalent to broader terms of higher frequency by the applicable thesaurus mapping.

This discussion then raises the possibility of providing different thesauruses for different types of questions. Specifically, if it is expected that the user is interested in reasonably complete retrieval, including most everything that is likely to be useful, then the thesaurus with broad categories which provides high recall and low precision should be used. On the other hand if only a few items are to be retrieved, but the user insists that these items must be relevant, then the specific thesaurus categories will prove more useful. This then confirms the well-known fact that any kind of retrieval tool must be constructed with the retrieval environment in mind in which it is expected to operate.

Concerning now the problem of where a given term is to be put within a given thesaurus organization, this depends largely on the type of user which may be expected to avail himself of the retrieval systems. As an example, dictionaries constructed for a population of students may be expected to require an organization somewhat different from that which would be useful to advanced research scientists. The latter might, for example, be interested in the specific physical characteristics of certain devices, whereas the former are more interested in the uses of the devices. A "transistor" could then appear in a category under "three terminal switching devices", if the users were to be engineers, but it would appear under "computer components", for a user population consisting of computer programmers.

The following principles of thesaurus construction may then be enunciated:

1) no very rare concepts should be included in the thesaurus since these could not be expected to produce many matches between documents and search requests;

2) very common high frequency terms should also be excluded from the dictionary, since these produce too many matches for effective retrieval (it is in fact possible to replace individual high frequency terms by much more specific compound or hyphenated terms; for example, terms such as "computer" or "control" might

well be eliminated in favor of a term such as "computer-control", since the former are clearly ambiguous in many contexts whereas the latter is much more specific);

3) non-significant words should be studied carefully before any are included in the list of words to be eliminated (for example, a term such as "hand" should be included in a thesaurus dealing with biology, but it should not be included if its high frequency count is due to expressions such as "on the other hand");

4) ambiguous terms should be coded only for those senses which are likely to be present in the document collections to be treated (for example, at least two category numbers must be shown for the term "field", corresponding on the one hand to the notion of subject area, and on the other hand to its technical sense in algebra; however, no category number need be shown to cover the notion of "a patch of land" if the dictionary deals with the mathematical sciences or related technical fields);

5) each concept class should only include terms of roughly equal frequency so that the matching characteristics are approximately the same for each term within a category.

Consider as an example some of the synonym dictionaries constructed for use with the SMART retrieval system. In that system it was found useful to operate with a reasonably large number of concept classes (of the order of 700 for a given restricted subject field), and to use also a large list of non-significant words to be excluded from the content indications. This list includes in particular verbs such as "begin", "contain", "indicate", "call", "designate" etc., which could not be depended upon to provide safe content indication. It was also found useful to isolate high frequency terms into separate categories so that these terms would not impair the retrieval effectiveness of other more specific terms.

Consider as an example of the kind of analysis which is normally necessary for dictionary construction the concept number 101 representing the notion of "tag". The word list attached to this concept originally included terms such as "call", "designate", "identify", "identifier", "identification", "index", "indicate", "label", "mark", "name", "point", "signal", "sign", "subscript", and "tag". The concept occurred in 94 documents out of some 500, with the following distribution of significant terms:

Term	Frequency	Number of Documents
index	17	7
signal (pulse)	20	14
identify	6	4

All other terms under concept 101 occurred a total of 91 times, accounted for almost exclusively by the terms "pointed out", "indicated", and "call". As a result of the analysis, the words "indicate", "call", "name", and "designate" were removed from category 101 and were included in the list of common words; the words "sign" and "signal" were also removed from category 101, since they seemed to occur in the document collection only in the sense of "pulse signal" and therefore not in the sense of "tag"; words with stem "identi" accounting for "identifier", "identification", "identify", etc., were moved to a new concept number representing the idea of recognition. At the end only the terms "index", "label", "subscript" and "tag" remained under category 101.

Performance figures which measure the efficacy of various types of dictionaries are given later in this report. Several methods of semi-automatic thesaurus construction using aids in the form of frequency lists and word concordances are also described.

B) The Null Thesaurus and Suffix List

One of the earliest ideas in automatic information retrieval was the suggested use of words contained in documents and search requests for purposes of content identification. No elaborate content analysis is then required, and the similarity between different items can be measured simply by the amount of overlap between the respective vocabularies. While one should not expect that word matching techniques alone will normally provide adequate retrieval performance, it is nevertheless useful to consider a word matching technique as part of a retrieval system, since this provides a standard against which various types of dictionary procedures may be measured. This was one of the reasons for including in the SMART system the so-called <u>null thesaurus</u>.[2,3]

The null thesaurus consists simply of a list of word stems, constructed by using the words included in a typical document collection, each distinct word stem being furnished with a different sequence number. The sequence numbers in the null thesaurus are then equivalent to the concept numbers included in the regular thesaurus, with the exception that each sequence number, of course, has only a single correspondent (words or word stem) in the null thesaurus, compared to the possible multiple correspondences in the regular thesaurus. A typical sample from a null thesaurus is shown in Fig. 3, where the word stems are listed in the order of increasing frequency of occurrence within a document collection, rather than in the usual alphabetic order.

Clearly, the operation which consists in using the sequence numbers obtained from a null thesaurus for purposes of document and request identification leads effectively to a word matching technique for document retrieval, since sequence numbers and text words are in effect isomorphic. The main virtues of the null thesaurus <u>per se</u> result from the fact that the dictionary look-up routine programmed for the regular thesaurus will serve also for the null thesaurus (because the structure of the two thesauruses is the same), and that the null thesaurus permits the word matching operation to be confined to only those words actually included in the thesaurus (since the others will not have an assigned sequence number).

This raises a question about the type of null thesaurus which should be used as a standard for the word matching operations. The following alternatives appear of principal importance in this connection:

1) the null thesaurus can include complete English words, or can alternatively be made up from <u>word stems</u>, obtained from the original words by a suffix cut-off;
2) an entry can be included in the null thesaurus for each text word included in a certain document collection, or expected to be important in a given topic area; or, alternatively, function words and other words not easily used for content identification may be excluded, or marked with a special identifying code;
3) all non-common words, or word stems may be used, or only those words which have certain predetermined frequency characteristics (for example, words occurring more than 5 times but less than 100 times in a given document collection).

In the SMART system, all dictionaries (including regular and null thesauruses) are based on word stems rather than original words; furthermore, common words appear on an exclusion list, and are thus not

FRE-QUENCY	STEM	SUFFIX	SEQUENCE NUMBER	FRE-QUENCY	STEM	SUFFIX	SEQUENCE NUMBER
11	MOCULE	S	2099	12	MANIPUL	ATION	2129
11	PLACE	S	2100	12	MECHAN	ISM	2130
11	RESPONSE		2101	12	MODUL	ATION	2131
11	RF		21C2	12	MUCH		2132
11	SOURCE		2103	12	OSCILL	ATORS	2133
11	THICK		2104	12	PHYS	ICAL	2134
11	TRUNC	ATION	21C5	12	PREV	IOUS	2135
11	WAVE		2106	12	RANGE		2136
11	WHEREB	Y	2107	12	RECCRD		2137
11	WIR	ING	2108	12	RELAX	ATION	2138
12	ALPHABET	ICAL	2109	12	REPCRT	ED	2139
12	BASE		2110	12	REVERS	ED	2140
12	CAP	ABLE	2111	12	RULE	S	2141
12	CENT		2112	12	SATIS	FY	2142
12	CONCEPT		2113	12	SHCW		2143
12	DECIS	ION	2114	12	STUC	Y	2144
12	DEPCSIT	ED	2115	12	SYSTEMAT	ICALLY	2145
12	DUE		2116	12	TREE	S	2146
12	ECCNOM	ICAL	2117	12	TUNNEL		2147
12	ESAKI		2118	13	10		2148
12	EXAMIN	ED	2119	13	650		2149
12	FUNCTICN	AL	2120	13	ANISOTROP	Y	2150
12	GRAPH		2121	13	ASSUM	ED	2151
12	HAV	ING	2122	13	CARRI	ER	2152
12	IMPROVE	MENT	2123	13	CAR	RY	2153
12	IMPROV	ED	2124	13	COMM	ON	2154
12	INDIVIDU	AL	2125	13	COMMUNIC	ATIONS	2155
12	LEAST		2126	13	COMPOS	ITION	2156
12	MAGNETIZ	ATION	2127	13	DEPCNSTR	ATE	2157
12	MAIN		2128	13	DENS	ITY	2158

WORD STEM FREQUENCY LIST
(NULL THESAURUS)

Fig. 3

included in any of the dictionaries. Experiments were conducted with the SMART system, using both unrestricted vocabularies (full null thesaurus), as well as frequency restricted entries (partial null). A sample set of document abstracts of some 50,000 total running words, would typically produce a full null thesaurus of about 2,800 distinct word stems, and a partial null dictionary of about 900 stems (assuming a frequency of at least four occurrences for each entry listed).

If it is desired to list word stems, rather than full words, these must of course first be generated by a suffix cut-off system. To this effect, a suffix dictionary is built, a typical example of which is shown in Fig. 4. The lookup procedure in this suffix dictionary is described in the next chapter together with the lookup procedures for the other dictionaries. The structure of the suffix dictionary may, however, be examined immediately. It may be seen from Fig. 4 that each suffix is listed with a sequence number and with one or more syntactic codes. The latter may be used if it later becomes necessary to recombine stems and suffixes into complete, acceptable words, as may be required, for example, to carry out a syntactic analysis.

The syntactic codes included in the suffix dictionary represent only partial homographs which must be combined with complementing codes attached to the word stems in order to determine which suffixes match which stems. (The syntactic codes attached to the word stems included in the null thesaurus are not shown in the output of Fig. 3.) For example, a partial homograph such as OT1O from the null dictionary will combine with a partial homograph code from the suffix list, such as VOOSO, to form a complete homograph. In this case the complete code is VTISO, indicating a single object transitive verb in the third person singular.

Alphabetic Suffix List		Syntactic Suffix Codes				
FICATION	058	058	NØUS			
FICATIONS	059	059	NØUP			
FIED	060	060	VOOCO	POO O	ADJ	
FIER	061	061	NØUS			
FIERS	062	062	NØUP			
FIES	063	063	VOOSO			
FOLD	064	064	ADJ	NØVC		
FUL	065	065	ADJ	NØVC		
FULLY	066	066	AV1			
FY	067	067	VOOPO	IOO O		
FYING	068	068	ROO O	GOOSO	NØVS	ADJ

Typical Suffix Dictionary Entries

Fig. 4

A typical suffix dictionary for English suffixes may contain about 200 entries. To simplify the look-up algorithm, noun suffixes may be entered in the plural as well as singular forms, and adjectival suffixes may also be listed in the adverbial form. Verb suffixes should include the common endings "ed", "ing", and "s", as well as true verb suffixes such as "fy" with their inflected forms. (Multiple suffixes, such as "fying" could be detected by a dual scanning of the suffix list, looking first for "ing" and then for "fy"; a dual scan is avoided if such multiple suffixes are also entered in the suffix dictionary.)

In general, it is possible to encode word stems and suffixes in such a way that no ambiguity results when the fragments are combined into full words. For example, the stem "recti" is coded as a potential verb because it can form "rectify"; the stem "reduct", on the other hand, is carried without syntax codes, since it can be combined only with common suffixes such as "ion" and "ible" which by themselves are carried as complete homographs, representing respectively "noun singular" and "adjective".

In a limited number of cases, partial syntactic coding may introduce an ambiguity: if the word "capital", for example, is coded as a potential verb to accept the suffix "ize", the plural noun "capitals" will receive the extraneous coding of a verb in the third person singular. This difficulty may be prevented by entering the stem "capit" with a partial verb code. The suffix "als" properly carries with it only the plural noun code, and "capitalize" can then be found by a double scan of the suffix list.[2]

C) The Phrase Dictionaries

Both the regular as well as the null thesauruses are based on entries corresponding either to single words or to single word stems. In attempting to perform a subject analysis of written text, it is possible however, to go further by trying to locate "phrases" consisting of sets of words which are judged to be important in a given subject area. For example, in the field of computer science, the concepts of "program" and "language" may mean many things to many people. On the other hand, the phrase concept which results from a combination of these individual words, that is, "programming language" has a much more specific connotation. Such phrases can be used for subject identification by building phrase dictionaries to be used in locating combinations of concepts, rather than individual concepts alone. Such phrase dictionaries would then normally include pairs, or triples, or quadruples of words or concepts, corresponding in written texts to the more likely noun and prepositional phrases which may be expected to be indicative of subject content in a given topic area.

Many different strategies can be used in the construction of phrase dictionaries. For example, it is possible to base phrase dictionaries on combinations of high-frequency words or word stems occurring in documents and search requests; alternatively, one may want to use a thesaurus before appeal is made to a phrase dictionary. Under those circumstances, the phrase dictionary would then be based on combinations of concept categories included in the thesaurus, rather than on combinations of words.

Furthermore, given the availability of a phrase dictionary one can recognize the presence of phrases in a given text under a variety of circumstances: for example, the existence of a phrase may be recognized whenever the phrase components are present within a given document, regardless of any actual syntactic relation between the components; alternatively, the presence of a phrase may be inferred whenever the components are located within the same sentence of a given document, rather than merely within the boundaries of the same document; finally, even more stringent restrictions can be imposed before a phrase is actually accepted, by checking that a pre-established syntactic relation actually exists between the phrase components in the document under consideration.

In the SMART system, the phrase dictionaries are based on co-occurrences of thesaurus concepts, rather than text words, and two principal strategies are used for phrase detection: the so-called "statistical phrase" dictionary is based on a phrase detection algorithm which takes into account only the statistical co-occurrence characteristics of the phrase components; specifically a statistical phrase is recognized, if and only if all phrase components are present within a given document or within a given sentence of a document, and no attempt is made to detect any particular syntactic relation between the components; on the other hand, the "syntactic phrase" dictionary includes not only the specification of the particular phrase components which are to be detected, but also information about the permissible syntactic dependency relations which must obtain if the phrase is to be recognized. Thus, if it were desired to recognize the relationship between the concept "program" and the concept "language", then any possible combination of these two concepts such as, for example, "programming language", "languages and programs", "linguistic programs", would be recognized as proper phrases in the statistical phrase dictionary; in the syntactic dictionary, on the other hand, an additional restriction would consist in requiring that the concept corresponding to "program" be syntactically dependent on the concept "language". This eliminates phrases such as "linguistic programs", and "languages and programs", but would permit the phrases "programming languages", or "programmed languages".

A typical excerpt from a statistical phrase dictionary used in connection with the SMART system is shown in Fig. 5. It may be seen that up to six phrase components are permitted in a given phrase, but that the usual phrase specification consists of two, or at most three, components. With each phrase included in Fig. 5 is listed a phrase concept number which replaces the individual component concepts in a given document specification whenever the corresponding phrase is detected by the phrase processing algorithm in use. For example, the first line of Fig. 5 shows that a phrase with concept number 543 is detected whenever the concepts 544 and 608 are jointly present in the document under consideration. Whenever such a phrase concept is attached to a given document specification, the weight of the phrase concept can be increased over and above the original weight of the component concepts to give the phrase specification added importance.

Since the phrase components used in the SMART system represent concept numbers rather than individual words, a given phrase concept number does then in fact represent many different types of English word combinations depending of course on the number of word stems assigned to each component concept by the original thesaurus mapping.

The syntactic phrase dictionary has a more complicated structure as shown by the excerpt reproduced as Fig. 6. Here, each syntactic phrase also known as a "criterion tree" or "criterion phrase", consists not only of a specification of the component concepts, but also of syntactic indicators, as well as of syntactic relations which may obtain between the

PHRASE CONCEPT	COMPONENT CONCEPTS					
543	544	608	-0	-0	-0	-0
282	280	281	-0	-0	-0	-0
282	306	281	-0	-0	-0	-0
280	69	648	-0	-0	-0	-0
280	69	215	-0	-0	-0	-0
694	1285	1284	-0	-0	-0	-0
291	265	290	-0	-0	-0	-0
291	265	406	-0	-0	-0	-0
422	646	185	-0	-0	-0	-0
640	309	290	-0	-0	-0	-0
294	21	293	-0	-0	-0	-0
393	21	635	-0	-0	-0	-0
393	635	106	-0	-0	-0	-0
294	21	245	-0	-0	-0	-0
605	44	150	-0	-0	-0	-0
78	572	565	-0	-0	-0	-0
411	370	328	-0	-0	-0	-0
411	370	389	-0	-0	-0	-0
411	370	476	-0	-0	-0	-0
666	46	601	-0	-0	-0	-0
666	330	53	601	-0	-0	-0
666	347	46	-0	-0	-0	-0
666	347	290	-0	-0	-0	-0
666	347	601	-0	-0	-0	-0
666	357	290	-0	-0	-0	-0
666	46	353	-0	-0	-0	-0
666	347	353	-0	-0	-0	-0
666	430	353	-0	-0	-0	-0
377	347	478	-0	-0	-0	-0
666	347	406	-0	-0	-0	-0
666	478	406	-0	-0	-0	-0
381	381	150	-0	-0	-0	-0
287	618	14	-0	-0	-0	-0
287	618	509	-0	-0	-0	-0
620	618	621	-0	-0	-0	-0
355	326	62	-0	-0	-0	-0
690	460	689	-0	-0	-0	-0
119	202	209	-0	-0	-0	-0
693	352	175	-0	-0	-0	-0
693	352	290	-0	-0	-0	-0
693	352	330	-0	-0	-0	-0
292	631	108	-0	-0	-0	-0
316	316	619	-0	-0	-0	-0
296	512	600	-0	-0	-0	-0
539	1267	538	-0	-0	-0	-0
534	1267	255	-0	-0	-0	-0
650	1267	640	-0	-0	-0	-0
475	1267	473	-0	-0	-0	-0
541	1267	267	-0	-0	-0	-0
736	350	30	-0	-0	-0	-0
301	26	53	-0	-0	-0	-0
301	26	114	-0	-0	-0	-0
301	350	26	-0	-0	-0	-0
301	350	215	-0	-0	-0	-0
728	350	62	-0	-0	-0	-0
639	97	20	-0	-0	-0	-0
644	97	121	-0	-0	-0	-0
398	1213	560	-0	-0	-0	-0

EXCERPT FROM
STATISTICAL PHRASE DICTIONARY

Fig.

included concepts. For example, the first phrase shown in Fig. 6 carries the concept number 422, and the mnemonic indicator MAGSWI to indicate that this phrase deals in one way or another with magnetic switches. Fig. 6 also shows that the first component of the phrase must consist either of concepts 185 or 624, while the second phrase component must represent concept 225. The indicators after the dollar sign in the output of Fig. 6 carry the syntactic information. In particular, the information given for the phrase MAGSWI indicates that this particular phrase must be either of syntactic types 7, or 15, or 16.

More specifically, there exist four mail classes of syntactic specifications, corresponding respectively to noun phrases, subject-verb relations, verb-object relations, and subject-object relations. The four syntactic classes are in turn subdivided into approximately twenty syntactic types, each of which specifies a particular syntactic relation between the components. The particular relations which apply to a sample phrase, labelled SYNTAX, are shown in Fig. 7. It may be seen in the figure, that the first component of the phrase must correspond either to concepts 11 or 158, whereas the second component corresponds to concepts 102, 188, or 170. Also specified in Fig. 7 are the four allowable format types namely 1, 3, 4 and 13. These formats are specified in the center of Fig. 7 in the form of syntactic dependency trees.

Dependency trees are characterized by the fact that vertical displacement along a given path of the tree denotes syntactic dependence, the dependent structures being always listed below the corresponding governing structures. This can be illustrated by using the example of Fig. 7, where the format type 1 specifies that the second component,

NAME OF TREE	OUTPUT CONCEPT	FIRST NODE CONCEPTS	SECOND NODE CONCEPTS	TYPE 7 SERIAL 143	TYPE 15 SERIAL 398	TYPE 16 SERIAL 399

```
MAGSWI=422(185,624)/(225)$7/143,15/398,16+
MANMCH=517(600)/(516)$7/144,15/400
MANROL=286(290)/(113)$7/145,5+,15/401,16+,19+
MATHOP=594(615)/(7,116,376)$7/147
MCHBKD=69(689)/(600)$1/148
MCHCOD=304(102,281)/(14,41,600,601)$1/149,15/404
MCHOPE=93(615)/(600)$7/150
MCHORI=41(513)/(600,601)$7/151,15/405
MCHTIM=691(617)/(52,600,601,605,1281)$7/152
MCHTIM=691(617)/(77,615)$1/153
MCHTRA=303(98)/(119,600)$1/154,4+,5+,6+,10+,15/406,16+,19+
MEMACC=593(672)/(121)$1/159,15/409
MEMCOR=557(669)/(121)$7/137,15/395
MEMEFF=284(64)/(121)$1/160,6+,15/410
MEMSPA=552(212)/(121)$1/162,13+,15/411
MHTOOL=471(327)/(600)$7/164
MINSTA=294(245)/(220)$7/165,15/412
MISTRA=668(46,341)/(346)$1/166,3+,15/413
MISTRA=668(341)/(344)$1/168,15/414
MISTRA=668(406)/(341,657)$7/169
MISTRA=668(657)/(346)$7/170
MLTACC=481(672)/(55)$7/171,15/415
MTHMOD=722(1273)/(116)$7/172
MTHSIM=722(353)/(116)$7/173
NATLNG=283(102)/(35,179)$7/174,15/416,16+
NETENC=630(618)/(623)$7/175
NLNSYS=727(496,1273)/(297)$7/176
NOGCMP=91(601)/(604)$7/177,15/418
NOGDIG=288(604)/(603)$1/178,6+,15/419
NOGSIM=679(353)/(604)$7/180
NOGVRT=289(254)/(288,603,604)$7/181,15/420
NRETIM=354(617)/(482)$7/182,15/421
NRTROL=667(46,290,347,569)/(521)$7/183,15/422
NUCPOW=441(649)/(441)$7/184,15/423
NUMBAS=15(620)/(518)$7/185
NUMERI=375(610,619)/(625)$7/186,15/424
NUMQAD=723(307)/(625)$7/187
NUMRAT=529(255)/(50,313)$1/188
NUMSTR=262(428)/(625)$7/189
NUMTEG=723(384)/(625)$7/190
OPERES=597(162,613)/(615)$7/191,15/425
PARCHK=280(207)/(271)$1/192,3+
PATGEN=683(334)/(340,625)$1/194,3+,15/426,17+
PATGEN=683(509)/(340,563)$1/196,3+,15/428,17+
PATREC=567(332)/(340,563)$1/198,3+,4+,15/430,17+,19+
PHMTCH=586(585)/(77,564)$1/201,3+,15/433,17+
PLXNUM=390(518)/(241)$7/203,15/435
PLXVAR=390(621)/(241)$7/204,15/436
PNTCON=408(685)/(150)$7/205,15/437
POBCRI=519(513)/(602)$7/206,15/438
POBSLV=292(612)/(602)$7/207,15/439
POGOPE=93(596,615)/(608)$7/208
POLPOL=398(13)/(628)$7/209
POWDIS=652(349,648)/(649)$1/210,3+,15/440
POWLOS=653(389)/(649)$1/212,3+
POWSOU=655(514,656)/(649)$1/214,3+
POWSPC=733(556)/(649)$7/216
POWTSS=654(406,624)/(649)$7/217
```

EXCERPT FROM CRITERION TREE DICTIONARY

corresponding in this case to either concept numbers 102, 188 or 170, be syntactically dependent on the first component corresponding to concept 11 or 158; furthermore, the second component is specified as an adjective, whereas the first component is specified as a noun. Examples corresponding to each of the syntactic format frames listed are shown on the right-hand side of Fig. 7. For instance, the first tree of format type 1 might correspond to English phrases such as "syntactic analysis", "syntactic synthesis", "phrase relations", "subject correspondence", and so on. Because of the multiple assignment of concepts to phrase components, and the multiplicity of syntactic format types specified for each phrase, a given criterion phrase generally represents many hundreds of English phrases or sentences. This feature is used to match the many sentence parts in the language which are semantically similar, but syntactically quite distinct.

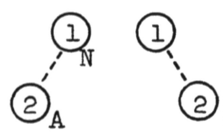
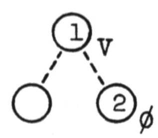
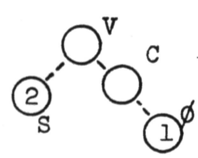
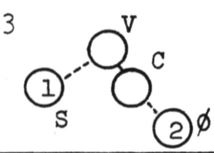
Since the syntactic dependency specifications are always directed from a dependent component to a governing component, the grammatical structure of a syntactic phrase, unlike that of a statistical phrase, is well determined. For the first example of Fig. 7 (format type 1) the string "phrase relations" is an acceptable interpretation, but not "relational phrase"; similarly for format type 13 an acceptable interpretation is "this analysis is applicable to Russian grammar", but the transposed "this grammar is applicable to Russian analysis" would not be accepted.

D) The Concept Hierarchy

Hierarchical arrangements of subject headings have been used for many years in library science and related documentation activities. In general, such arrangements make it possible to classify more specific topics under

PHRASE SPECIFICATION:

$$\text{SYNTAX}\quad \underbrace{(11,158)}_{\substack{\text{CONCEPT}\\ \text{NODE 1}}}/\underbrace{(102,188,170)}_{\substack{\text{CONCEPT}\\ \text{NODE 2}}}\ \underbrace{\$\ 1,3,4,13}_{\text{FORMATS}}$$

NODE 1	NODE 2	FORMATS	SAMPLE PHRASES
11 ANAL SYNTHESIS SYNTHES	102 INTERLINGU LANGUAGE	1 	1 SYNTACTIC ANALYSIS PHRASE RELATIONS ANALYSIS OF SENTENCES
SYNTHET	170 PHRASE SENTENCE	3 	3 WE CAN ANALYZE THE LANGUAGE ...SYNTHESIZE A SYNTAX
158 CLASS CORRESPOND GROUP INDEPEND	SUBJECT WORD	4 	4 THE GRAMMAR IS NOW AVAILABLE FOR ANALYSIS
RELATE	188 GRAMMAR SYNTAX SYNTACTIC	13 	13 THIS ANALYSIS IS APPLICABLE TO RUSSIAN GRAMMAR

Criterion Phrase Specification

Fig. 7

more general ones, and to formulate a search request by starting with a general formulation, and progressively narrowing the specification down to those areas which appear to be of principal interest. Thus, one can start with a topic area such as "mathematics", and from there proceed to "algebra" which is a subdivision of mathematics, from where in turn one can go to "graph theory", which then leads to "tree structures", from where finally one can obtain the syntactic dependency trees previously illustrated in Fig. 7.

In a content analysis system, a hierarchical arrangement of words or word stems can be used both for information identification and for retrieval purposes. Thus, if a given search request is formulated in terms of "syntactic dependency trees", and it is found that not enough useful material is actually obtained, it is possible to "expand" this request to include all tree structures or indeed all abstract graphs, by using a hierarchical subject classification.

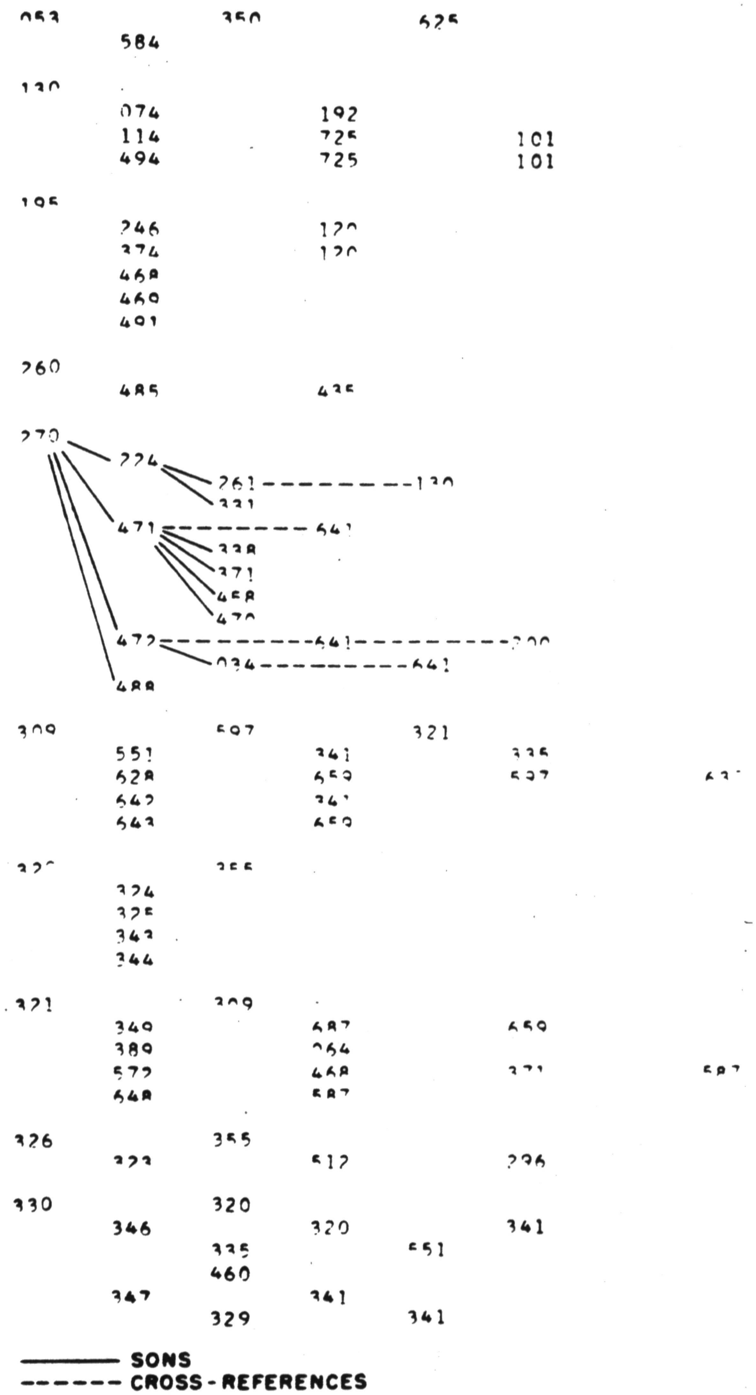
A hierarchy of concept numbers is included in the SMART system, and it is assumed that a thesaurus look-up operation precedes any hierarchical expansion operation. A typical example from the SMART concept hierarchy is shown in Fig. 8. The broad, more general concepts appear on the left side of the figure, corresponding to the "roots" of the hierarchical tree; and the more specific concepts appear further to the right. For example, concept 270 is the root of a sub-tree, this concept has four sons on the next lower level, namely concepts 224, 471, 472, and 488. Concept 224 in turn has two sons, labelled 261 and 331; similarly, concept 471 has four sons, including 338, 371, 458 and 470. It may be seen from Fig. 8, that the sons of a concept, representing more specific terms, are shown below their parents and further to the right.

The hierarchy of Fig. 8 also provides for the inclusion of cross references from one concept to another, which are connected to the original concept by broken lines. Such cross references represent general, unspecified types of relations between the corresponding concepts, and receive in general a different interpretation than the generic inclusion relations normally represented by the hierarchy.

It would be nice if it were possible to give some generally applicable algorithm for constructing hierarchical subject arrangements. This is, in fact, a topic which has preoccupied many people including mathematicians, philosophers, and librarians for many years. In general, one can say that broad concepts should be near the top of tree, whereas specific concepts should be near the bottom; furthermore there appears to be some relationship between the frequency of occurrence of a given concept in a document collection, and its place in the hierarchy. More specifically those concepts which exhibit the highest frequency of occurrence in a given document collection, and which by this very fact appear to be reasonably common, should be placed on a higher level than other concepts whose frequency of occurrence is lower.

Concerning the specific place of a given concept within the hierarchy, this should be made to depend on the user population and on the type of expansion which is most often requested. Thus, a concept corresponding to "syntactic dependency tree" would most reasonably appear under the broader category of "syntax", which in turn could appear under the general class of "language", assuming that the user population consists of linguists or grammarians; on the other hand, if the users were to be mathematicians or algebraists, then the "syntactic dependency trees" should probably appear

CONCEPT NUMBER	SEQUENCE NUMBER
53	1
584	2
-0	3
130	4
74	5
114	6
494	7
-0	8
195	9
246	10
374	11
468	12
469	13
491	14
-0	15
260	16
485	17
-0	18
270	19
224	20
261	21
331	22
471	23
338	24
371	25
458	26
470	27
472	28
34	29
488	30
-0	31
309	32
551	33
628	34
642	35
643	36
-0	37
320	38
324	39
325	40
343	41
344	42
-0	43
321	44
349	45
389	46
572	47
648	48
-0	49
326	50
323	51
-0	52
330	53
346	54
335	55
460	56
347	57
329	58



HIERARCHY EXCERPT

Fig. 8

under "abstract trees", which in turn would come under "graph theory", a branch of algebra. It does not appear reasonable to expect that a hierarchical arrangement of concepts will serve equally well for all uses under all circumstances. Rather any hierarchy will serve its function, if it can be counted upon to suggest ways of broadening or narrowing a given search request or a given interpretation of the subject matter under most of the circumstances likely to arise in practice.

4. Dictionary Performance

In order to obtain an idea of the relative effectiveness of the various dictionaries in a retrieval situation, some experimental results may be presented, based in each case on averages obtained with 17 search requests used in conjunction with a document collection of some 500 document abstracts in the computer literature. The retrieval performance is measured by two parameters, known respectively as <u>recall</u> and <u>precision</u>. Recall is defined as the proportion of relevant material actually retrieved and a high recall score therefore implies that much of what is useful in a collection has actually been produced during the search operation. Precision, on the other hand, is the proportion of retrieved material which is actually relevant, and a high precision score implies that very little useless material had been obtained as a result of a given search. Clearly both of these parameters are important, and a perfect search would therefore exhibit both a high recall and a high precision.

Recall and precision results can be presented in many different forms. One of the simplest ways in which to exhibit the performance measures is in the form of recall-precision graphs. Such graphs are obtained by looking at many recall points for each search request, and computing in each case the corresponding precision. For example, recall may be computed after retrieving five documents, and again after ten documents, and so on, in increments of five documents; in each case, the recall presumably increases, as more relevant documents are retrieved, and the precision may decrease at the same time if additional irrelevant documents are also produced. In any case, these several recall-precision points can be plotted on a curve, and the curves obtained can be averaged for many search requests. This produces the typical recall-precision graphs used in the present section.

A) The Null Thesaurus

As previously explained, the null thesaurus is used as part of a word matching, or word stem matching procedure. This dictionary can, however, be used in various different ways: for example, it is possible to apply the dictionary look-up procedure to whole documents, that is, to all word stems contained in a given document, or to only certain document excerpts such as titles or section headings; furthermore, a given sequence number from the null thesaurus can be assigned to a document specification with a uniform weight if, and only if, the corresponding word stem appears in the given document; alternatively, the sequence numbers can be weighted in such a way that the weight of a sequence number reflects the frequency of occurrences in the document of the corresponding word or word stem.

Typical results obtained with the null thesaurus are shown in Figs. 9 and 10, respectively. Fig. 9 exhibits the average output obtained by using the null thesaurus, first only for word stems occurring in the titles of the documents, and then for all word stems contained in the complete document abstracts. Fig. 10, on the other hand, illustrates the effect of the weighting procedure. In each case, a perfect result would be indicated by having both a recall and a precision of 1, which in the recall-precision graph implies a curve concentrated in the upper right-hand corner of the grid. The fact that the curves actually vary between a precision of 0.8 and 0.9 for a recall of 0.1, and a precision of 0.1 to 0.4 for a recall of 1 shows that the retrieval results were less than perfect.

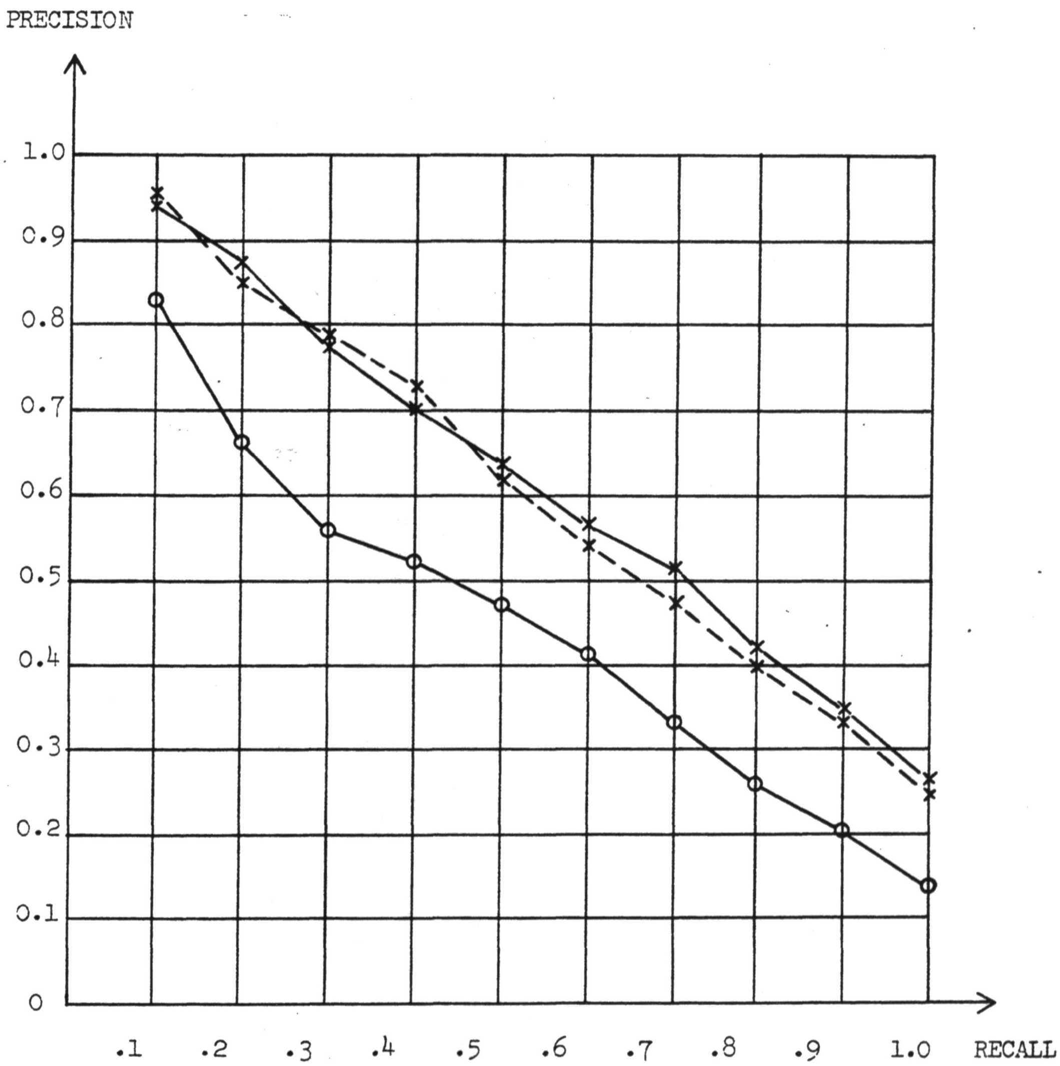
Fig. 9 indicates first of all that the null thesaurus procedure, when applied to the document titles only, performs much less well than when the thesaurus look-up is extended to complete document abstracts. Indeed the so-called "null title only" process produces a precision inferior by about 20 to 30 percent for a given recall level, compared to the other "full null" and "null title 2" processes. It is interesting to note, in this connection, that the "null title only" procedure is effectively equivalent to the use of a so-called KWIC index (keyword-in-context) which is widely advocated and used for retrieval purposes. Permuted document titles are listed in a KWIC index in such a way that a given title appears in the proper alphabetic position corresponding to <u>each</u> of the principal words contained in the title (for example, a title such as "Information Retrieval" will be listed under I for information and again under R for retrieval). It may be that a KWIC index is more useful than no index at all, but it is quite clear — as reflected in the results of Fig. 9 — that a process which takes into account only the words from document titles is not nearly as effective as an equally simple process which matches word stems from full text.

The other two curves included in Fig. 9 cover the already mentioned

Null Title Only	
0.1	0.8307
0.2	0.6800
0.3	0.5720
0.4	0.5323
0.5	0.4816
0.6	0.4142
0.7	0.3489
0.8	0.2687
0.9	0.2016
1.0	0.1463

Null Title 2	
0.1	0.9446
0.2	0.8853
0.3	0.7881
0.4	0.7049
0.5	0.6437
0.6	0.5812
0.7	0.5148
0.8	0.4262
0.9	0.3518
1.0	0.2761

Full Null	
0.1	0.9563
0.2	0.8648
0.3	0.7968
0.4	0.7381
0.5	0.6371
0.6	0.5589
0.7	0.4877
0.8	0.4086
0.9	0.3426
1.0	0.2613

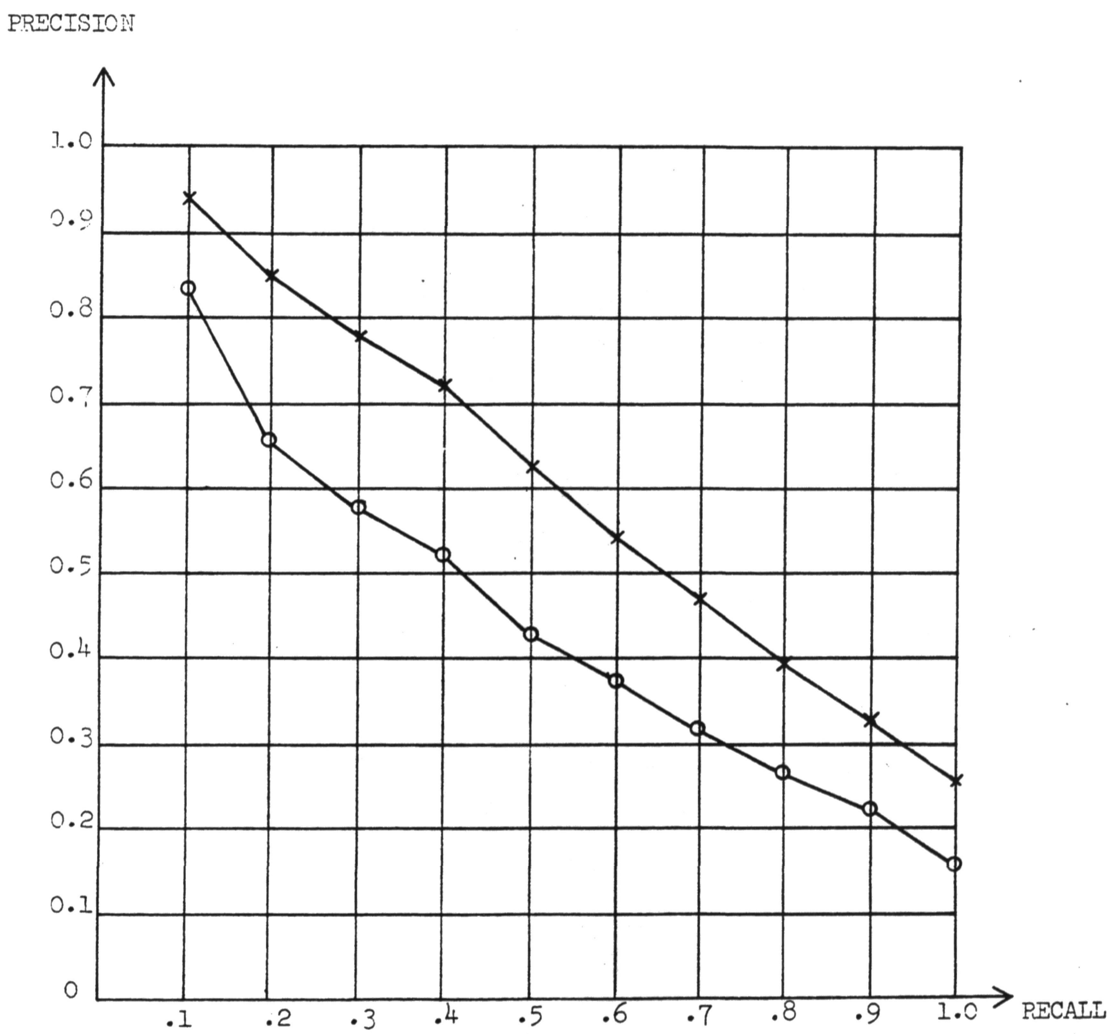


Comparison Based on Document Length
(averages over 17 search requests)

Fig. 9

Null LogVec	
0.1	0.8460
0.2	0.6841
0.3	0.5926
0.4	0.5216
0.5	0.4399
0.6	0.3897
0.7	0.3288
0.8	0.2762
0.9	0.2241
1.0	0.1643

Full Null	
0.1	0.9563
0.2	0.8648
0.3	0.7968
0.4	0.7381
0.5	0.6371
0.6	0.5589
0.7	0.4877
0.8	0.4086
0.9	0.3426
1.0	0.2613



Comparison Based on Stem Weights
(averages over 17 search requests)

Fig. 10

cases where all word stems included in the complete document abstract are matched (full null), and where all word stems are used, but stems included in document titles are weighted twice as heavily as other word stems (null title 2). As can be seen there is not much to choose between these two methods, although the increased title weights seem to perform slightly better for high recall points. It should be noted that both of the complete word matching procedures produce very high precision when the recall is low. This reflects the fact that the documents which exhibit the highest similarity with the search requests, and which therefore are retrieved early in a given search operation — assuming that documents are retrieved in decreasing order of similarity with the search requests — may be expected to be almost all relevant to the given request. Or, differently expressed, a word matching procedure will be useful if the requestor desires to see only a few documents, and does not insist on obtaining everything that is relevant within a given collection. The more sophisticated thesaurus procedures may then be expected to be useful mainly for the purpose of raising the precision for high recall values, that is, to retrieve documents which cannot be immediately obtained by a word matching process.

Fig. 10 shows that the word matching procedure which assigns weights to the stems in proportion to their frequency within a given document (full null) is much more effective than the equivalent matching process in which weights are disregarded (null logvec). The logical vector process is one where each word stem is assigned the same weight, namely 1, and no distinction is made between more and less important stems.

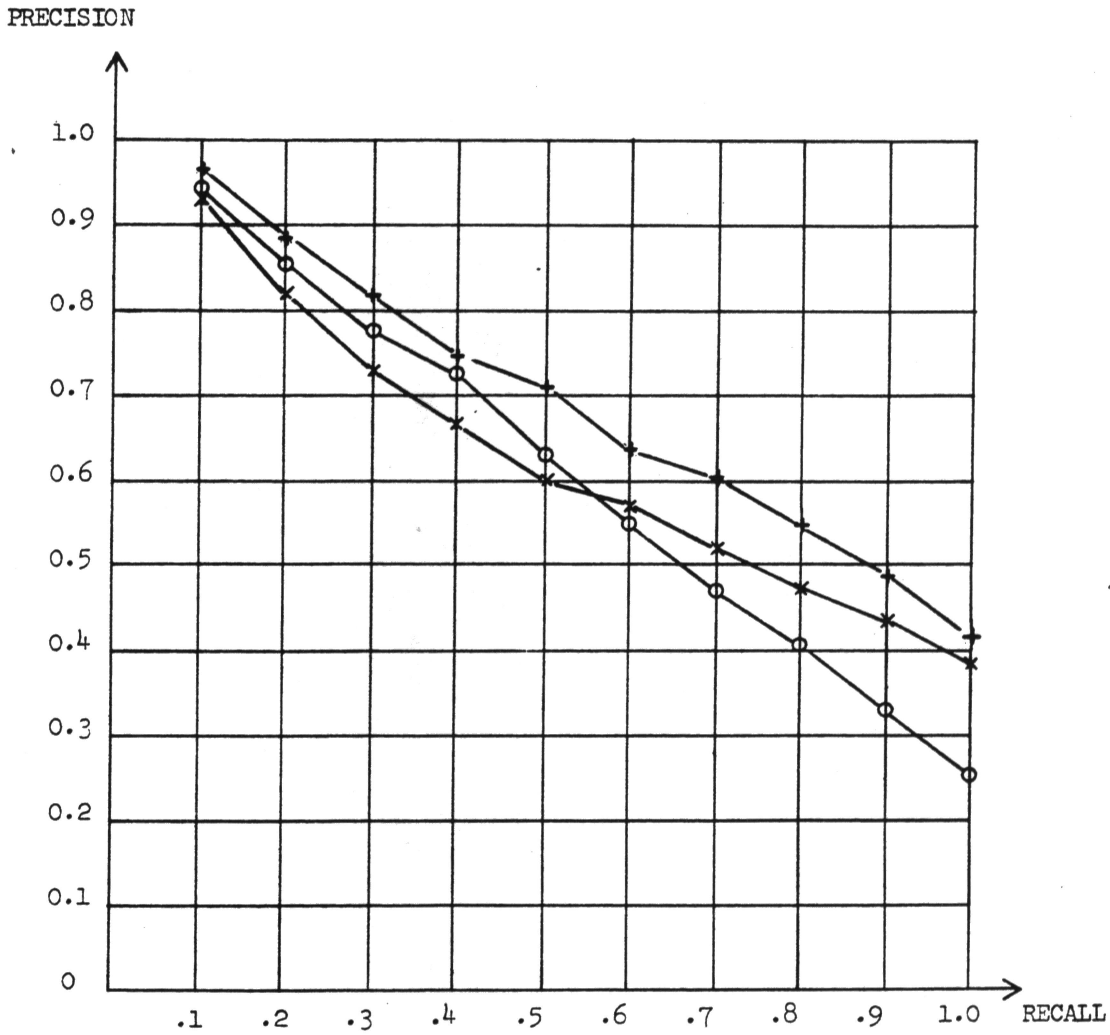
To summarize then, the word stem matching procedure performs best when all word stems are used from null document abstracts, or full documents, and when the stems are weighted in accordance with their frequency within the document. Furthermore, this process produces high precision if a less than complete recall performance is desired, because documents whose word stems match the stems present in the search requests are generally found to be useful to the requestor.

B) The Regular Thesaurus

The regular thesaurus provides synonym recognition and may therefore be expected to be useful in retrieving some documents which cannot be easily obtained by a word matching procedure alone. The results obtained with two synonym dictionaries constructed for the computer literature are shown in Fig. 11. The first dictionary, called "Harris 2", is a thesaurus constructed by hand using ad hoc methods to group the terms included in the thesaurus. The other dictionary, termed "Harris 3", was built using the thesaurus construction principles, outlined in the preceding part, which provide for the isolation of high frequency words and for the elimination of many words whose information content is unclear. Fig. 11 shows a comparison between the retrieval effectiveness of the full null thesaurus and the two regular thesauruses previously referred to.

It may be noticed first of all that the performance of the Harris 3 thesaurus is better throughout than that of the Harris 2 dictionary, thus indicating the effectiveness of the thesaurus construction procedures compared to ad hoc methods. Fig. 11 also indicates that the performance of the null dictionary degrades as the recall values become larger. Initially, the null thesaurus produces a higher precision than the Harris 2 dictionary, since false retrievals due to questionable synonyms

Full Null		Harris Two		Harris Three	
0.1	0.9563	0.1	0.9551	0.1	0.9735
0.2	0.8648	0.2	0.8242	0.2	0.8973
0.3	0.7986	0.3	0.7389	0.3	0.8245
0.4	0.7381	0.4	0.6796	0.4	0.7551
0.5	0.6371	0.5	0.6070	0.5	0.7146
0.6	0.5589	0.6	0.5702	0.6	0.6499
0.7	0.4877	0.7	0.5233	0.7	0.6012
0.8	0.4086	0.8	0.4821	0.8	0.5514
0.9	0.3426	0.9	0.4452	0.9	0.4973
1.0	0.2613	1.0	0.3951	1.0	0.4118



Comparison Based on Thesauruses
(averages over 17 search requests)

Fig. 11

included in the regular thesaurus cannot be generated by the null process. Eventually, as more documents are retrieved, the performance of the null thesaurus which offers no synonym detection at all becomes less attractive. The Harris 3 dictionary is competitive with the null dictionary for precision, but also maintains the recall advantage by careful isolation of high frequency words, and by the corresponding promotion of important low frequency words.

As an example of the performance of synonym dictionaries, consider the search result obtained with a collection on aeronautical engineering for a request whose text reads "how does scale height vary with altitude in an atmosphere". The ranked output in decreasing correlation order with the search request shown in Table II indicates that more relevant documents have low ranks (and therefore high correlation with the request) for the regular thesaurus procedure than for the null thesaurus. Moreover, the regular thesaurus has succeeded in promoting a number of relevant documents, such as documents number 617, 621, 15+ and 302. One of the promoted documents, number 621 is found to contain the sentence "variations in air density between day and night in the region 190 to 280 km are found to be small". This sentence contains no matching words with the request, and is therefore useless for a word matching procedure. The regular thesaurus, however, contains both "air" and "atmosphere" in the same concept class, thus explaining in part why the rank of document 621 improves from 14th for the null thesaurus to 4th for the regular synonym dictionary. The same type of analysis reveals that the relevant document 15+ contains a sentence reading "density data are given for the altitude range of 370 to 400 km", which is again used by the thesaurus since "altitude" and "height" are grouped in a common class.

Null Thesaurus			Regular Thesaurus			
Rank	Document	Relevant	Rank	Document	Relevant	Promoted
1	622	yes	1	622	yes	
2	616	yes	2	616	yes	
3	10C		3	617	yes	yes
4	578		4	621	yes	yes
5	619	yes	5	578		
6	617	yes	6	619	yes	
7	613		7	15+	yes	yes
8	620	yes	8	10C		
9	614		9	620	yes	
10	15+	yes	10	613		
11	719		11	614		
12	618		12	302		yes
13	436		13	618		
14	621	yes	14	436		
15	371		15	710		

Query Text: "How does scale height vary with altitude in an atmosphere ?"

Example of Thesaurus Performance

Table II

Fig. 12 does for the "Harris 3" thesaurus what Fig. 9 did for the null dictionary: specifically, it shows the effect of using the thesaurus for title words only, compared to using it throughout, and of applying higher weights to the title than to the remainder of the text. The results are substantially in agreement with those previously obtained for the null thesaurus: the "title only" process is again much poorer, indicating that synonym recognition for title words alone, while better than no synonym recognition at all, is still not nearly so effective as full synonym detection; also as before, the increased weighting of title words does not substantially add to the retrieval effectiveness.

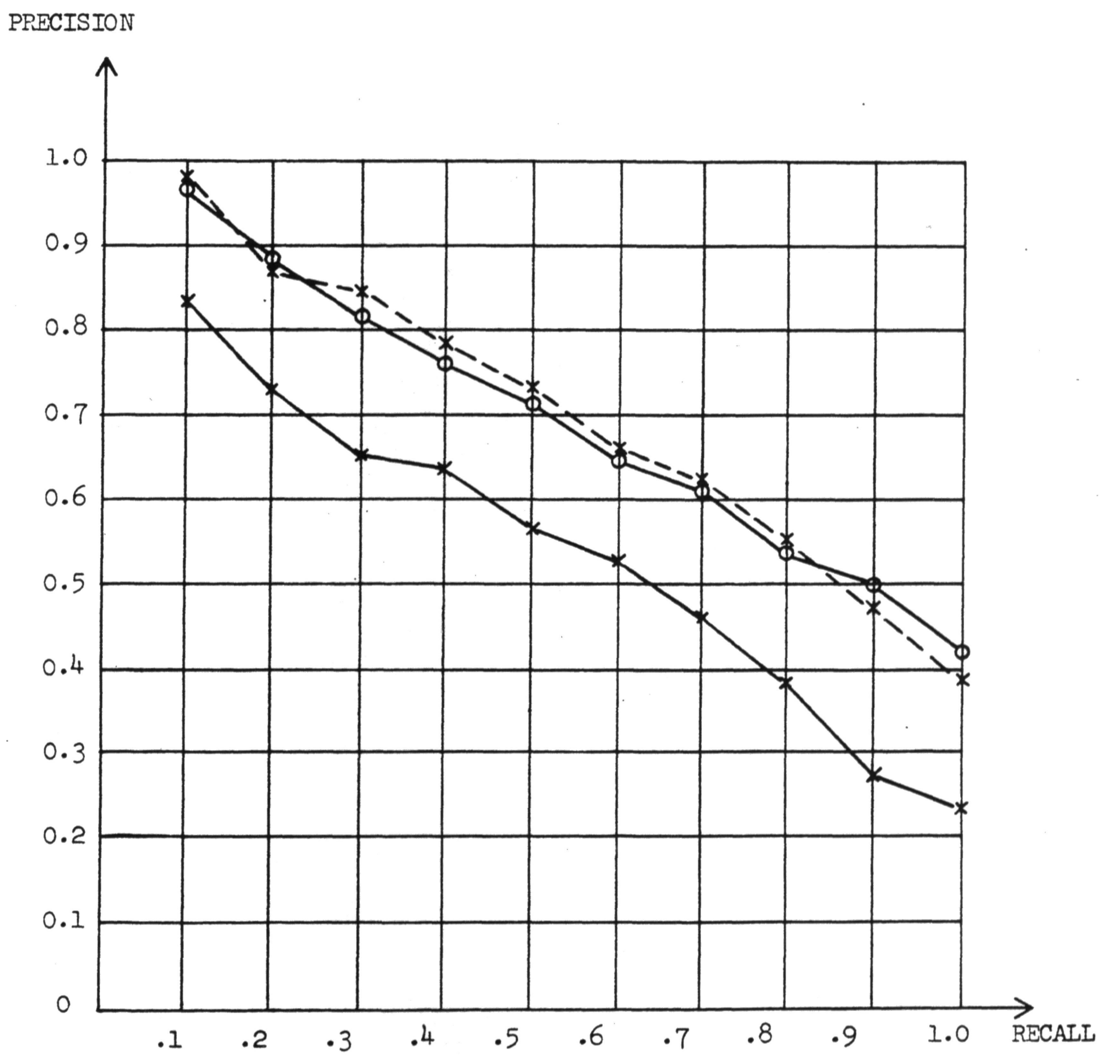
C) The Phrase Dictionary

The performance of the statistical phrase dictionary may be evaluated by using the output of Figs. 13 and 14. Fig. 13 presents a comparison between the early "Harris 2" thesaurus, and the same thesaurus supplemented by statistical phrases of equal weight. The same procedures are compared in Fig. 14 for the more powerful "Harris 3" thesaurus. Fig. 14 also includes performance figures for two combined searches consisting first of the regular thesaurus look-up followed by a statistical phrase look-up, in which phrases are weighted one and a half times as much as individual concepts.

Fig. 13 shows that the statistical phrase process affords a noticeable improvement in retrieval effectiveness, compared with the "Harris 2" thesaurus alone; a much smaller improvement is obtained over "Harris 3", as seen in Fig. 14. The third dictionary includes fewer ambiguities, thus explaining why the phrase process is less important in this case.

For both synonym dictionaries it may be noticed that for very high

Harris Three		H3 Title Only		H3 Title 2	
0.1	0.9735	0.1	0.8437	0.1	0.9804
0.2	0.8963	0.2	0.7436	0.2	0.8953
0.3	0.8189	0.3	0.6733	0.3	0.8535
0.4	0.7782	0.4	0.6547	0.4	0.7907
0.5	0.7137	0.5	0.5828	0.5	0.7324
0.6	0.6517	0.6	0.5328	0.6	0.6570
0.7	0.6102	0.7	0.4739	0.7	0.6154
0.8	0.5492	0.8	0.3925	0.8	0.5579
0.9	0.5002	0.9	0.2874	0.9	0.4855
1.0	0.4201	1.0	0.2320	1.0	0.3969



Comparison Based on Dictionary with Title Weights
(averages over 17 search requests)

Fig. 12

precision, the dictionary without phrases is preferable. This result reflects the feeling, already expressed in connection with the null thesaurus, that the first few documents are best retrieved by the simplest possible methods, when the chances of erroneous analysis are smallest. The statistical phrase procedure, as well as the regular thesaurus look-up, may always generate an occasional concept which is in error. Such concepts may affect the retrieval results, thus depressing precision. On the other hand, the increasingly more sophisticated text analysis which becomes possible through the phrase detection procedure is undoubtedly responsible for retrieving at least some documents which cannot be brought to the surface by other simpler methods. This accounts for the beneficial effect of all well-built dictionaries in improving the recall performance, usually at a loss in precision.*

The observed usefulness of synonym and phrase dictionaries raises the important question of how such dictionaries are best prepared. This question is examined in more detail in the next part.

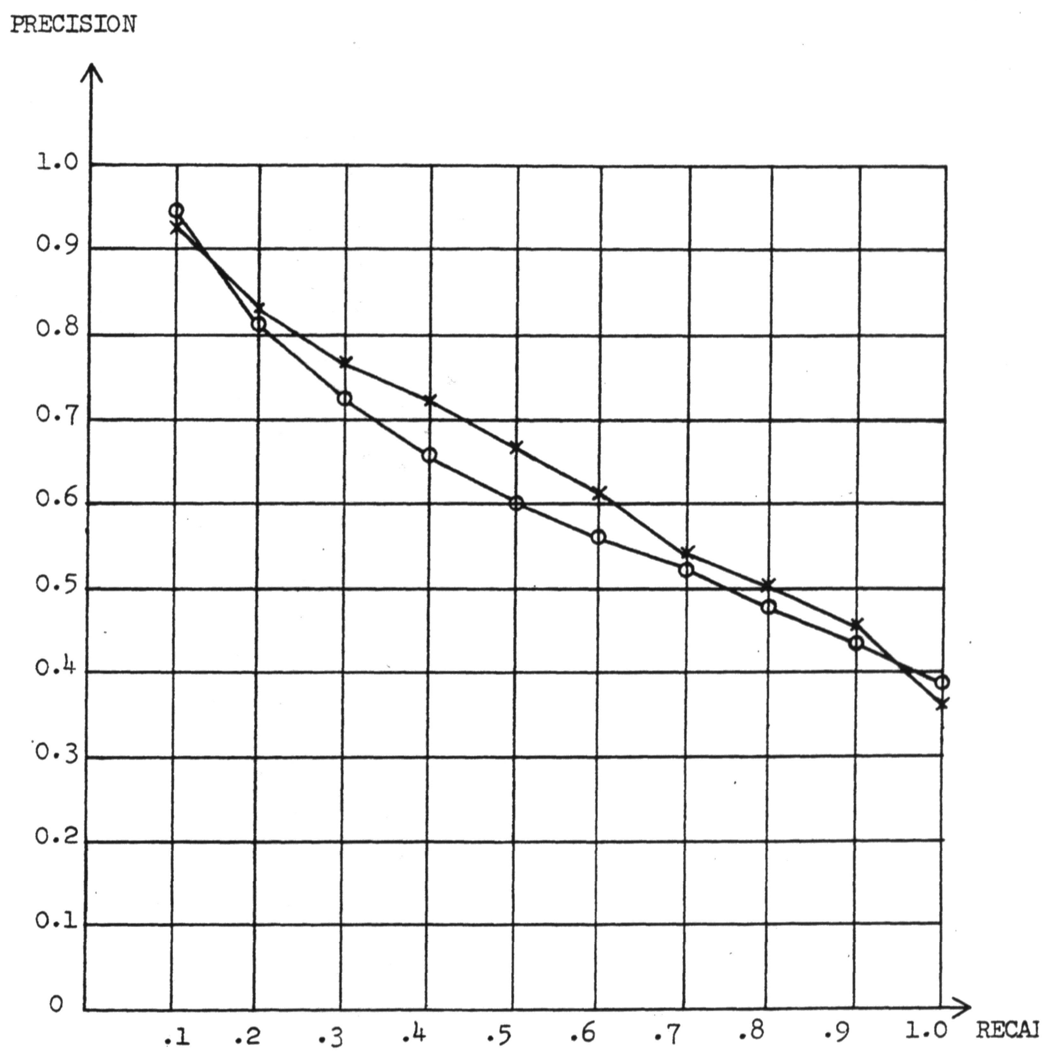
5. Automatic Thesaurus Construction

Under normal circumstances, the task of constructing a subject dictionary for a given topic area is one which demands many skills, including also a great deal of persistence and tenacity. It is not usually enough to be a subject expert in a given area, but training is also normally expected in linguistics and philosophy. Furthermore, since the task is of large proportions, a committee is often appointed which thrashes out controversial questions and eventually produces a suggested standard

---

* The search results exhibited in this report for documents and dictionaries in the computer literature have been confirmed for other subject areas, including aeronautical engineering and documentation, also processed with the SMART programs.

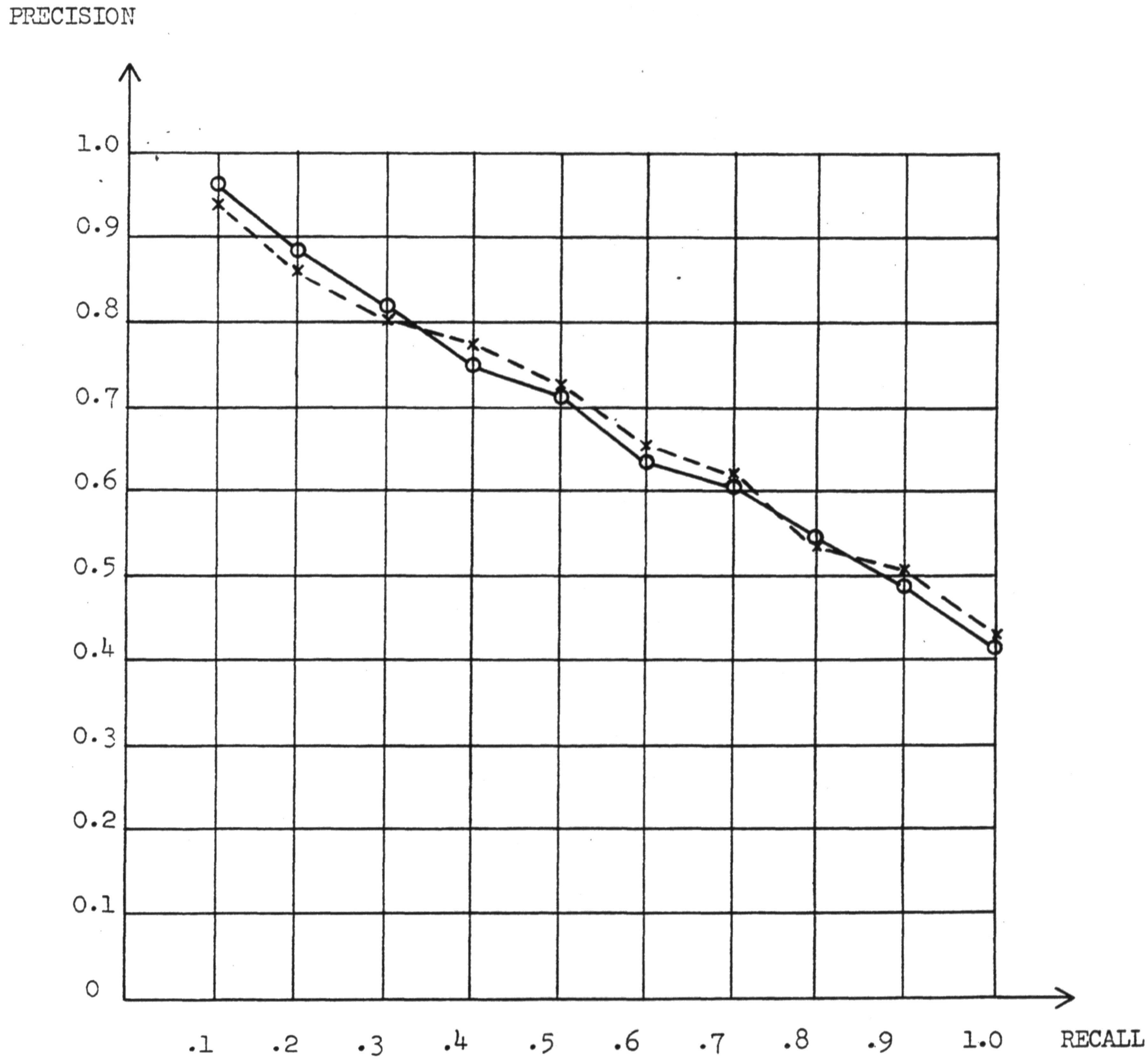
Harris Two		H2 Stat. 1	
0.1	0.9551	0.1	0.9471
0.2	0.8242	0.2	0.8372
0.3	0.7398	0.3	0.7786
0.4	0.6796	0.4	0.7242
0.5	0.6070	0.5	0.6717
0.6	0.5702	0.6	0.6182
0.7	0.5233	0.7	0.5464
0.8	0.4821	0.8	0.5034
0.9	0.4452	0.9	0.4670
1.0	0.3951	1.0	0.3704



Comparison Based on Phrase Dictionary (Harris 2)
(averages over 17 search requests)

Fig. 13

Harris Three		H3 Stat. 1		Harris Three H3 Stat. 1.5	
0.1	0.9735	0.1	0.9588	0.1	0.9735
0.2	0.8973	0.2	0.8706	0.2	0.8963
0.3	0.8245	0.3	0.8169	0.3	0.8189
0.4	0.7551	0.4	0.7836	0.4	0.7782
0.5	0.7146	0.5	0.7205	0.5	0.7137
0.6	0.6499	0.6	0.6526	0.6	0.6517
0.7	0.6012	0.7	0.6152	0.7	0.6102
0.8	0.5514	0.8	0.5510	0.8	0.5492
0.9	0.4973	0.9	0.5035	0.9	0.5002
1.0	0.4118	1.0	0.4213	1.0	0.4201



Comparison Based on Phrase Dictionary (Harris 3)
(averages over 17 search requests)

Fig. 14

dictionary. Such a committee produced standard frequently ends by satisfying no one, despite the enormous effort which goes into its construction.

Clearly, if it were necessary to follow this particular pattern in order to build a useful dictionary for retrieval purposes, then any saving which might result from automatic search and retrieval methodology would promptly be lost through the elaborate preparations required to build dictionaries.

This situation has led to many efforts calculated to produce dictionaries either fully-automatically, or in any case by more systematic procedures than a committee-controlled process. Any reasonably standardized method for dictionary construction not only saves time and decreases costs, but also permits a great deal more latitude in the type of retrieval procedures which can be implemented. The following principal advantages are evident:

1) the retrieval procedures can be extended to collections in many different areas, since the dictionary problem no longer constitutes an impediment;

2) it becomes possible to investigate differences in vocabulary between different subject areas, notably the frequently heard assertion that the vocabulary in some subject areas is "soft" (that is, not well standardized and ambiguous), whereas in other areas it is "hard";

3) it removes any possible differences in retrieval effectiveness between different subject areas due to disturbances introduced by varying methods of thesaurus construction;

4) it becomes possible to investigate the retrieval effectiveness of a variety of thesauruses for a given collection, including variations in the thesaurus size, in the number of concept classes, and in the correspondents assigned to each class.

No matter what particular method of thesaurus construction is adopted, the main virtue of an automatic process is to eliminate the human element, either completely if a fully-automatic method can be found, or partially if the process is semi-automatic. In the latter case, it is desirable to restrict the human activities to questions which require only local decisions within the given subject area, rather than global considerations involving linguistic knowledge, and experience in subject classification and indexing.

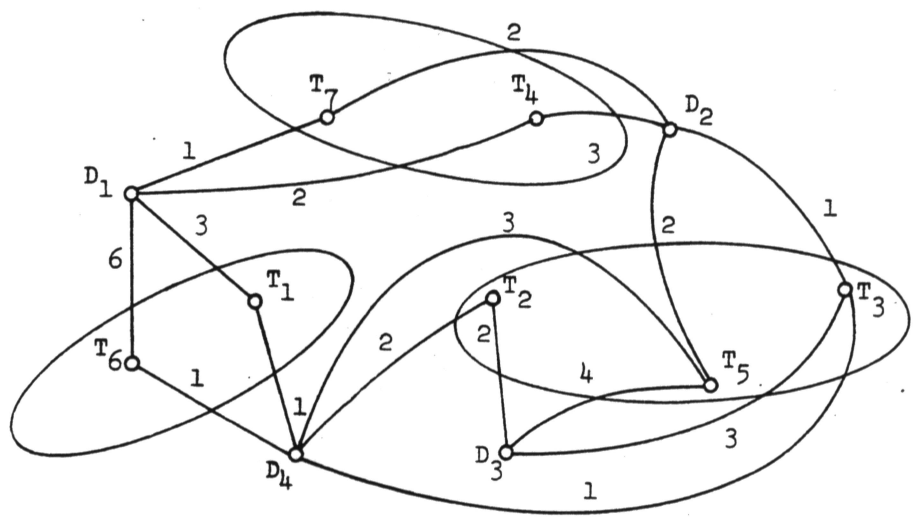
Some systematic procedures for thesaurus construction are described in the next few paragraphs, and a simplified example is given of one particular semi-automatic process.

A) Fully Automatic Methods

Most automatic methods for thesaurus construction are based on the vocabulary contained in a sample document collection assumed to be typical for a given subject area.[4,5,6] In particular, a frequency count is made of the words contained in a set of documents, and each document is identified by certain high frequency words included in it. The choice of these words may be based strictly on frequency characteristics, or alternatively on more complicated properties of the word distribution for the given collection. In any case, the sample collection is initially represented by a term-document matrix, or a term-document graph as shown in Fig. 15. The matrix element at the intersection of row i and column j of the matrix represents the weight of term j in document i ; this same weight is represented in the graph of Fig. 15 (b) by the labelled branch between nodes T_j and D_i.

		terms assigned to documents							
		T_1	T_2	T_3	T_4	T_5	T_6	T_7	....
document vectors	D_1	3	0	0	2	0	6	1	
	D_2	0	0	1	3	2	0	2	
	D_3	0	2	3	0	4	0	0	
	D_4	1	2	1	0	3	1	0	

(a) Term-Document Matrix Showing Frequency of Terms Assigned to Documents



(b) Term-Document Graph for Matrix of Fig. 15 (a)

Term-Document Graphs and Matrices

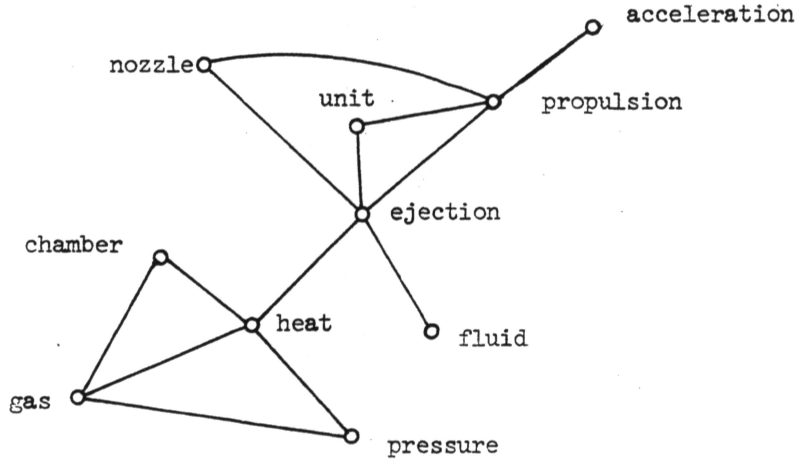
Fig. 15

Given such a term-document matrix or graph, it is now possible, by well-known statistical association methods, to compute similarity coefficients between terms, based on co-occurrence characteristics of the terms in the documents of the collection. The similarity coefficient between each pair of terms can then be made to depend on the frequency with which the terms are jointly assigned to the documents of a collection. In Fig. 15, for example, it may be noted that terms T_1 and T_6 are both assigned to documents D_1 and D_4 (although with differing weights), while they are both not assigned to documents D_2 and D_3. As a result, the term association process may assign these two terms to a common thesaurus category.

For the example of Fig. 15 an associative procedure might result in the formation of three term (thesaurus) groups, consisting respectively of terms T_1 and T_6 (because of joint assignment to documents D_1 and D_4), terms T_7 and T_4 (because of joint assignment to D_1 and D_2), and finally terms T_2, T_3 and T_5 (because of joint assignment to D_3 and D_4). The result of a term association process may then be displayed as an association map, in which branches between terms represent term relations, or, alternatively, thesaurus groupings. An excerpt from a typical term association map is shown in Fig. 16.[4,7,8] The thesaurus groupings suggested by the map of Fig. 16 can be found by inspection.

B) Semi-Automatic Methods

The methods outlined in the preceding part are based on the assumption that term co-occurrences in documents, or joint assignment of terms to documents are indicative of term similarity or relatedness. This assumption



Excerpt from Word Association Map

Fig. 16

may not always hold, and if it holds, its applicability may be restricted to a given document collection rather than to a complete subject field. For this reason, it is of interest to consider also somewhat less radical procedures which avail themselves of a certain amount of human judgment. These methods are generally based on various automatic aids, but use subject experts for the basic task of defining the meaning of each term being introduced into the thesaurus.[9,10,11,12]

The basic idea is to start with a word frequency list, as before, for the words included in a given document collection. In addition, it is also useful to have available a listing which exhibits the words in context, so that a distinction may be made between individual word-uses for ambiguous terms. For example, a word such as "base" may be broken down into "$base_1$", "$base_2$",and "$base_3$", to represent, respectively "army base", "lamp base", and "baseball base" (assuming that those three uses of the term are in fact present in a given collection). A standard "keyword-in-context" (KWIC) list may be prepared automatically, to permit a human observer to ascertain the individual word-uses for the terms included in a collection. An example of a typical KWIC index list, used in conjunction with the SMART system is shown in Fig. 17.[13]

Fig. 17 shows that the term "spectral" is used in the given collection in only one sense, namely that of a "spectral norm"; the term "square" is, however, used in two senses in the concordance excerpt, first as a rectangle of equal sides (square matrix), and then as a power of two (square root). The list of word-uses to be constructed would then include a single instance of the term "spectral", but two separate examples of "square".

```
                                                                                                   DOCUMENT
                                                 SPECIFIED                                          NUMBER
BY ANY SIPPLE EXTENSICN CF THE PETHCDS USEC FCR CCMPLETELY SPECIFIED FUNCTICNS . AN ANALYSIS CF THE PRCBLEP IS PRESENTE     90
 NUPBER CF CCCURRENCES 2

                                                 SPECTRAL
NA PERGING TECHNIGUE . SEVERAL EXAPPLES ARE GIVEN . ON THE SPECTRAL NORPS CF SEVERAL ITERATIVE PRCCESSES                   178
RPS CF SEVERAL ITERATIVE PRCCESSES                   . THE SPECTRAL NORP OF A SQUARE SYPPETRIC PCSITIVE CEFINITE MATRIX    179
TIVE PETHCD CF PATRIX INVERSICN IS CERIVEC IN TERMS OF THE SPECTRAL NORP . VARICUS THECREPS CCNCERNING THE SPECTRAL NOR    182
RPS OF THE SPECTRAL NCRP . VARICUS THECREMS CONCERNING THE SPECTRAL NORP ARE PRCVED THE RESULTS OBTAINEC ARE APPLIEC TO    183
 NUMBER CF CCCURRENCES 4

                                                 SPLITTING
TORED INITIALLY CN THE P+IST TAPE . AN ITERATIVE SCHEPE OF SPLITTING BLCCKS CF DATA WHOSE DESIGNATION HAS THE HIGHEST N     24
NO WRITING IN THE REVERSE CIRECTICN ANY GRCUP REQUIREC FOR SPLITTING IS ALWAYS UNDER THE HEAD CF THE APPRCPRIATE TAPE U     29
 NUMBER CF CCCURRENCES 2

                                                 SQUARE
ATIVE PRCCESSES                    . THE SPECTRAL NORP OF A SQUARE SYMMETRIC PCSITIVE DEFINITE PATRIX IS DEFINEC AS THE    179
METRIC PCSITIVE DEFINITE PATRIX IS CEFINEC AS THE POSITIVE SQUARE ROOT OF THE PAXIPUP PAGNITUCE CF ITS CHARACTERISTIC R    180
ORGANIZE SUCH A CALCULATICN IS FLCWCHARTEC . CN TAKING THE SQUARE ROOT OF A CCPPLEX NUPBER                        . A    214
TC CANCELLATICN RESULTING FRCP A SUBTRACTICN IN TAKING THE SQUARE RCOT CF A CCPPLEX NUPBER IS NCTEC . CNE SOLUTION IS T    216
LEX NUMBER IS NCTED . CNE SCLUTION IS TC TAKE INTERMECIATE SQUARE ROOTS TO DCUBLE LENGTH ACCURACY ANCTHER IS TC FINC TH    217
 NUMBER CF CCCURRENCES 5

                                                 ST
ESCRIBED . THE UNSCRTED CATA IS STCREC INITIALLY ON THE P+IST TAPE . AN ITERATIVE SCHEPE CF SPLITTING BLCCKS OF CATA WH     24
 NUMBER CF CCCURRENCES 1

                                                 STAGE
DYNAMIC PRCGRAMMING SHCW HCW TC PRCCEEC CPTIMALLY FROM ONE STAGE TO THE NEXT WITH THE NUMBER CF CCPPUTATICNS INCREASING    263
 NUMBER CF CCCURRENCES 1

                                                 STAGES
F CCMPUTATICNS INCREASING CNLY LINEARLY WITH THE NUMBER OF STAGES CONSIDERED . THE GENERAL PRINCIPLES ARE ILLUSTRATEC B    263
 NUMBER CF CCCURRENCES 1

                                                 STARTING
QUENTIAL PACHINES                                        . STARTING FROP PEALY S PCDEL CF A SECUENTIAL PACHINE A CONNEC    144
 NUMBER CF CCCURRENCES 1

                                                 STATE
PBERS ARE SUGGESTEC FCR FUTURE CATA PRCCESSING COMPUTERS . STATE LOGIC RELATICNS IN AUTCNOPCUS SEQUENTIAL NETWCRKS          51
SSED . THE RELATICNSHIP BETWEEN THE INTERNAL LCGIC ANC THE STATE SEQUENTIAL BEHAVICR OF SUCH NETWORKS IS EXAMINEC THROU     53
CNDITIONS FCR THE NETWCRK TC BE NCNSINGULAR I.E. TO HAVE A STATE DIAGRAP WHICH IS DETERMINISTIC EVEN IN REVERSEC TIME A     56
SIGULARITY CCNDITICN ARE CEPCNSTRATEC . THE EFFECTS CN THE STATE DIAGRAM OF SEVERAL KINDS CF CCNSTRAINTS IMPCSEC CN THE    58
 NUMBER CF CCCURRENCES 4

                                                 STATES
FUNCTICNS IS TAKEN INTC ACCCUNT . MINIMIZING THE NUMBER OF STATES IN INCCPPLETELY SPECIFIED SECUENTIAL SWITCHING FUNCTI     84
LY DESCRIBES THE PACHINE IS CEVELCPEC . THE ECUIVALENCE OF STATES OF A SEGUENTIAL PACHINE IS ANALYZEC SYSTEMATICALLY BY    146
 NUMBER CF CCCURRENCES 2

                                                 STATISTICAL
ATICN CF AN ADDRESS FUNCTICN IS CESCRIBEC . USUALLY ONLY A STATISTICAL APPRCXIPATICN TC THE ACCRESS FUNCTION IS KNOWN .      9
 NUMBER CF CCCURRENCES 1
```

CONCORDANCE EXCERPT

Fig. 17

After the list of word-uses to be included in the thesaurus is available, it becomes necessary to group them into thesaurus classes. This can be done in various ways:

1) an informal judgment can be made for each pair of word-uses to decide whether in the subject area under consideration, they are synonymous, and if so, they can be grouped in the same thesaurus class;

2) a set of "syntactic frames" can be used, and those word-uses which fit into the same frames can be collected in the same thesaurus group, or, equivalently, a decision is made based on whether term A can always replace term B in a given context X.[9] This decision is of course not mechanized, but the dictionary maker is faced only with local choices within certain narrow limits;

3) a set of questions can be prepared designed to elicit answers about the terms to be grouped, and each term can be identified by the set of answers obtained in response to the proposed questions; for example, one might ask "does this term represent a physical object or process, or does it represent an abstraction, or is this question inapplicable"; a score of 1 may then be assigned for a physical object, 2 for an abstraction, and 3 if the question is not applicable.

At the end of such a procedure, each term is then identified by a set of properties (in the form of contexts which fit a given term, or in the form of answers to questions about the terms), and the complete vocabulary may be represented by a property matrix, as shown in simplified form in Fig. 18. It remains, then, to find the semantic distance between terms by comparing the rows of properties representing the respective word-uses.

Specifically, rows which are completely identical can be coalesced into a single group immediately; terms which are not identical may be

$$
\begin{array}{ll|cccccc}
 & & \multicolumn{6}{c}{\overbrace{\text{properties identifying word uses}}} \\
 & & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
\hline
\left.\begin{array}{c}\text{word-uses}\\ \text{obtained}\\ \text{from collection}\end{array}\right\{ & T_1 & 1 & 0 & 0 & 2 & 1 & 0 \\
 & T_2 & 0 & 1 & 0 & 1 & 0 & 1 \\
 & T_3 & 2 & 0 & 1 & 1 & 2 & 0 \\
 & T_4 & 1 & 2 & 0 & 1 & 0 & 1 \\
 & \vdots & & & & & & \\
\end{array}
$$

0 property inapplicable

1 property applies somewhat

2 property applies strongly

Typical Term-Property Matrix

Fig. 18

for a total frequency of n/N, assuming that classes of approximately equal frequency are wanted. The process of generating N classes from P initial property sets may now be carried out as follows:

1) a $P \times M$ word-use versus property matrix (similar to that shown in Fig. 18) is constructed;
2) the property vectors are sorted into numeric order, and the set of P property vectors is reduced to only the distinct property vectors, say $Q_1 \leq P$;
3) since each of the Q_1 distinct vectors is to account for a word-use frequency of n/N, each vector is examined to see whether the total frequency represented by that vector is approximately n/N;
4) if a given concept vector occurs with a frequency smaller than n/N, it represents too small a class and should be combined with other vectors; this is done by deleting a sufficient number of questions (columns of the property matrix) to obtain a resulting combined concept class of frequency approximately equal to n/N; let the number of distinct property vectors which result be equal to $Q_2 < Q_1$;
5) some property vectors account for too large a frequency count, and ought to be broken up by using the concordance to formulate additional questions so as to resolve finer shades of meaning; this eventually produces Q_3 distinct vectors $(Q_3 > Q_2)$;
6) by alternately using the procedures of parts 4) and 5), the frequency count of each of $Q_i = N$ vectors eventually may approach n/N, at which point the process terminates.

Consider, as an example, the list of word-uses shown in Fig. 19 (a), accounting for a total frequency count of 2198 word instances, and assume that it is desired to form a thesaurus with 5 concept classes. Each concept vector should then cover approximately $2200/5 = 440$ word

Original Word-Uses	Frequency in corpus
computer	508
system	263
digital	186
operate	139
circuit	130
program	127
machine	124
generate	121
function	112
design	106
equation	102
logic	98
memory	94
data	88
	2198

(a) Original List of Available Word-Uses

Fig. 19

occurrences.

After applying the three questions of Fig. 19 (b) to the original corpus, one obtains the set of property vectors shown in Fig. 19 (c). After ordering the property sets in increasing numeric order, and combining the word-uses with identical property vectors, a reduced property matrix is obtained, as shown in Fig. 19 (d). This matrix contains 9 property vectors instead of the desired 5.

In order to reduce the number of vectors, the class with the smallest frequency count is examined (consisting of the term "logic" with a frequency of 98 instead of the desired 440). The elimination of question B will not avail, since the reduced property vector (3,2) does still not combine with any other row. Eliminating question A, however, produces the reduced matrix of Fig. 19 (e), consisting of five classes with frequencies varying between 288 and 632, close enough to the desired value to terminate the process.

Whether the suggested process is always manageable remains to be seen; however, in view of the obvious simplifications involved, and the need for context-limited local decisions only, it seems worthwhile to attempt an implementation in an operational situation.

6. Semi-Automatic Hierarchy Formation

The need for a hierarchical arrangement of terms, or concept classes, as part of an information retrieval system is by no means obvious, although it is easy to find useful applications for a well-constructed hierarchy, particularly when search strategies are considered which are designed to proceed from more general to more specific search formulations or vice-versa.

Question Number	Formulation
A.	Is this word used in connection with computer design and construction, or rather in connection with computer use and programming ? 1. Construction and design 2. Use and programming 3. Both of the above 4. Does not apply
B.	Does this word refer to a physical object or to an abstraction ? 1. Real, physical object 2. Abstraction or process 3. Does not apply
C.	Does the use of this word require that the object of discussion be multiple, rather than single; or, equivalently, does it imply interconnections of some sort ? 1. Subject may be single 2. Multiplicity is implied 3. Does not apply

(b) Multiple Choice Questions Applied to Words of Figure

Fig. 19 (continued)

Word-Uses	Frequency	Questions A	B	C
computer	508	3	1	1
system	263	1	1	2
digital	186	3	3	2
operate	139	2	2	1
circuit	130	1	1	2
program	127	2	2	2
machine	124	3	1	1
generate	121	2	2	1
function	112	4	2	1
design	106	1	2	2
equation	102	4	2	3
logic	98	3	2	2
memory	94	1	1	2
data	88	2	2	2

(c) Original Set of Property Vectors

A	B	C	Frequency	Components
1	1	2	487	system, circuit, memory
1	2	2	106	design
2	2	1	260	operate, generate
2	2	2	215	program, data
3	1	1	632	computer, machine
3	2	2	98	logic
3	2	3	186	digital
4	2	1	112	function
4	2	3	102	equation

(d) Ordered Property Vectors

Fig. 19 (continued)

Questions B	Questions C	Frequency	Components
1	1	632	computer, machine
1	2	487	system, circuit, memory
2	1	372	operate, generate, function
2	2	419	logic, program, design, data
2	3	288	digital, equation

(e) Reduced Classes after Elimination of Question A

Fig. 19 (continued)

It has been remarked in this connection, that when words, or word-uses, of unequal frequency are included in a thesaurus, or represented on an association map of the type shown in Fig. 16, a hierarchical arrangement results almost inevitably, since frequent words can be made into categories, and words of lesser frequency into subcategories.[4] Hierarchical association maps have in fact been constructed, using the frequency characteristics of the words as a criterion.[15] In any case, no matter what procedure is actually adopted, it would seem that a useful hierarchy which places general concepts near the top of the tree, and specific ones near the bottom, must exhibit the expected frequency characteristics which generally hold between broad and specific terms.

Since the construction of a complete hierarchy without any guidelines is at the least a thankless task, and at worst an impossible one, methods must be investigated to generate hierarchical arrangements semi-automatically. Three different procedures are outlined, all of which are based on a term-property matrix of the type shown in Fig. 18, or a term-document matrix as shown in Fig. 15 (a).

The first process directly uses the questions also used for thesaurus construction, and breaks down the initial vocabulary as a function of the responses elicited. An initial question is asked first, and classes of word-uses are formed based on the responses to this question; the next question is then applied to each of the resulting word classes which are thereby broken down again, and so on, until the subdivision is sufficiently fine.

The process is applied to the vocabulary of Fig. 19 (a) in conjunction with the questions of Fig. 19 (b). The resulting hierarchy is shown in Fig. 20, which shows the word-use frequency attached to each node.

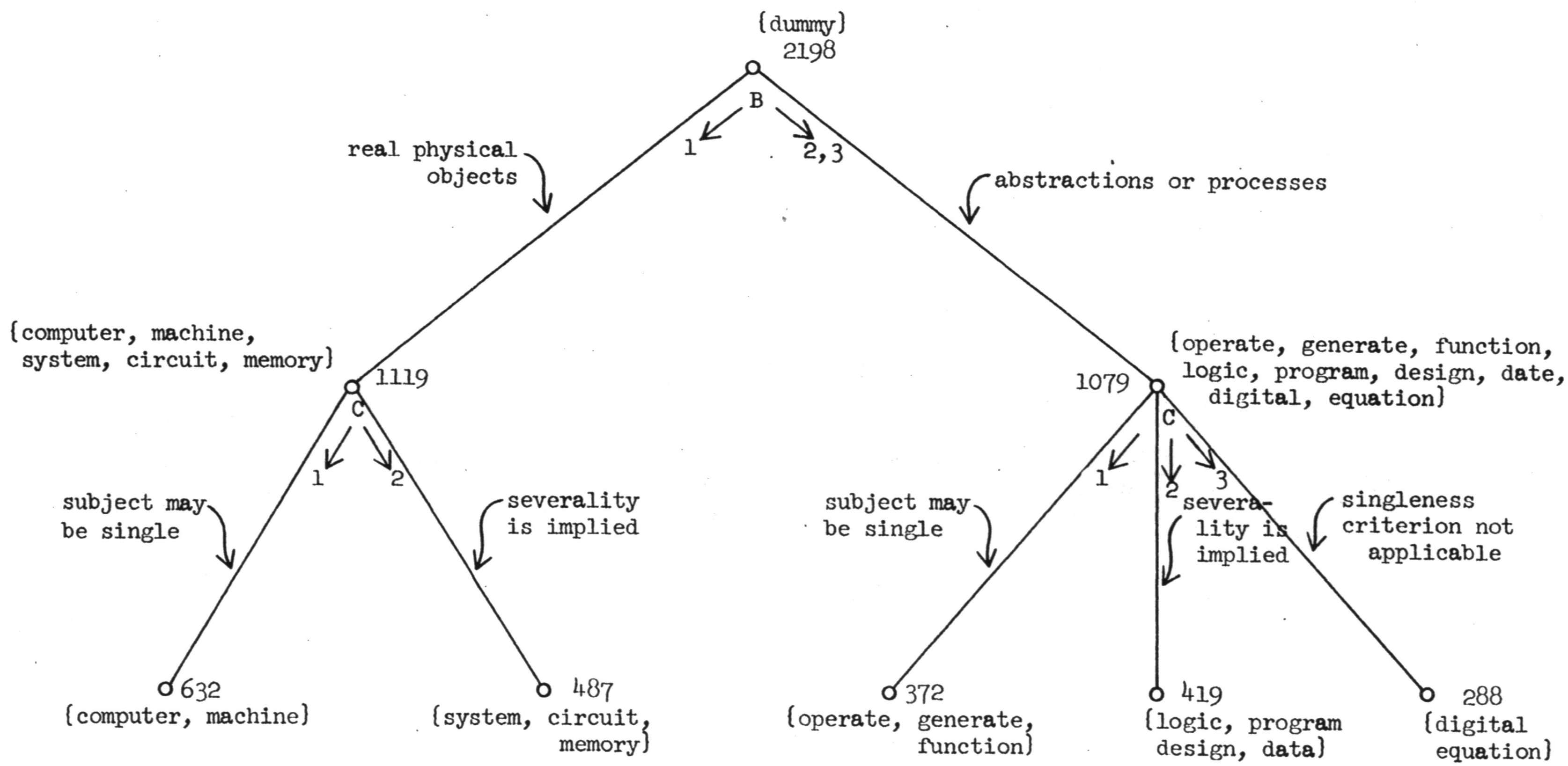
Question B is first applied to the complete vocabulary, thus forming two groups of "physical objects" and "abstractions or processes", with a frequency of 1119 and 1079, respectively. Question C is then used to furnish the five classes already shown in Fig. 20.[14]

A somewhat different process operates directly from the word-use frequencies, and is therefore not based on the thesaurus groupings as is the previous method. Instead, the hierarchy is constructed first, and the thesaurus is later based on the previously available hierarchy. A start is made as before, with a concordance and a word frequency list, and the word-uses are selected for inclusion in the hierarchy. The two-way hierarchy is now started by choosing the word-use with highest frequency, say word T_i, and letting one node represent word T_i plus all words like it, the second branch representing all "other" words not related to T_i. The word group of highest total frequency is now chosen, and its high frequency word is again used as a criterion for partitioning; this procedure continues until all word groups are small enough to be entered as concept classes into the thesaurus.

At each point in the partitioning process the following local decisions must be made;

1) the highest frequency word in the high frequency word group is chosen, and it is used as the "central" word of the subbranch; the other words in the same word group are then examined to see if they fall into the same subbranch by being related in one way or another to the central word; no relations need exist among the words which form the "other", unrelated class;

2) if a given word cannot properly be placed in one of the two categories (either related to the central word, or unrelated), it is left at the present level as a parent of the words in the



Hierarchy Construction by Property Separation
(word-sue frequencies are shown)

Fig. 20

subbranches;

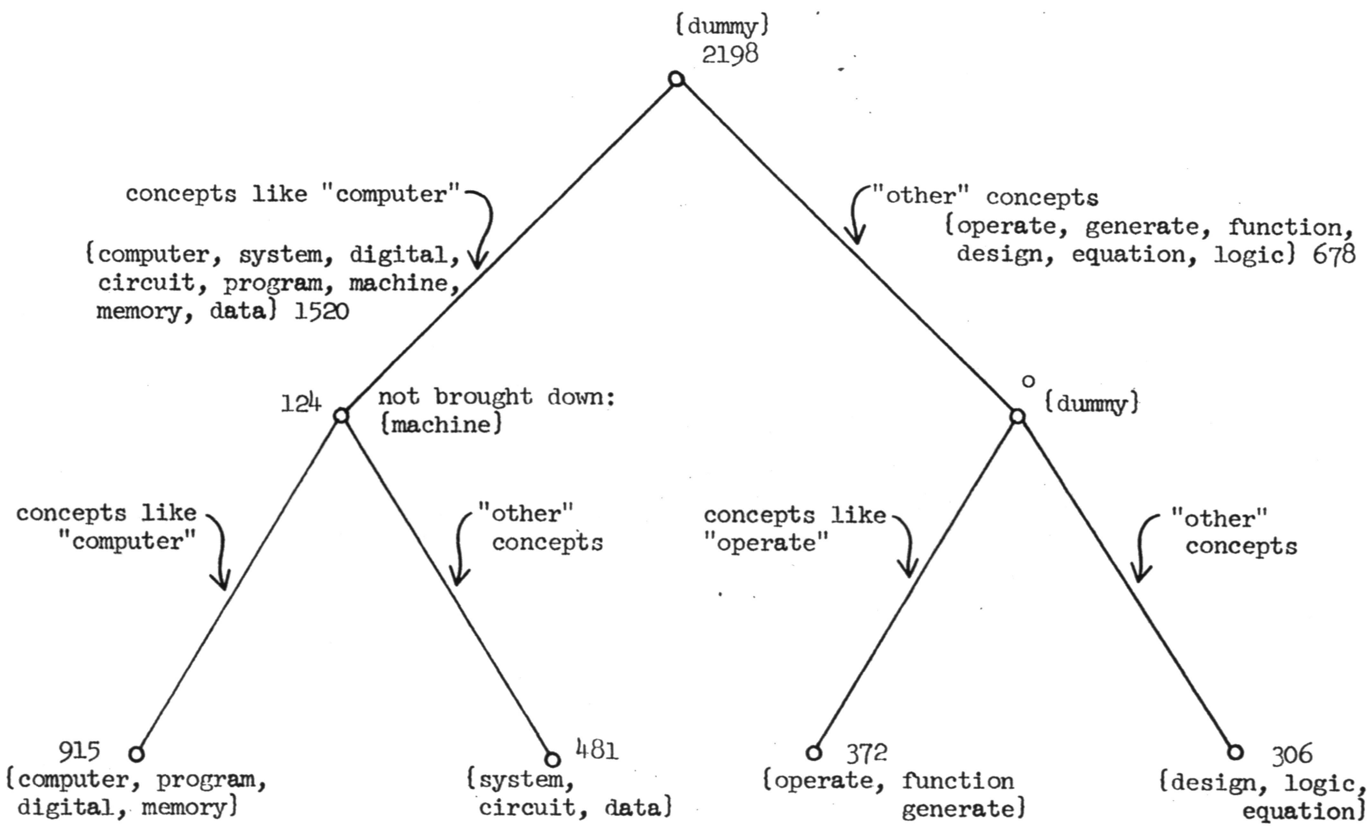
3) if all words in a given word group are being placed in the same branch with the high-frequency word, this word belongs one level up as a parent of all the remaining words.

Consider again the vocabulary of Fig. 19. The highest frequency word is "computer" (frequency 508), and two classes are first formed of words like "computer", and of the "other" words (see Fig. 21). The high frequency class is the one containing the term "computer", so that it is subdivided again using the word "computer" as a criterion. This produces two classes consisting respectively of "computer, program, digital, memory" and "system, circuit, data"; the term "machine" which is generic to the whole class is left on the second level. The original "other" category can also be subdivided, using the included high-frequency word "operate" as a guide, and producing the complete hierarchy shown in Fig. 21.

A comparison of the hierarchies of Figs. 20 and 21 reveals that the word groups produced by the thesaurus question method of Fig. 20 may be more reasonable; however, the frequency procedure is more systematic and may conceivably be easier to apply.

The last hierarchy formation process is also based on a term-document or a term-property matrix. In this case, however, the process of forming the hierarchy is completely automatic, even though the original property matrix may have been constructed by hand. Consider two arbitrary terms identified by weighted property vectors. The following conditions may then obtain:

1) terms A and B are identified by different properties, and as such are not related;
2) terms A and B are identified by the same properties, and the



Hierarchy Construction by Frequency Algorithm

Fig. 21

weights of the properties are reasonably similar for both terms, so that neither term dominates the other, and they are placed in the same concept class;

3) terms A and B are identified by the same properties, but the property weights are higher for term A than for term B; then A may be said to dominate B, and may be placed on a higher level in the hierarchy;

4) terms A and B are identified by the same properties, and B dominates A.

In order to be able to make a decision concerning the similarity between two property vectors, it is necessary to compute a similarity coefficient between them. In the present context, it is best to use an asymmetric coefficient such that the similarity between term i and term j is not necessarily the same as between term j and term i. Given property vectors $\underline{v}^i$ and $\underline{v}^j$, representing terms T_i and T_j respectively, a possible similarity measure is

$$\underline{c}_{ij} = \frac{\sum_k \min(\underline{v}_k^i, \underline{v}_k^j)}{\sum_k \underline{v}_k^i}.$$

Using this measure, a term-term correlation matrix can now be constructed, giving for each pair of terms the similarity measure c. It may be noticed, that if the two vectors $\underline{v}^i$ and $\underline{v}^j$ are identical, then $\underline{c}_{ij}$ equals 1, and when $\underline{v}^i$ and $\underline{v}^j$ have no common properties, then $\underline{c}_{ij}$ equals 0. A cut-off value K may now be applied to the similarity coefficients, and a hierarchy may be formed based on the following algorithm:[11]

if $\underline{c}_{ij}$ and $\underline{c}_{ji}$ are both below the cut-off value K, then terms i and j are unrelated;

if $\underline{c}_{ij}$ and $\underline{c}_{ji}$ are both above cut-off, then terms i and j are synonymous and are placed in the same thesaurus category;

if $\underline{c}_{ij}$ is below cut-off and $\underline{c}_{ji}$ above cut-off, then term i is a parent of term j in the hierarchical arrangement;

finally, if $\underline{c}_{ij}$ is above cut-off and $\underline{c}_{ji}$ below cut-off, then term j is a parent of term i.

This system may not generate a true tree structure, since a given term may have more than one assigned parent. The method is, however, fully automatic, and a manual revision after the initial generation can be used to modify the resulting hierarchy to make it acceptable. This can be accomplished, for example, by introducing cross-references between terms in the hierarchy to replace the connections which are not compatible with the tree organization. A set of sample vectors is treated in the suggested manner in Fig. 22. It is seen that property vectors which intuitively appear to be similar will in fact be classified as synonymous (case 1), vectors which appear unrelated are classified as unrelated (case 2), and vectors for which an inclusion relation is apparent are assigned a hierarchical ranking.

Various procedures have been suggested for updating hierarchies and dictionaries by addition of new terms and deletion of old ones.[11,12] These must be used in conjunction with the dictionary look-up operations in any operating situation.

Case 1 : synonymous terms

$$\underline{v}^i = (3, 0, 0, 5, 1, 0)$$

$$\underline{v}^j = (2, 0, 1, 5, 2, 0)$$

$$\underline{c}_{ij} = \frac{\Sigma(2, 0, 0, 5, 1, 0)}{\Sigma(3, 0, 0, 5, 1, 0)} = \frac{8}{9}$$

$$\underline{c}_{ji} = \frac{\Sigma(2, 0, 0, 5, 1, 0)}{\Sigma(2, 0, 1, 5, 2, 0)} = \frac{8}{10}$$

Assuming cut-off $K = 0.7 \Longrightarrow \underline{c}_{ij}$ and $\underline{c}_{ji} > K$

Case 2 : unrelated terms

$$\underline{v}^i = (3, 0, 0, 5, 1, 0)$$

$$\underline{v}^j = (0, 1, 3, 0, 1, 0)$$

$$\underline{c}_{ij} = \frac{1}{9} \qquad \underline{c}_{ji} = \frac{1}{5}$$

For cut-off $K = 0.7 \Longrightarrow \underline{c}_{ij}$ and $\underline{c}_{ji} < K$

Case 3 : term i is a parent of term j

$$\underline{v}^i = (3, 0, 0, 5, 1, 0)$$

$$\underline{v}^j = (1, 0, 1, 3, 2, 0)$$

$$\underline{c}_{ij} = \frac{6}{9} \qquad \underline{c}_{ji} = \frac{6}{7}$$

Here $\underline{c}_{ij} < K$ and $\underline{c}_{ij} > K \Longrightarrow$ term i is parent of j

---

Sample Automatic Hierarchy Formation

Fig. 22

References

[1] G. Salton, Automatic Phrase Matching, 1965 International Congress on Computational Linguistics, New York, May 1965.

[2] C. Harris, Dictionary and Hierarchy Formation, Information Storage and Retrieval, Report ISR-7, to the National Science Foundation, Section III, Harvard Computation Laboratory, June 1964.

[3] C. Harris, Dictionary Construction and Updating, Information Storage and Retrieval, Report No. ISR-8, to the National Science Foundation, Section VII, Harvard Computation Laboratory, December 1964.

[4] L. B. Doyle, Is Automatic Classification a Reasonable Application of Statistical Analysis of Text, Journal of the Association for Computing Machinery, Vol. 12, No. 4, October 1965.

[5] S. F. Dennis, The Construction of a Thesaurus Automatically from a Sample of Text, presented at Symposium on Statistical Association Methods for Mechanized Documentation, Washington, March 1964.

[6] G.Salton, Data Manipulation and Programming Problems in Automatic Information Retrieval, Communications of the ACM, Vol. 9, No. 3, March 1966.

[7] L. B. Doyle, Semantic Road Maps for Literature Searchers, Journal of the ACM, Vol. 8, No. 4, October 1961.

[8] L. Rolling, Euratom Thesaurus - Keywords Used within Euratom's Nuclear Energy Documentation Project, Report EUR 500.e, Euratom Center for Information and Documentation, 1964.

[9] K. Sparck Jones, Experiments in Semantic Classification, Mechanical Translation, Vol. 8, No. 3-4, October 1965.

[10] F. Lévery, Organisation et consultation d'un thesaurus, 1965 FID Congress, Washington, October 1965.

[11] C. T. Abraham, Techniques for Thesaurus Organization and Evaluation, in Information Science, M. Kochen, editor, Scarecrow Press 1965.

[12] P. Reisner, Semantic Diversity and a Growing Man-machine Thesaurus, in Information Science, M. Kochen, editor, Scarecrow Press 1965.

[13] Guy T. Hochgesang, SOCCER - A Concordance Program, Information Storage and Retrieval, Report No. ISR-11 to the National Science Foundation, Section III, Cornell University, 1966.

[14] M. Lesk, Semi-automatic Semantic Classification Systems, Harvard Computation Laboratory, unpublished manuscript, 1965.

[15] L. B. Doyle, Expanding the Editing Function in Language Data Processing, Communications of the ACM, Vol. 8, No. 4, April 1965.