AUTOMATED LANGUAGE ANALYSIS

1973-1974

Report on research for the period

September 1, 1973 - December 31, 1974

(Final Report under this Contract)

Sally Yeates Sedelow, Principal Investigator

This research was supported by
the Office of Naval Research

through

Contract N00014-70-A-0357-0001
Under Project No. NR348-005

The University of Kansas
Departments of Computer Science and Linguistics
Lawrence, Kansas 66045

AUTOMATED LANGUAGE ANALYSIS

1973-1974

Report on research for the period

September 1, 1973 - December 31, 1974

(Final Report under this Contract)

Sally Yeates Sedelow, Principal Investigator

This research was supported by
the Office of Naval Research

through

Contract NOOO14-70-A-0357-0001
Under Project No. NR348-005

The University of Kansas
Departments of Computer Science and Linguistics
Lawrence, Kansas 66045

AUTOMATED LANGUAGE ANALYSIS

1973-1974

Report on research for the period

September 1, 1973 - December 31, 1974

Sally Y. Sedelow, Principal Investigator
Walter Sedelow, Consultant
Robert Bryan
Frank Ford
Herbert Harris
Scott Taylor

Table of Contents

Abstract

This report includes (i) a summary of research pursued under this contract; (ii) a description of the editing of a computer-accessible version of Roget's International Thesaurus; (iii) a discussion of mathematical approaches to the modelling of Thesauri, with Roget's serving as an instantiation; (iv) a Master's Thesis exploring graph theory applications to the study of the structure of Roget's. Articles include "Brief Overview of Research Under This Contract" by Sally Yeates Sedelow, "The Automated Version of Roget's International Thesaurus: A Description with Suggestions for Future Editing" by Herbert R. Harris, "Further Discussion of the Use of Brackets and Parentheses in Roget's Thesaurus" by Scott Taylor, "Modeling in Thesaurus Research" by Robert Bryan, and "Selected Graph Theory Applications to a Study of the Structure of Roget's Thesaurus: A Data Base for Automated Language Analysis" by Scott Taylor.

PREFACE

I would like to express my appreciation to the Information Systems Staff, Office of Naval Research, for their unfailing helpfulness and administrative support during the period of this contract. All of us associated with this project also owe a considerable debt of gratitude to Mr. Mark Katz, University of Kansas Computation Center, who has helped us in many ways, including monitoring the compatability between our programs and the operating system. Finally, the excellent secretarial support, as well as other "back up" from the Department of Computer Science and Linguistics should be acknowledged.

Brief Overview of Research Under This Contract

Research undertaken during this three year contract period has concentrated on the nature of thesauri and upon a particular instantiation of thesauri--Roget's International Thesaurus.[1] The work on thesauri evolved from earlier research supported by the Office of Naval Research which entailed the development of programs directed toward stylistic analysis of documents and other written discourse. One package of these programs called VIA (Verbally-Indexed Associations) was developed to investigate that aspect of style having to do with word choice and word associations. Such investigation is clearly a form of content analysis and the VIA package indeed has been used for various forms of content analysis. Analagously to the needs of special-purpose information retrieval programs (concerned with, for example, analyzing the content of titles in an effort to see whether the document so referenced should be retrieved) for lists of words with which the retrieval program could operate, it became obvious that for a general-purpose content analytic program package, it would be desirable to have a general-purpose reference work of semantically-related words which could be used by the VIA programs as well as by other research efforts demanding access to a store of semantic nets for the English language as a whole.

As possible answers to this need for a general-purpose reference work, synonym dictionaries, word lists, and various forms of thesauri were investigated. Finally, a computer-based comparative study of Webster's Dictionary of Synonyms, Roget's University Thesaurus, and Roget's International Thesaurus was undertaken with the result that a decision was made to put Roget's International Thesaurus into computer-accessible form.

[1]Third Edition, Thomas Y. Crowell Co., New York, 1962.

The decision to put Roget's International Thesaurus into computer-accessible form implied a massive effort, the least of which was the original key punching of the Thesaurus. The editing (still not completed) of the Thesaurus has been one major activity under the current contract. A summary of the editing process and a description of the current, edited version of the Thesaurus is provided by Harris in this report. The article by Taylor in this report on brackets and parentheses also relates to the editing process. In addition, articles by Harris, Taylor and Sally Sedelow in the previous two annual reports detail aspects of the editing process. The crux of the problem has been that the Thesaurus incorporates many of the ambiguities incorporated by human beings in their use of natural language and the Thesaurus also introduces inconsistencies in its own format. Re the latter type of inconsistency, it is not at all clear, for example, why some listings of words appear physically as lists in the Thesaurus and others appear in paragraph form. A problem related to the former type of inconsistency concerns the reader's ability to infer and supply semantic information as well as syntactic information. An example of the former type of inference would be the use of etc. to suggest a continuation of a list in some dimension to be supplied by the knowledgeable reader. An example of the latter type of inference would be the deletion of words (e.g. the deletion of the second "jump" in the entry "jump or off") when the syntactic (and possibly semantic) structure makes clear to the human reader that the deletion is taking place. We have made a genuine effort to edit the Thesaurus so that it can be used for a wide range of linguistic investigations. We are regretful that it has not been possible to complete the "entire" job during this contract period. It should be stressed, however,

that the current version is usable for some research undertakings and that we have compiled the data, including error listings, which will make a very usable version available. In some cases, further editing is a matter of key punching and inserting the corrections for errors which have been identified and for which corrections have been listed.

Computer-based empirical investigation of the Thesaurus has been constrained by the pace of the editing. Nonetheless, a master's thesis developing an approach to the study of connectivity in the Thesaurus has been written (see Scott Taylor in this volume) and it was possible to use a sub-section of the Thesaurus to test the methodology outlined in the thesis. In addition, Bryan's paper in this report is based, in part, upon a computer-based investigation of sub-sections of the Thesaurus.

Another research effort, reported by Warfel in the September, 1971, report for this project, was, in effect, both an exploration of a possible application of the computer-accessible Thesaurus and an empirical investigation of subsets of the Thesaurus. This exploration looked into the possibility of using the Thesaurus to determine when a string of characters which sometimes functions as a prefix is, in fact, functioning as a prefix. The hypothesis explored was that an un-prefixed and prefixed version of the same lexical root would occur within nearby branches of the thesaural tree, while words such as "vent" and "prevent" in which the two "vents" do not have the same meaning would appear at some distance from each other in the thesaural tree. This hypothesis was borne out by Warfel's investigation and, at the same time, his investigation suggested some possibly desirable modifications in the printed Thesaurus as it now stands.

It should be noted that the interest in prefixation developed under an earlier contract, again with reference to the VIA programs, because we

believed that grouping words together with the same root represented a first pass at pulling together words which are semantically related. Dealing with suffixes was considerably more straightforward than handling prefixes. In fact, we discovered that linguistics had relatively little to offer on the theory of prefixation, and a doctoral dissertation by Warfel--Studies in the Semantics of English Prefixation--was written during this current contract period. The dissertation was printed in last year's annual report for this project.

Just as it is our feeling that there ought to be theory to facilitate a decision as to when a character string is functioning as a prefix and when it isn't, it is our belief that in order to use and possibly modify any instantiation of thesauri it is desirable to have a theoretical model of thesauri. Under this contract, Robert Bryan has worked on this problem, publishing papers both in this report and last year's report. Again, it is extremely desirable to test aspects of the theoretical model through empirical investigation. Under this contract, we have been able to make only the most modest start towards testing some of the hypotheses embodied in the general model. It is my suspicion, for example, that much more extensive connectedness will be revealed than is suggested by Taylor's investigation (see his thesis in this report), if the interpretation of separate items in a list as comprising a semi-colon group is modified or effectively ignored.

We have barely scratched the surface of the study of the connectivity structure(s) of the Thesaurus. The implications of this work for measures of semantic distance, and the implications, in turn, of such measures for a wide range of information retrieval, information analytic, and other natural language applications cannot be over-emphasized. It is our hope that a continuation of this effort will be possible.

The Automated Version of Roget's International Thesaurus: A Description with Suggestions for Future Editing

by Herbert Harris

Part I THE PARSED FORM OF ROGET'S INTERNATIONAL THESAURUS

The computer accessible version of Roget's International Thesaurus[1] has now been parsed. This parsing consisted of identifying each comma delimited word or set of words as an entry in the Thesaurus. A file has been created with each entry identified explicitly. A description of the format[2] of that file, a discussion of the decisions made in building the file, and work that remains to be done follows.

Each record in this file is composed of the entry itself, and all the information about its location in the Thesaurus and any other information that accompanied the entry in the Thesaurus. Each record has a set of twenty-two (22) fixed length fields preceding a variable length field which contains the entry, itself. The format of the record and the contents of each field are as follows:

1. The first field is eight characters long. In the parsed version, this field contains eight zeroes. When the parsed version is sorted, this field will contain a sequence number for the sorted entries. Including this field now provides that the same format can be used for both the parsed version and a sorted version.

2. The second field also contains eight characters. The number occurring in this field is the sequence number of the entries as they are encountered in the Thesaurus from page one to page 657. For example the entries in the first paragraph are: existence, subsistence,

being, entity, essence, occurrence, presence, and life. These would be numbered as follows:

existence	00000100
subsistence	00000200
being	00000300
entity	00000400
essence	00000500
occurrence	00000600
presence	00000700
life	00000800

This number is incremented by one hundred to allow for insertion of entries on the file. When the corrections are made the numbers will not increment smoothly from one entry to the next although the last entry will always be less than the present one.

3. The third field will designate what version of the Thesaurus is being worked on. It is anticipated that there will be a number of versions of the Thesaurus -- at least three. In version 1, the entries are encountered in the same order as that of the Thesaurus from front to back. This version is the one being described here and is labeled with a one in this field. In version 2, the entries will be sorted more or less alphabetically. This version would provide an index for human use. Several versions of this sort order are currently being examined for their usefulness and ease of use. And version 3 would be a sorted version using a site specific sort order for machine internal processing. It is anticipated that for output, only versions 1 and 2 will be useful; therefore, fields are provided for only two sequence numbers per entry. No matter what internal array is present, entries for output can be sorted on either fields one or two for human perusal.

Field 3 which designates the order of the entries, or version of the Thesaurus, contains a one (1) for the parsed version, which is the only version currently available.

4. Field four is a one character field that can be used to indicate a format other than the current one. As editing progresses, this field can be used to indicate different versions of the Thesaurus. Or for some purposes a version with a subset of the information in the complete version would be useful. These differences can be indicated in this field. It currently contains a 1.

5. The fifth field, five characters in length, contains the hierarchy code for each entry. The Thesaurus itself is a series of some 1043 semantic fields, e.g., existence, non-existence, substantiality, unsubstantiality, etc. Each of these thesaural categories is divided into paragraphs. However, the "Synopsis of Categories" at the beginning of the Thesaurus provides a hierarchical tree, from which the 1043 categories "branch". The hierarchy number indicates where in this hierarchical tree, the entry occurs. In the 'Synopsis', there are three levels. In the highest level there are eight branches; the first number in the five digit hierarchy number denotes the relevant branch at the top level, the middle two characters denote the second-level branch, and the final two characters indicate the relevant branch at the third level. For example, a hierarchy number of 10202 would point to the first branch at the top level, Abstract Relations (Branch 1), to the second subdivision (02) of that branch, Relation, and to the second subdivision (02) under Relation, Partial Relation. The next (fourth) level in the tree is one of the 1043 categories in the Thesaurus.

6. The sixth field is the five character representation of the category number mentioned above. All but three of the categories in the book are numbered from 1 to 1040; three categories (306a, 413a, 520a) have an a appended to the number of the category immediately

preceding them. These categories are given an all numeric representation in our parsed version of the Thesaurus. Categories 1-1040 are represented as follows: 00010, 00020, 00030, etc. The numbers are incremented by ten. This arrangement not only allows the addition of categories in the future, but makes all-numeric representation possible by substituting a 1 for the a in those categories that have an a. In the Thesaurus the numbers run 305, 306, 306a, 307, etc. On the parsed version these same numbers run 03050, 03060, 03061, 03070, etc. All numbers have leading zeroes.

7. The seventh field is the number of the paragraph in which the entry occurs. As mentioned before, every category is divided into a series of paragraphs. This field is two characters in length.

8. The eighth field contains the semi-colon group number. Each paragraph is subdivided into groups of entries separated from each other by semi-colons. The semi-colon groups are not explicitly numbered in the Thesaurus. In the parsed version, each of the semi-colon groups is given an explicit number beginning with one at the first of each paragraph. The eight field is three characters long.

9. The ninth field contains a number for the entries within a semi-colon group. None of the entries within the semi-colon group is explicitly numbered in the Thesaurus. In the parsed version these entries also are explicitly numbered, beginning with 1 in each semi-colon group. The ninth field is three characters long.

In connection with the designation of entries within semi-colon groups, some editorial decisions had to be made. As an illustration, take the subsection of the list 48.8, labeled 'knots':

cat's paw
clinch
clinch inside, clinch outside
clove hitch

As can be seen the list is a collection of items, in some generic category, but some lines of the list contain more than one item. There were at least two ways that these lists could be interpreted in the above format. As one option, the whole list could be taken as being simultaneously a paragraph and a semi-colon group, and each item in the list as a separate entry. The other option would consider each line in the list as a semi-colon group and multiple entries within lines as separate entries in that semi-colon group. This latter course has been chosen since it seems to be more consistent semantically with the usage of these designations in the paragraph sections of the Thesaurus. Given this approach, each line in the list may be considered a 'concept' which may be instantiated by more than one specific entry and has to be instantiated by at least one. If there is more than one entry for the concept, then these entries are likely to be substitutable for each other in suitable sentence frames to form nearly synonymous paraphrases. In the paragraphed sections of the Thesaurus, it seems to be the case that entries in a semi-colon group can be substituted for each other to form near paraphrases.

10. The tenth field, one character in length, is open (currently unused) and is filled with a zero.

11. The next field, one character in length, is called the variant sequence number. The eleventh and thirteenth fields work together and must be discussed together. In the Thesaurus there are entries joined with an or. In some cases, the or indicates spelling variations for the entry, in others the entries joined by or appear to be pronunciation

variants. In others, one entry is an abbreviation for another. (For a full discussion of these distinctions see Harris, 1973.[3]) In the parsed version of the Thesaurus, the leaves of the tree formed by the Thesaurus are conceived of as semantic concepts. These concepts can be complex and have internal structure. They can be represented in several ways. For example, they may be represented by a spoken word or a written word. In the realm of the written word, there may be spelling variations for any one word. For example, theater and theatre are spelling variants of each other, the latter being a British spelling. Since we only designate as spelling variations single words, we leave open whether there are exact paraphrases which would in effect be spelling variants of the same concept. But it would be possible for completely different words to be spelling variants of the same concept. In everyday usage a spelling variation is of some specific word. As we have represented spelling variation in the parsed version of the Thesaurus, spelling variation is a variant representation of some semantic concept. A second type of variation entails one spelling as an abbreviation of another. A third type of variation takes place when a writer is trying, through a difference in spelling, to indicate a difference in pronunciation which in turn, indicates a social or geographical dialect. These three presumed reasons for the differences in the character sequences (spelling) representing any one concept have been incorporated in the parsed version of the Thesaurus.

The thirteenth field indicates what kind of variant multiple entries with the same location in the Thesaurus tree (i.e. same category, paragraph, semi-colon and entry sequence number) are. Two entries that are spelling variants of each other will be labeled 1,

pronunciation variation is labeled 2, and abbreviation 3. The eleventh field merely contains a unique number assigned to each variant of one concept location. If there are three spelling variants of a term like chili, chilli, chile, these would be assigned numbers 1, 2, and 3, respectively, in the eleventh field and a 1 in field thirteen. In tabular form the above example would be:

	field 11	field 13
chili	1	1
chilli	2	1
chile	3	1

These variant sequence numbers are assigned in the order in which the variants are encountered in the Thesaurus. With this arrangement all variants of one location must be of the same type.

This representation is not entirely satisfactory since it cannot represent all of the logically possible combinations. It works for all the examples encountered so far. If there are two or three character sequences representing alternate spellings for the concept whose number they bear, the system works. But if it is possible for one of these sequences to have an abbreviation that the others do not, then the system cannot represent this fact. This situation is theoretically possible since there are geographical distributions of names for the same item. For example, the long sandwich served on a hard bun in some places is called a submarine, in others a hoggie, etc. These terms can all be thought of as spelling variants of the same concept. Each of the spellings has a geographical distribution which can be pinpointed in the bracket field, field 16. One of these terms could also have a spelling variant or an abbreviation that applied only to it. With the use of such completely different words, there is the possibility for

an array of more complex relationships than can be represented by the scheme in use here.

12. The twelfth field is a one character open field and is filled with zero.

13. The thirteenth field is a one character field indicating the type of variant and has already been discussed. Recapitulating its contents: One (1) means a spelling variant, two (2) a pronunciation variant, and three (3) an abbreviational variant.

14. The fourteenth field is a one character open field filled with a zero.

15. The fifteenth field is a two character field indicating the typeface and context used for the entry in the printed Thesaurus. In the Thesaurus, there are four different typefaces for individual entries. Words that occur in boldface occur frequently in English. Italicized entries are foreign words or phrases. There are boldface italicized words which are presumably common foreign phrases. And finally there are plainfaced entries.

Entries can occur in one of three contexts. 1. They can be a quotation. No quotation marks have been included in the parsed version except where quoted material occurs within a longer entry. 2. Entries can be in a paragraph or, 3. in a list. All of these cases are distinguished in the following way. A one (1) in the second character position indicates a boldface item, a two (2) italics, and a three (3) boldface italics. A zero (0) in the first character position is plain typeface, a one (1) is a list item, and a two (2) a quotation. There are, so far as I know, no examples of italicized, boldfaced or boldfaced italicized quotations. There are, however, entries where only

one word in a multi-word entry is italicized. There are also quoted words within a quotation. The second set of quotation marks is kept in the entry as well as any internal punctuation marks. Putting the above scheme in a table we have:

00 - - - - paragraph plainface
01 - - - - paragraph boldface
02 - - - - paragraph italics
03 - - - - paragraph boldface italics
10 - - - - list plainface
11 - - - - list boldface
12 - - - - list italics
13 - - - - list boldface italics
20 - - - - quotation

16. The sixteenth field is a four character field designating the type of bracketted information. Some entries are followed by information within brackets. The brackets contain such items as the author of a quotation, level of social usage (i.e. colloquial, slang, etc.), or geographical distribution. (For a review of both the use of brackets and parentheses see Taylor in this volume.) Each type of bracketted information has been assigned a unique number, which is stored in the sixteenth field on the parsed file. For example, [Coll.], a bracket that follows 5,393 entries, has been assigned number 0003. This number will occur in field sixteen anytime an entry is followed by this bracketted comment in the Thesaurus.

17. The seventeenth field, also four characters in length, deals with information contained within parentheses. Parentheses occur in two places in the Thesaurus. They can follow an entry, as in:

brass tacks (as, to get down to brass tacks [coll.])

They can also occur at the beginning of paragraphs:

1.7 (Science of Existence) ontology, metaphysics cosmology.

Parentheses have a number of semantic uses. The second example above appears to be a list heading. (For a discussion of paragraph versus list format see Harris, 1973). There are cases where the parenthesis is used to designate some social fact about every item in the paragraph.

34.12 (colloquialisms) tremendous, terrific, terrible horrible, dreadful, awful, fearful, frightful, deadly.

Every item in the above paragraph would be labeled colloquial by this method. Other parenthetical items give the semantic area in which the entries in the paragraph operate, as in:

26.3 (General Agreement) consensus, . . .

All of the parenthetical items occurring at the beginning of paragraphs have been ignored in the parsing. However, these items were compiled in the file of all parentheses and were assigned a unique number. The parentheses at the beginning of paragraphs were ignored because they are not consistently used; without extensive editing which would require examination of each occurrence individually they do not appear to be usable in any foreseeable way.

Paragraphs that are disguised lists with the list heading included in a parenthesis will be designated as lists by putting a 10 in field fifteen of each of the entries of those paragraphs. Entries that occur in paragraphs that are parenthetically labeled slang, colloquial, etc., will be given this designation by including the appropriate number in field 16. This practice will make these entries consistent with the major usage in the Thesaurus. For example, in paragraph 34.12 given above, every entry in that paragraph will have a 0003 in field sixteen.

When this seventeenth field has no relevance for an entry (because the entry has no parenthetical information) it is filled with zeroes.

18. Field 18 is a two character field designating the part of speech for each entry. Some paragraphs begin with a part of speech designation for the whole paragraph. In the majority of cases, this designation holds for all succeeding paragraphs until a new designation occurs. This continuity is not however the case for lists. The items in a list in most cases are collections of nouns. Consequently, all lists have been labeled as nouns. There are, however, occasional lists which are collections of adjectives, etc., which have been labeled incorrectly. These will have to be identified and corrected.

In the parsed version, each designation of a part of speech has been assigned a unique number. These part of speech designations may be ambiguous. For example, there are paragraphs in which the entries are not all one part of speech. This ambiguity will also have to be edited out.

One other editing problem remains: the input file to the parsing program contained parts of speech, as follows:

1.	nouns	15.	advs. etc.
2.	verbs	16.	nouns etc.
3.	adjs.	17.	conjs. etc.
4.	advs.	18.	adjs., advs.
5.	preps.	19.	nouns *
6.	prons.	20.	nouns etc. *
7.	verb *	21.	advs., preps.
8.	interjs.	22.	preps., advs.
9.	phr.	23.	adjs., advs. *
10.	conjs.	24.	nouns preps.
11.	interrogs.	25.	verbs *
12.	proverbs	26.	adjs., preps.
13.	phrs.	27.	noun *
14.	preps. etc.	28.	interjs. etc.

Items marked with asterisks were caused by errors on the input file. These errors will be corrected during the current editing pass. The above numbers are the numbers that are currently on the uncorrected file. Ambiguous cases such as 14, 15, 16, etc., must be dealt with later.

When the file is corrected to correspond to the Thesaurus the part of speech designations for each entry will be:

1. nouns
2. verbs
3. adjs.
4. advs.
5. preps.
6. prons.
7. interjs.
8. phrs.
9. conjs.
10. interrogs.
11. proverbs
12. preps. etc.
13. advs. etc.
14. nouns etc.
15. conjs. etc.
16. adjs. advs.
17. advs. preps.
18. nouns, preps
19. adjs. preps.
20. interjs. etc.

19. Field 19--the cross-reference field--is twenty-four characters long. Following some entries, there are cross-reference numbers which give one or more locations where additional words with similar meaning also occur. It is not clear how to use these cross-reference numbers. In most cases the word with the cross-reference number occurs in the paragraph referred to by the cross-reference number. But this is not always the case. And in some cases, a whole category is cross-referenced. In such a case, should the whole category be used or only the entries having the same part of speech designation? If a paragraph is referenced, should only the words in the semi-colon group in which the cross-referenced word occurs be used, should all of the words in the paragraph be used, or should only the words following the cross-referenced word in the paragraph be used? Before these cross-reference numbers can be used a decision will have to be made on these issues and a check made on consistency. A more trivial editorial problem arises

from the fact that the a's following category numbers have not been eliminated in the cross-reference numbers. These will have to be changed to a one (1).

Field 19 is filled with blanks unless there is a cross-reference number. In one instance,

514.9 Dice Throws 90.3, 93.2, 96.1, 99.2-7

the cross-reference number exceeds the length of the field. At the moment, this number has a * as the final character in the field. The problem can be eliminated by closing up intervening blanks between the comma-separated paragraph numbers.

20. The twentieth field is a two character blank field which contains zeroes. It is anticipated that during the sorting the entries will be scanned for a set of 'illegal' characters and a count of such characters will be stored in this field. For example, the presence of brackets will be checked for. This will give an additional 'proof reading'.

21. The twenty-first field, two characters in length, contains the number of English words in the entry. Of the 198,423 entries identified by the parsing program, 57,307 were multiple word entries such as matter of fact or the case as opposed to fact.

22. The twenty-second field is a two character field giving the number of characters in the entry. The longest valid entry identified so far is a quote from Cicero with 91 characters, although there are mistakes presently on the file that extend to 96 characters. This character count includes the blanks between words in the entry, but not the blanks following the end of the entry (see below).

23. The twenty-third, and final, field is a variable length field which contains the entry itself. It contains all the characters that occur in the Thesaurus including internal punctuation and characters which designate the accent marks on foreign words. A list of these accent mark characters has been given in earlier reports (Harris, 1972)[4]. They precede the accented letter.

For convenient use of the University of Kansas Honeywell 635 computer, the entry is blank filled to the next computer word. In order to insure that the file fits on one tape, the block size has been raised to 4094, one word less than the maximum.

Putting the above information about the format of each record in a table we have:

Field	Begins	Ends	# Char.	
1	01	08	08	''00000000''
2	09	16	08	Parsed Output Sequence
3	17	17	01	Origin Code/Sort Code
4	18	18	01	Record Type Code
5	19	23	05	Hierarchy Code
6	24	28	05	Category Code
7	29	30	02	Paragraph Code
8	31	33	03	Semi-Colon Group Number
9	34	36	03	Entry Sequence Number
10	37	37	01	Open
11	38	38	01	Variant Sequence Number
12	39	39	01	Open
13	40	40	01	Variant Type Code
14	41	42	02	Open
15	43	44	02	Typographic Quality Code
16	45	48	04	Bracket Code
17	49	52	04	Parenthesis Code
18	53	54	02	Part of Speech Code
19	55	78	24	Cross Reference Text
20	79	80	02	'00''
21	81	82	02	Count of English Words in Entry
22	83	84	02	Count of Characters in Entry
23	85	(180)	(96)	Text of Entry

PART II FUTURE EDITING

In this section, I will discuss further editing necessary for the comprehensive use of Roget's. The work can be divided into sections so that with the completion of each task another feature of the Thesaurus can be used. Some of these editing tasks have been mentioned or briefly discussed in the previous section. In this section I will try to be as comprehensive as possible.

1. As the file now stands, it can be used to reflect the semantic grouping of the general vocabulary, insofar as the Thesaurus is consistent and correct in its semantic grouping. Errors present on the file will lead to 'funny' entries, which will not be matched against an incoming text. For example, there are several 'entries' which are all blanks, some incorrectly parsed entries, and of course some misspellings.

The estimated error rate of from 2 to 4 percent that occurred in the parsing consists mostly of entries that were skipped in the parsing. Therefore except for about .05% of the above rate of 2 to 4 percent, the error is a matter of inconsistencies when the file is proofed against the book. The program which parsed the original file also identified those sections that were skipped for one reason or another, scanned the identified entries for possible errors, and made note of these possible errors. These listings can now be used to correct the file. When these corrections are completed, the error rate (estimated informally) will be about .05%, consisting largely of misspellings.

2. It was noted in the previous section that some correction is necessary in order to use the part of speech designations in the Thesaurus. This work again divides into two parts: making the file correspond exactly to the Thesaurus and making the Thesaurus more

explicit. The corrections necessary to make the file reflect the facts of the Thesaurus have already been described. We might also include here the relabeling of improperly labeled lists.

There are two types of ambiguous categories in the Thesaurus. First there are categories which have two designations, e.g. adjs. advs. Since many words in English can serve in several syntactic categories, it could be the case that entries in these paragraphs are meant to have two part of speech designations. However, this possibility does not seem to hold for the following example:

966.20 advs., preps. in favor of, for, pro, all for.

All of these except pro seem suseptible to a prepositional interpretation. Pro itself is ambiguous being interpretable as either an adverb, adjective, or noun depending on the context. But given that it must be either an adverb or a preposition, it would have to be an adverb as in a pro American stance. Whether this practice is consistent through the book I do not know.

The other type of ambiguity entails an ambiguous marking at the beginning of the paragraph, e.g. conjs, etc. One would assume that such a paragraph is a collection of different category types. But each case would have to be looked at individually and each item marked.

If it is the case that some entries should have an ambiguous category marking, then the list of category markings will have to be expanded to include these multiple category cases.

3. There are some problems with the use of the lists. It has already been mentioned that some lists have received the wrong category marking because of the assumption by the parsing algorithm that all lists are collections of nouns. The regular algorithm for the assignment

of part of speech designations in the Thesaurus would not work with lists. Lists are placed at the end of the 1043 categories in the Thesaurus. Therefore although the list is a collection of nouns and has no part of speech label it may follow any other part of speech. Since most of the lists are collections of nouns, they were arbitrarily assigned to the noun category. But there are instances where lists are not collections of nouns. These erroneous assignments will have to be identified and corrected. An example of such a list is:

28.8 Quantities
armful
bag, bagful
barrel, barrelful
. . . etc.

There is also one very erratic list, 182.10. Principal Cities of the World. In most lists like that above when there is more than one entry per line a comma is used to separate those entries. In 182.10 the comma is used to separate the city from the county. These were interpreted by the parsing program as being two separate entries and they are so listed on the file as it now stands. After correction such entries will have an internal comma, e.g. Accra, Ghana. This particular list also has another peculiarity. In the Thesaurus a * is placed in front of cities that are capitals. Since a * was used as a symbol for the following letter to be capitalized, two *'s (**) were used to denote capitals. Therefore there are entries in this list with three preceding *'s (***).

4. As noted earlier, parenthesized material occurs in two different places. Some parentheses occur after entries and others are at the beginning of paragraphs. Both of these cases have been discussed in the previous section. In fact, the parentheses have several semantic

functions, some of which overlap with the use of brackets. In a complete editing of the parentheses, they need to be looked at together with the brackets in order to differentiate their semantic functions and to devise ways to code these functions. The brackets also serve several semantic functions which need to be differentiated.

5. The cross-reference numbers need to be edited and perhaps expanded. First, cross-reference numbers followed by an a, e.g. 27.5 (verbs) dissent 520a.4, need to have the a changed to a 1 and the whole format of the number changed to conform to the usage of the category and paragraph numbers. The cross-reference numbers occur in several formats:

1.	life	406
2.	authenticity	515.5
3.	mother	169.8-12
4.	equal	30.8,9
5.	mammoth	194.20, 21

The corrections necessary range from the trivial, such as making four and five uniform by either including a space or deleting it, to the need for the specific interpretation of one through three above (for a discussion of this matter see page 17 above). If the cross-reference number is interpreted as a device which will add possible entries to a semantic field, the level of the Thesaurus tree being referred to by the cross-reference number is not clear. In the first example above an entire category, 406, is referenced. The cross-reference comes from the lowest level of the tree, from one of the leaves. But if a paragraph is referenced, it is probably not the case that all of the entries in the paragraph are meant to be added to the semi-colon group. For if this were the case either the internal structure of these paragraphs would be lost or new lower levels would be added to the

Thesaurus tree. It might be the case, however, that if the word with reference number also occurs in the paragraph referenced, only the entries co-occurring in the newly referenced semi-colon group should be added to the first semi-colon group.

It appears that the order of entries in the semi-colon group and the order of semi-colon groups in the paragraph is not important. Therefore, entries or semi-colon groups could be added without too much attention to ordering. However, when the cross-reference refers to the heading for an entire category, (e.g. 406 Life, above), we have the problem of whether or not to add this whole category to the semi-colon group from which the reference came and where to add the category or any part of it. For example, life is in a paragraph labeled nouns. It will add many levels to the Thesaurus tree if the entire category is added as a structured branch under the semi-colon group level. There could be an infinite number of levels if these references turned out to be circular. If the branching distinctions of the referenced category are interpreted as parallel to the category from which the reference came, then at least one additional level will be introduced to encompass these two categories. For example, in the above example of life, the categories of Life and Existence would have to be joined by a new level just above the category level.

This short discussion of some of the possibilities will at least adumbrate the problems involved in the use of the cross-reference numbers. In anticipation of this editing, during the parsing a complete file was made up of all the entries having a cross-reference number. This file can now be used for this editing.

6. The decision to introduce the three semantic distinctions: 1. entries as spelling variants of each other, 2. entries as pronunciation variants of each other, and 3. an entry as an abbreviational variant of another, has introduced some inconsistencies into the file. These distinctions were first introduced during the editing when eliminating the or separating two entries. A search of the entire file has not been made for such variants. These distinctions have only been included when examples were encountered. Therefore there will be examples of these distinctions that have not been explicitly marked. The abbreviations will probably constitute the greatest number of unprocessed entries since there were fewest of these present with an or separating the full and abbreviated form.

PART III SEMANTIC TESTING

Although it has semantic implications, the previously discussed editing has been aimed at making the information contained in the Thesaurus explicit. But the semantics of the Thesaurus should also be examined. Two semantic questions that should be raised are: First, how complete is the file?, and second, how consistent is it? The first question can be asked in two ways. How large a sample of words of the whole language does the Thesaurus contain? This question can be answered by comparing the Thesaurus with other word lists. When the Thesaurus is to be used against incoming texts, the most useful comparison would be against lists that contain words that have actually been gathered from texts. The Thorndike-Lorge list is an example of such a list. With this kind of comparison the probability of matching words from an incoming text can be ascertained.

Comparison can also be made against dictionaries. Comparison to a small dictionary will give a rating of how well the Thesaurus will cover the incoming text, insofar as the dictionary contains the most used subset of the words of a more complete dictionary. Larger dictionaries will give measures of completeness with respect to the whole language. Since idioms and quotations are included in the Thesaurus, a comparison with idiom dictionaries as well as dictionaries of quotations would give useful measures concerning these language strings.

An incomplete list of quotations would not appear to inhibit semantic comprehension as much as idioms since in most cases they, unlike idioms, are the sum of their parts.

The second kind of completeness comparison is whether the range of entry locations for any one entry covers the range of meanings possible for that entry. This question is related to the number of definitions any entry may have in a dictionary. It might be that comparing these two numbers, the number of times an entry occurs in the Thesaurus and the number of meanings that this entry has in a dictionary, might give a quick and dirty measure of this kind of semantic coverage.

This comparison is complicated by the fact that small dictionaries combine noun and verb entries and do not list morphological variants separately. Larger dictionaries list all of these separately, but they also include jargon meanings giving a longer list of meanings. Ideally one would want to compare the exact meanings in the dictionary with the meanings implied in the Thesaurus. If this last kind of comparison could be done, it would be possible to uncover systematic omissions in the Thesaurus, that is sections of the semantic space of the whole language that are not covered.

To test the consistency of the Thesaurus is more difficult. First, judgments concerning consistency presuppose knowledge of how the decisions about the grouping of entries ought to have been made. It is probable that the decisions actually made about grouping have not been explicitly verbalized by the people compiling the Thesaurus. At least two avenues present themselves as entries to this kind of research. The easiest is to develop a program that will take incoming sentences and form paraphrases of these sentences by substituting words in the same semi-colon group for the content words that occur in the sentence. This substitution can be controlled to exclude those

entries from foreign countries or perhaps from the wrong sociological level. Looking at the output from such a program would provide one kind of test of the reliability of the Thesaurus. Failures among the paraphrases might also give clues to the principles of grouping that were employed.

As was mentioned before, it does not appear that the sequencing of entries in the semi-colon group or the sequencing of semi-colon groups were controlled. Some of the inconsistencies of the placement of the negation of an entry have already been discussed in Warfel (Warfel, 1974)[5]. To look at higher level (paragraph, category, etc.) sequencing, it is necessary to know the semantic dimensions of their arrangement. To give an idea of what is involved here take category 699. This category is labeled Preservation. Looking only at the nouns in this category, there are six paragraphs with groups of nouns. The first paragraph consists of two semi-colon groups listing types of preservation:

> 699.1 preservation, preseval, conservation, saving, salvation, keeping, safekeeping, maintenance, support; protection 697.

We can cast these words in different sentence frames to focus on how they differ from each other.

1. He is interested in the preservation of his strength.
2. He is interested in the conservation of his strength.
3. He is interested in the protection of his strength.

For me the first is the most general kind of saving. The second is a saving from use. And the third is a saving from some external agent. For me it is questionable whether 'safekeeping' should be in the first semi-colon group since it seems to involve the implication of an external agent. It should be with protection, the second semi-colon

group.

The second paragraph lists different processes for preservation; these processes are separated into semi-colon groups on the basis of method: 699.2.

> 699.2 curing, seasoning, slating, brining, pickling, corning, kippering, jerking, marination; drying, dry-curing; dehydration, anhydration, evaporation, desiccation; smoking, fuming, smoke-curing; refrigeration, freezing, quick-freezing; embalming, mummification; canning, tinning [chiefly Eng.]; bottling, potting.

Taxidermy is not included in this paragraph, although presumably it would go with embalming. This paragraph is in fact labeled by a parenthesis, 'Means of Preservation'.

The third paragraph lists the chemical substances used in the processes of the previous paragraph. These inanimate, material causes are called instruments in linguistics. They are contrasted with animate effective causes called agents. The first semi-colon group with two entries gives the general terms preservative, and conservative. The second semi-colon group lists more specific substances like salt, brine, vinegar, formaldehyde, etc. Presumably a definition for each of the listed substances could be formulated which would not mention that in our culture these substances are used in preservatives. It is the case, however, in most dictionaries that this fact is included in their definitions either directly or by implication as in the case of brine.

The fourth paragraph could be either more instruments or they could be agents:

699.4 preserver, saver, keeper, safekeeper; lifesaver.

Some of the entries could be either, but if the entries are consistent, then these potentially ambiguous entries can be interpreted as being

agents.

The fifth paragraph again is a collection of instruments.

699.5 life preserver, life jacket, life belt, safety belt, swimming belt, cork jacket, etc.

All of these have to do with the sea. As instruments they should have been included in paragraph three. But if this had been done, it would have been impossible to subdivide them and still group them together as being associated with the sea. There are six semi-colon groups in this paragraph.

The last paragraph labeled nouns is also labeled with a parenthesis, 'Place set apart for conservation'. Linguistically these entries would be labeled locatives. Some examples are:

preserve, reserve,. . . national park;. . . Indian reservation, . . . etc.

The six paragraphs labeled nouns in category 699 therefore seem to be differentiated from each other along five semantic dimensions: 1. result, 2. process, 3. instrument, 4. agent, 5. locative. But in this one example the instruments are not grouped in the same paragraph nor are they grouped together. It should be emphasized that this list of dimensions does not recur in every category. If however some set of dimensions is used throughout the Thesaurus, then paragraph nodes in the Thesaurus can be labeled and their ordering controlled. If, however, every category presents a unique set of dimensions, then ordering is not possible and the grouping in the Thesaurus would be by some other principle such as clustering of usage around real world events or by psychological association. It is obvious that some clustering around real world associations is present, i.e. the instruments concerned with the sea above.

This brief discussion is not meant to solve any problems, but

only to suggest some possibilities. Now that the editing to make entries explicit is nearly complete, the file can be put to use and even though it may not be 'perfect' semantically it may be useful and the limitations discovered will suggest future changes.

Bibliography--Footnotes

1. Roget's International Thesaurus, 1962, third edition (Thomas Y. Crowell Company: New York)

2. I would like to thank Mark Katz of the University of Kansas Computation Center for his patience, help and advice. That the format being described here works at all is largely the result of his efforts. It would probably be better, if he didn't have to contend with me.

3. Harris, Herbert, 1973, "The Conversion of Roget's International Thesaurus to an Automated Data Base," in Automated Language Analysis 1972-73, Sally Yeates Sedelow, et. al., (University of Kansas) pp. 5-34.

4. _____, 1972, "Further Editing of Roget's International Thesaurus Tape and Some Observations on Further Studies of the Thesaurus, in Automated Language Analysis, 1971-72," Sally Yeates Sedelow, et. al. (University of Kansas) pp. 19-30.

5. Warfel, Sam, 1972, "The Value of the Thesaurus for Prefix Identification, in Automated Language Analysis 1971-72,) Sally Yeates Sedelow, et. al. (University of Kansas) pp. 31-49.

Further Discussion of the Use of Brackets and Parentheses in Roget's Thesaurus

by Scott Taylor

An important source of information in Roget's Thesaurus is the bracketted and parenthesized information associated with particular entries and groups of entries. In general, this information serves as an additional source of semantic and syntactic information in the Thesaurus. Oftentimes, bracketted and parenthesized information appearing with entries serves to clarify and/or modify the meaning and usage of entries, and can greatly affect the use of the Thesaurus for automatic language analysis. As briefly discussed in last year's report [1], during the process of constructing a machine-accessible version of the Thesaurus considerable attention has been focused on an effective means of dealing with this qualifying information. The major goals of the work involved in developing a means of handling the qualifying information appearing bracketted and parenthesized are:

1) to identify the kinds of information presented in brackets and parentheses.

2) to determine how this information is used in the Thesaurus and perform any normalization and/or corrections in usage which may be necessary.

3) to determine how to incorporate this information into a machine-accessible version of the Thesaurus in order to allow efficient access to and use of this information.

As reported last year, there are essentially five major kinds of information found enclosed in brackets in the Thesaurus.

1) The origin and/or usage of entries.

463.1 minstrel [poetic]

563.6 dry nurse [slang]

564.7 debutante [fem]

2) Geographic information -- countries, locales within countries, and parts of the world associated with entries.

44.9 marabou [Louisiana]

825.5 trade-in [U. S.]

787.6 scissorbill [slang, West. U. S.]

3) Point in time associated with usage of entries.

125.6 flapper [1920's]

300.11 mizzle [slang, Eng.; circa 1781+]

4) Translations of foreign phrases and words.

520.9 hurler avec les loups [F., howl with the wolves]

542.7 auspicium melioris aevi [L., omen of a better wage]

5) Sources of quotes.

139.1 "the ringing grooves of change" [Tennyson]

139.1 "the changes and chances of this mortal life [Book of Common Prayer]

In addition, the following kinds of information may be found parenthesized in the Thesaurus.

1) Translations of acronyms.

276.24 LCC (landing craft, control)

348.31 UNIVAC (Universal Automatic Computer)

2) Descriptions of entries -- examples of usage.

74.5 pride (of lions)

39.4 rollback (as in price [coll.])

113.2 sudden (as, all of a sudden)

218.16 hanging (as, hanging gardens)

3) Word forms and usage.

44.8 mestiza (fem. [Sp.])

4) Explanations of entries -- often involving more generic equivalents of technical or "unusual" entries.

370.6 jade (green)

374.2 Caelus (deified sky [Rom. Myth])

391.9 (science of humidity) hygrology, hygrometry, psychrometry, ...

414.4 Porifera (sponges)

413.21 Copenhagen (Wellington's charger at Waterloo)

Examination of the above examples should provide some insight into the importance of bracketted and parenthesized information when using the Thesaurus. Such information associated with thesaural entries can provide the necessary information for identifying contextual usage of words and phrases when examining text automatically. For example, consider the following two occurrences of the word "assembly" in the Thesaurus.

74.2 assembly (of persons)

74.13 (a putting together, as parts of a machine) assembly, assemblage; ...

Consideration of the information contained in parentheses in these two examples reveals two distinctly different meanings (and uses) of the word "assembly." As a result, this parenthesized information provides one means of regulating the search of the Thesaurus for words and phrases related to the word "assembly," based on the context of the particular occurrence of "assembly" in the text being analyzed.

During the development of the machine-accessible version of the Thesaurus all instances of bracketted and parenthesized information have been identified and tape files of the distinct occurrences of each type of information have been created. Likewise, the number of times each distinct piece of bracketted or parenthesized information occurs in the

Thesaurus has been computed, as has been the number of characters comprising these occurrences. For example, the five character word "slang" occurs bracketted in the Thesaurus a total of 4,261 times. As a result, a total of 21,305 characters comprise all occurrences of the bracketted word "slang." Such information is an important consideration in determining the means of handling bracketted and parenthesized information most efficiently in the final machine-accessible version of the Thesaurus. Specifically, such information is necessary to adequately assess the desirability of creating some form of index to bracketted information, and possibly an index to parenthesized information. Such an index could prove very useful as an aid in preventing the size of the machine-accessible version of the Thesaurus from becoming too large for efficient use in automatic language analysis.

As reported in last year's report [1], all occurrences of bracketted information modified with BOTH and ALL were to be distributed among the entries modified with this form of bracketted information. This work has been completed, and as a result, the numbers of brackets and the character counts presented in last year's report have been updated. There are now a total of 21,619 bracketted words and phrases. Of this number, 1,602 are distinct. The number of characters comprising these 21,619 occurrences is 151,818. In the case of parenthesized information it has been found that there is a total of 1,482 occurrences of parenthesized information. Of this total, approximately 1,300 are distinct, and the number of characters comprising the 1,482 occurrences is 23,690. As these figures illustrate, the majority of occurrences of parenthesized words and phrases in the Thesaurus are unique, with very

few of the distinct words and phrases occurring more than once. However, in the case of bracketted information most of the words and phrases are used several/many times, as with the word "slang" mentioned earlier. As a result, the feasibility of constructing an index to bracketted words and phrases appears to be a much more practical solution to the problem of how to handle this information, while in the case of parenthesized information the desirability of such an index is more questionable.

There are a number of special cases, primarily in the use of parentheses, which must be dealt with when attempting to develop a means of handling bracketted and parenthesized information. Some examples of these special cases follow.

First of all, there are cases of cross-references occurring within parentheses which refer the user of the Thesaurus to other locations where information related to the entry or entries may be found. For example, consider the following case in which a cross-reference is used to reference information which serves to "expand" the "etc.".

137.9 ; triennial, decennial, etc. (above 137.4);

Sub-category 137.4 consists of entries which serve to complete the set of entries comprising 137.9.

137.4 ... ; biennial, triennial, quadrennial, quinquennial, sextennial, ...

It would seem that this case warrants replacing the "etc." and the parenthesized cross-reference with the information found in sub-category 137.4.

Secondly, there are many cases of bracketted information occurring in conjunction with words and phrases appearing in parentheses.

For example:

39.4 rollback (as in price [coll.])

44.8 mestiza (fem. [Sp.])

125.8 dogie (motherless calf [West. U. S.])

In the case of the first example, 'rollback (as in price [coll.])', "coll." refers to a particular usage of rollback such as "price rollback." That is, the bracketted information must be used in conjunction with that information occurring with it in parentheses when referencing the entry "rollback." However, the association between the bracketted information and entry in the latter two examples appears more direct and less dependent upon the parenthesized information also associated with the entry. As a result, the bracketted and parenthesized information in the latter two examples can be referenced independently of one another with very little effect on information content. One means of determining the degree of dependency between the bracketted and parenthesized information is to consider other occurrences of the entry in the Thesaurus. For example, the bracketted phrase 'West. U. S.' is associated with every occurrence of "dogie" in the Thesaurus, while 'motherless calf' appears with "dogie" only in the example above. On the other hand, nowhere in the Thesaurus does 'coll.' appear with the entry "rollback," except in the example given here. Such distinctions are important, since those cases in which the bracketted and parenthesized information must be considered together can present special problems when attempting to construct and use an index to each type of information.

Thirdly, there are several cases of the use of "etc." as a shorthand notation for a list occurring within parentheses and brackets.

For example:

514.9 dice throws (snake eyes, etc.)

833.9 peso Argentina, Mexico, etc.

99.16 (multiply by five, etc.) ...

The use of the "etc." in this manner complicates the method in which these cases of parenthesized and bracketted information can be handled, especially if some form of index to the information is to be constructed. Refer to the article by Sedelow [2] in last year's report for a discussion of the use of "etc." in the Thesaurus and possible means of dealing with it.

Lastly, there are uses of brackets and parentheses which seem to warrant the creation of new entries using the bracketted and parenthesized information provided. The most obvious examples involve the occurrence of plural forms of an entry in brackets or parentheses.

41.1 addendum (pl. addenda)

216.5 telamon (pl. telamones)

It is suggested that such cases be handled by creating a new entry consisting of the plural form of the existing entry. That is, for the above examples the following changes would be made.

addendum, addenda, ...

telamon, telamones, ...

A major obstacle to dealing effectively with bracketted and parenthesized information in the Thesaurus concerns the normalization of the words and phrases occurring in brackets and parentheses. Specifically, there are numerous cases of the same information occurring in several different forms, usually in regard to the use of various abbreviations and shorthand devices. Many words, such as colloquial,

appear in the Thesaurus in several abbreviated forms (e.g., coll. and colloq.). Likewise, there are numerous cases of combinations of words and phrases, usually appearing within brackets, containing the same information content, but presented differently, whether in terms of punctuation or form. The handling of these cases is primarily an editorial task, but should be undertaken prior to any decisions regarding the creation of an index to this information.

Although the types of information listed earlier (pages 33-35) are representative of the information appearing in brackets and parentheses, there are numerous cases of discrepancies, whereby information normally appearing in parentheses is found in brackets, and vice versa. The following examples are cases of bracketted information which represent information which normally occurs in parentheses.

158.10 muscle [dial., move by muscular force]

619.4 willed [chiefly in composition, as strong-willed, self-willed, etc.]

834.12 bursary [of a college or monastery]

Likewise, examples of parenthesized information normally occurring in brackets follow.

34.6 (dialect) lashings, power, lavish, might, ...

182.3 (slang terms, chiefly U. S.) one-horse town, jerk-water town, tank town, ...

Usually, as illustrated above, these cases of parenthesized information seem to be a means of indicating that the information applies to a group of entries, as with cases of bracketted information modified with ALL (e.g., [ALL slang]). And, as with bracketted information modified with ALL, such parenthesized information should be distributed among the affected entries.

In conclusion, on the basis of the work completed thus far it appears that bracketted information can best be handled by the creation of an index to this information in the Thesaurus. Such an index will provide a great savings in the amount of space required for this information in the machine-accessible version of the Thesaurus. It seems, as a result of the large amount of repetition of much of the bracketted information, that construction of an index of this information can be fairly easily accomplished. Likewise, the nature of much of the bracketted information is such that users of the Thesaurus may not always require this information, depending upon the analysis being undertaken. In addition, in many cases it appears that for a given entry possessing bracketted information, all occurrences of the entry in the Thesaurus possess the same bracketted information. As a result, the additional semantic information provided by the bracketted information may be minimal, and in some instances nonexistant.

In the case of parenthesized information, however, the creation of an index to this information does not appear to be the most desirable means of handling it. Parenthesized information is responsible in many cases for introducing additional semantic information necessary to properly using entries possessing this information. Specifically, as mentioned earlier, such information is oftentimes critical to properly determining the contextual meaning of a word or phrase. Likewise, in many cases parenthesized information (e.g., when providing the generic equivalent to some technical term) possesses greater information content and may provide more usable information than the affected entry itself. In addition, there is very little repetition of much of the information appearing parenthesized, which minimizes the savings in the storage of this information which might be realized by an index.

This paper has attempted to illustrate some of the problems associated with incorporating bracketted and parenthesized information into the machine-accessible version of the Thesaurus. However, prior to any decisions regarding the means in which bracketted and parenthesized information can or should be incorporated into the machine-accessible version of the Thesaurus and used effectively, a considerable amount of editing and reformating should be performed.

R E F E R E N C E S

[1] Taylor, Scott R., "Handling of Bracketted Information" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1973.

[2] Sedelow, Sally Y., "Etc. in Roget's International Thesaurus" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1973.

Modeling in Thesaurus Research

by Robert Bryan

1. Introduction

Although the various reference works which go under the name "thesaurus" vary a good deal in structure and complexity, it is probably common to most people's concept of a thesaurus that it groups together words of like or identical meanings. This basic structure, a set of word-groups whose members are in some way semantically related, may be expanded and elaborated on in various ways, but it can be taken as the defining characteristic of those real world objects we would want to call thesauri, from a simple synonym dictionary to the more highly structured Rogets International Thesaurus (R.I.T.). The uses of such a data structure in language analysis are varied and complex and have been discussed in recent articles in a number of fields of language research. Unfortunately there is not yet a great deal of standardization in the use of terminology in this discussion and we must often rely on best guesses as to the meanings of terms used by authors who, in the best linguistic tradition, do not feel compelled to provide us with explicit definitions. A great aid in making more precise our statements about thesauri would be a model, in the framework of which ideas having to do with thesauri could be given explicit expression. Given the simplicity of the basic, defining characteristic of thesauri--a set of sets of semantically related words--such a model is readily available. While simple in concept, this model, like many mathematical configurations whose definition is equally simple, admits a rich and complex development.

Improved communication through explicit expression is clearly not

the only benefit we can expect to derive from carrying on the discussion of thesauri in the framework of such a model. Formulating a postulational system of which concrete thesauri are instantiations and bringing to bear the power of deductive reasoning in the development of that system will bring to light a great deal about thesauri that would otherwise remain obscure to us. It will lead us to pose questions about the structure of thesauri that we would otherwise fail to ask and suggests application and avenues of investigation which would otherwise remain unexplored. The reluctance on the part of some to formalize to this degree in the study of what seems to be a very vague and unstructured subject, the semantics of human language, is due, I think, to a misunderstanding of the use of modeling as a research tool. In studying thesauri with the construction of a model it is important to bear in mind that we are making no empirical claims about thesauri nor imposing any "ideal" structure on them. Neither is a model a simplified "approximation" of the real thing that may be wrong in part. Rather, a model, or "abstract thesaurus", is a "way of looking at" concrete thesauri. Specifically, we look at "entries" (word tokens) in the thesaurus as discrete entities, or points, and words and categories as sets of those entities. A "word" is the set of all entries identical in form to a given entry, while a "category" is a set of entries grouped together by the thesaurus as semantically related. We know of entries only information that derives from their membership in words and categories. That is, given entries e_1 and e_2, we know that there does or does not exist a word, w, of which both e_1 and e_2 are members and that there does or does not exist a category, c, of which both e_1 and e_2 are members, but we know nothing else of e_1 and e_2. They are simply distinct elements of a set of elements called entries. In short, we look at a thesaurus the way a computer does.

The human user of a thesaurus obviously knows a great deal more about the entries he encounters than whether or not they are in the same word or category with other entries. Besides knowing whether or not two entries have the same or similar meanings, he knows what those meanings are. He knows that the semantic relatedness signaled by the co-membership of entries in a category may be of varying kinds and intensities. He knows that two entries of identical form are sometimes semantically related in some way even though they are in different categories, such as the literal and metaphorical use of a word, but sometimes are completely unrelated semantically. A machine, on the other hand, does not approach the thesaurus armed with this knowledge of language and the world. The computer suffers from complete semantic aphasia. Entries are simply abstract symbols which are identical in form with certain other entries. The use of thesauri in language analysis by computer, therefore, requires that formal counterparts be formed which approximate the semantic devices available to the human user of the thesaurus. It is in the search for these formal devices that modeling promises to be of greatest benefit.

An "abstract thesaurus" is defined and several lines of development are explored in (Bryan, 73). Some portions of that paper which are relevant to the discussion in the following pages are repeated here. The reader is referred to that paper, however, for the definitions of certain terms that are borrowed from that paper and not redefined. An 'abstract thesaurus', as defined in that paper and below, is an accurate model of R.I.T., in the sense that R.I.T. is an instantiation of that model under at least one (in fact under several) interpretations of the primitive concepts of the model. It formalizes the idea of "thesaurusness" discussed above. It is not a "complete" model of R.I.T., however, and certain information carrying devices in R.I.T. have no formal correlates

in the model. In this paper a complete characterization of R.I.T. is given, formal correlates in that system of various semantic concepts are discussed, and directions of possible, further work in the area are suggested.

2. Abstract Thesauri

"Thesaurusness," as discussed above, is expressed formally in the following definition. Given a finite set S, $|S|$ denotes the number of elements in S.

Definition 2.1

A thesaurus, T, is a triple $\langle E,W,C\rangle$ where

i) E is a non-null, finite set,

ii) W and C are non-null collections of subsets of E,

iii) distinct elements of W are disjoint and distinct elements of C are disjoint,

iv) given any $e \in E$, $e \in w$ for some $w \in W$ and $e \in c$ for some $c \in C$, and

v) given $w \in W$, $c \in C$, $|w \cap c| \leq 1$.

Elements of E are called entries, elements of W are called words and elements of C are called categories. Elements of $M = W \cup C$ are called molecules. A thesaurus is thus a triple $\langle E,W,C\rangle$ where E is a non-null, finite set, W and C are non-null partitions of E, and a word and a category intersect in at most one entry.

A thesaurus lends itself nicely to pictorial representation. The T-graph of the thesaurus $\langle E,W,C\rangle$ is the geometrical configuration in which E and W are represented by sheafs of parallel lines and the intersection of the line corresponding to $w \in W$ and that corresponding

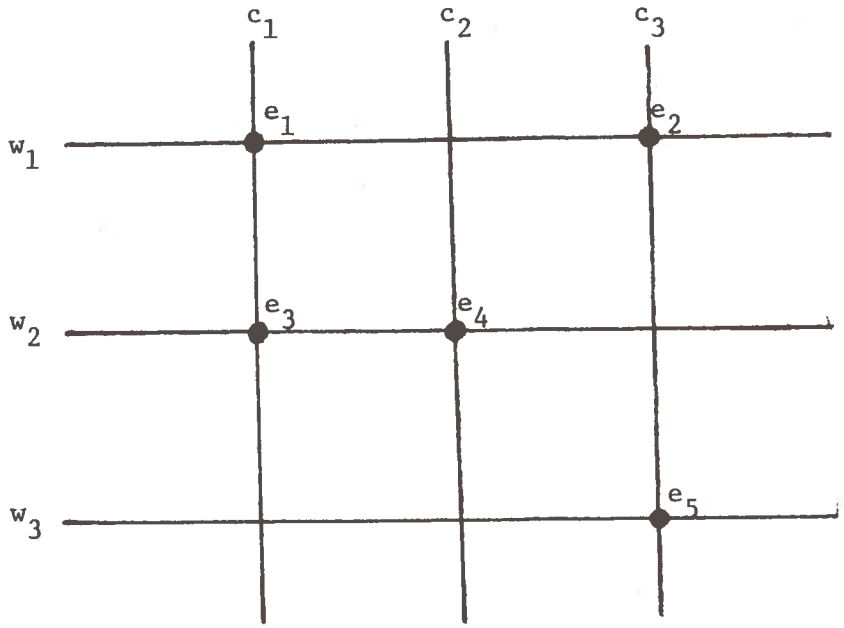
to $c \in C$ is marked with a dot if and only if $w \cap c \neq \emptyset$. Thus, the T-graph of the thesaurus $T = \langle E,W,C \rangle$ where

$$E = \{e_1,e_2,e_3,e_4,e_5\}$$

$$C = \{\{e_1,e_3\}, \{e_4\}, \{e_2,e_5\}\}$$

and $$W = \{\{e_1,e_2\}, \{e_2,e_4\}, \{e_5\}\}$$

is



in which the words and categories of T have been given labels.

Roget's International Thesaurus is an instantiation of an abstract thesaurus under several interpretations of the primitives "entry", "word", and "category" in terms of the definable elements present in Roget's. The "principal interpretation" is that under which "entries" are comma delimited character strings (e.g., the occurrence of the word "subsistence" in the first line of the first page of R.I.T.), "words" are sets of entries consisting of all and only those entries identical to a given entry (e.g., the set of all occurrences of "subsistence" as an entry in R.I.T.), and "categories" are sets of adjacent entries bounded by semicolons (e.g., the set consisting of the three entries "existence",

"subsistence", and "being" in the first line of R.I.T.). "Comma delimited" and "bounded by semicolons" in the previous sentences should be interpreted loosely enough to handle special cases, such as the first or last entry in a grouping larger than a "semicolon group", where the actual delimiter in R.I.T. may be something other than a comma or a semicolon. "Word" in this interpretation clearly has a special meaning distinct from its usual one since an entry in R.I.T. may contain several words in the usual sense and a "word" in the technical sense of this definition is a set of identical entries. That (i) through (iv) in the definition of a thesaurus are true if R.I.T. under the principal interpretation is immediate; (v) is true since no semicolon group in R.I.T. contains two identical entries. Other valid interpretations of the sets E, W, and C in terms of elements of R.I.T. include those under which E and W are interpreted as above and categories are taken to be sets of entries larger than semicolon groups. A precise determination of the full range of possible valid interpretations must be made mechanically.

Since in the definition of thesaurus nothing was assumed to be true of one of the sets W and C that was not also assumed to be true of the other, a duality principle applies to the discussion of abstract thesauri. Specifically, in the development of the abstract system, corresponding to each true statement containing a reference to one of W or C or to elements of that set, there is another true statement, called its dual, formed from the first by replacing each such reference by a reference to the other of W or C. Words and categories are therefore indistinguishable in terms of the properties they possess. They are different entities, but not different kinds of entities. (The duality principle ceases to be in force when, in a complete characterization of R.I.T., we add axioms to the system which are asymetric in that they ascribe different properties

to words and categories. While many statements will have a dual in that system, some will not.) While this may seem, at first glance, counter intuitive, it is an important fact in the development of the abstract system as well as in the application of the thesaurus to various language analysis tasks.

The abstract thesaurus turns out to be a fairly complex mathematical configuration with many interesting properties. Certain developed branches of mathematics, especially graph theory, have important implications in the theory of abstract thesauri. Each definition and each theorem in the development of the abstract system makes a corresponding statement about R.I.T. and about every other concrete thesaurus which, under some interpretation, is an instantiation of an abstract thesaurus. The following general principle describes the relationship which holds between abstract thesauri and any instantiation of an abstract thesaurus:

> A postulational system Σ is a model of a concrete object O if there exists an interpretation (an assignment of meaning to each primitive concept of Σ) under which each postulate of Σ is a true statement about O. ("Model" so defined corresponds to the linguist's, not the mathematician's, use of the term.) If Σ is a model of O under the interpretation I and if $\underline{t}$ is a theorem in the development of Σ (i.e., a statement in the terms of Σ derivable from the postulates of Σ), then the interpretation of $\underline{t}$ induced by I is a true statement about O.

Postulates in Σ and theorems in Σ together are called Σ-statements. An interpretation, I, is the aggregate of meaning assignment to the primitive terms of Σ. The interpretation induced by I of a theorem $\underline{t}$ in the development of Σ is an assignment of meanings to the terms in $\underline{t}$ by which each primitive term in $\underline{t}$ is assigned the meaning given that term by

I and each derived term of $\underline{t}$ is assigned the unique meaning derived from I and the definition of the derived term.

To illustrate, if Σ is the postulational system given in Definition 2.1 and O is the concrete thesaurus, T, with five entries that we used as an example above, then I is the interpretation of Σ under which entries are e_1, e_2, e_3, e_4, and e_5, categories are the sets $\{e_1,e_3\}$, $\{e_4\}$, and $\{e_2,e_5\}$, and words are the sets $\{e_1,e_2\}$, $\{e_2,e_4\}$, and $\{e_5\}$. The primitive terms "entry", "word", and "category" of Σ (or E, W, and C) have each been assigned a meaning. That Σ is a model of the thesaurus T under the interpretation I is verified by checking that each of the postulates of Σ ((i) through (v) in Definition 2.1) is a true statement about the thesaurus $T = \langle E,W,C\rangle$ where E, W, and C are defined by

$$E = \{e_1,e_2,e_3,e_4,e_5\}$$

$$W = \{\{e_1,e_2\}, \{e_2,e_4\}, \{e_5\}\}$$

and $$C = \{\{e_1,e_3\}, \{e_4\}, \{e_2,e_5\}\}$$

It is easily seen that this is, in fact, the case. The interpretation I, together with each of the following definitions, determines an assignment of meaning to the derived terms (the underlined word) defined by those definitions:

Definition A: Given a thesaurus $T = \langle E,W,C\rangle$, a molecule is an element of the set $M = W \cup C$.

Definition B: Given a thesuarus $T = \langle E,W,C\rangle$, with $M = W \cup C$, let $M_1 = \{m_1,m_2,\ldots,m_n\} \subseteq M$. Then $\sigma(M_1) = m_1 \cup m_2 \cup \ldots \cup m_n$.

Definition A and the interpretation I determine the assignment of meaning to the derived term "molecule" in any Σ-statement containing that term. The molecules of T are the sets $\{e_1,e_2\}$, $\{e_2,e_4\}$, $\{e_5\}$, $\{e_1,e_3\}$, $\{e_4\}$, and $\{e_2,e_5\}$. I and Definition B uniquely assign a meaning to the notation $\sigma(M_1)$ given any set, M_1, of molecules. For example, if

$M_1 = \{\{e_1,e_2\}, \{e_2,e_4\}, \{e_4\}\}$, $\sigma(M_1)$ is $\{e_1,e_2,e_4\}$. Finally, given the Σ-statement

Theorem C: If $T = \langle E,W,C\rangle$ is a thesaurus and $W_1 = \{w_1,w_2,\ldots,w_n\} \subseteq W$,

$$|\sigma(W_1)| = \sum_{i=1}^{n} |w_i|,$$

which follows immediately from the Σ-postulate (i) through (v), the interpretation of Theorem C induced by I is the statement

"Given any collection $w_1,w_2,\ldots,w_n$ of the words $\{e_1,e_2\}$, $\{e_2,e_4\}$, and $\{e_5\}$, the number of elements in their union is equal to the sum of the number of elements in each set in the collection",

which is easily seen to be a true statement about the concrete thesaurus T. While we have illustrated with a trivial Σ-statement whose interpretation may seem obviously true, the same considerations apply to every Σ-statement in the development of the postulational system given in Definition 2.1, many of which are far from obvious. And because Σ as set forth in Definition 2.1 is a model of R.I.T. under the principal interpretation, I, given above, the interpretation induced by I (or any other interpretation under which Σ is a model of R.I.T.) of every theorem in the development of Σ is a true statement about R.I.T. (and every other object of which Σ is a model). We will denote the postulational system of Definition 2.1 by Σ_T in the following discussion.

3. A Complete Characterization of R.I.T.

Technically, a postulational system, Σ, is "complete" if any two instantiations of Σ are isomorphic. We define isomorphism of two thesauri in the following definition:

Definition 3.1

Given thesauri $T_1 = \langle E_1,W_1,C_1\rangle$ and $T_2 = \langle E_2,W_2,C_2\rangle$, T_1 is isomorphic to T_2, if there exists a one-to-one mapping Φ from $E_1 \cup W_1 \cup C_1$ onto

$E_2 \cup W_2 \cup C_2$ such that the restriction of Φ to each of E_1, W_1, and C_1 is a 1-1 correspondence between that set and E_2, W_2, and C_2 respectively, and such that for all $e \in E_1$, $w \in W_1$, and $c \in C_1$,

$$e \in w \cap c \Rightarrow \Phi(e) \in \Phi(w) \cap \Phi(c).$$

Clearly Σ_T is not complete. R.I.T. and the five-entry thesaurus in section 2 are two instantiations of Σ_T which are clearly not isomorphic. (Isomorphic thesauri have, among other similarities, the same number of entries.) A complete model of R.I.T. in this strict sense would not be particularly desirable. It would characterize R.I.T. too fully, down to the number of entries in each word and each category. We would rather construct a model, of which R.I.T. must be an instantiation, which has formal counterparts of all of the information carrying devices in R.I.T., but which leaves decisions concerning the use of those devices open. That is, we would like to allow as an instantiation of this new model a thesaurus constructed under the same "ground rules" as R.I.T. and having the same provisions, such as a hierarchy of categories, cross-reference numbers, the distinguishing of certain entries by printing them in bold face or some other method, etc., but in which different decisions were made as to which words to include in the thesaurus, the assignment of entries to categories, which entries to print in bold face, etc., I will call this "pseudo-complete" model of R.I.T. an "R.I. thesaurus".

We must first add to the definition of thesaurus a formal correlate of the hierarchy of categories in R.I.T. Semicolon-groups in R.I.T. are grouped together into sets of semicolon groups (the division referred to by the number after the decimal point in the index of R.I.T.), which are further grouped into larger groups (referred to in the index by the number preceding the decimal), etc. There is a total of six levels in

the hierarchy of categories in R.I.T. The upper levels are given by the "synopsis of categories" in the front of the thesaurus, We notice that the classificiation at each level is a partition of the set of categories at the next (lower) level. We therefore define a "multi-level thesuarus" as follows:

Definition 3.2

An n-level thesaurus is a quadruple $\langle E,W,C,H \rangle$ where

i) $\langle E,W,C \rangle$ is a thesaurus

and

ii) H is a collection of n-1 sets, $H_2,H_3,\ldots,H_n$, having the property that

a) H_2 is a partition of C

and b) H_k is a partition of H_{k-1} for all $2 < k \leq n$.

An n-level thesaurus for any positive integer n is a multi-level thesaurus. $\langle E,W,C \rangle$ is the base thesaurus of $\langle E,W,C,H \rangle$.

A thesaurus is a multi-level thesaurus with n = 1 (in which case H is null). R.I.T. is an instantiation of a 6-level thesaurus. It is also an instantiation of a 5-level thesaurus if C is interpreted to be the set of numbered "paragraphs" (the number after the decimal in the index) and E and W are interpreted as before. It is not, however, an instantiation of an n-level thesaurus for $n < 5$ if E and W have their "principal" interpretations. (With any broader interpretation of C, the fifth postulate is violated.) R.I.T. is an instantiation of an n-level thesaurus with $n < 5$, however, if E and W are allowed to have different interpretations (e.g., the interpretation under which elements of E are semicolon groups, elements of W are sets of identical semicolon groups, and elements of C are paragraphs--as sets of semicolon groups--or some larger sets of

semicolon groups. "Words" in such an interpretation would probably be unambiguous).

We must also include in the definition of an R.I. thesaurus formal correlates of various devices for distinguishing or labeling certain entries, categories, or elements of the sets $H_2,\ldots,H_n$ (bold face and italic type, information in brackets or parentheses following some entries, part of speech designations, etc.) and for pairing certain entries with categories or elements of the sets $H_2,\ldots,H_n$ (cross-reference numbers). This is done with functions from the sets E, C, $H_2,\ldots,H_n$ to "label-sets" L_E, L_C, L_{H_2}, etc. and relations from E to $E \cup C \cup H_2 \cup \ldots \cup H_n$.

Definition 3.3

An R.I. thesaurus is a sextuple $\langle E,W,C,H,L,R\rangle$ where

i) $\langle E,W,C,H\rangle$ is an n-level thesaurus for some positive integer n

ii) L is an n+1-tuple $\langle L_E,L_C,L_{H_2},\ldots L_{H_n}\rangle$ where for each subscript X, L_X is a finite collection, $\{(f_{X,1},L_{X,1}),\ldots,(f_{X,m},L_{X,m})\}$ of ordered pairs such that for each i, $f_{X,i}$ is a function from X to $L_{X,i}$

and

iii) $R \subseteq E \times (C \cup H_2 \cup H_3 \cup \ldots \cup H_n)$.

R.I.T. is an instantiation of an R.I. thesaurus under a suitable interpretation. Specifically,

1) E, W, C, and H are interpreted as before,

2) $L_E = \{(f_{E,1},L_{E,1}), (f_{E,2},L_{E,2})\}$

and a) $L_{E,1} = \{1,2,3\}$ and for all $e \in E$, $f_{E,1}(e) = 1$, 2, or 3 depending on whether e is printed in bold face, italics,

or regular type,

and b) $L_{E,2}$ is the set of distinct bracket or parenthesis labels which appear in R.I.T. together with the symbol ϕ and $f_{E,2}$ is the set of ordered pairs of the form (e,x) where e is followed by the bracket or parenthesis label x in R.I.T. (ϕ if e is unlabeled)

3) $L_{H_2} = \{(f_{H_2,1}, L_{H_2,1})\}$ where $L_{H_2,1}$ is the set of possible part of speech designations in R.I.T. and $f_{H_2,1}$ is the set of all ordered pairs of the form (h,p) where h is a paragraph in R.I.T. and p is the part of speech designation given h in R.I.T.

and

4) R is the set of ordered pairs of the form (e,x) which have the property that e is followed by the number of $x \in C \cup H_2 \cup \ldots H_n$ in R.I.T.

As a mathematical configuration, an R.I. thesaurus is too cumbersome to be fruitfully developed. Elements of an R.I. thesaurus may be combined, however, with results from the development of thesauri and multi-level thesauri to produce refinements of those results.

4. Directions of Possible Research

Research on thesauri, it seems to me, can profitably proceed in three major directions. First, the model can be further developed as a mathematical system. That development should be guided and motivated by the need to discover results which can shed light on possible applications of the thesaurus. But it is sometimes impossible to forsee which lines of development will yield the most beneficial outcome. Often results with important applications derive from a long line of blind development that does not seem promising. The fact that, under various interpretations,

a thesaurus is a graph means that the large body of graph theory results have important implications in the theory of thesauri. A number of these are discussed in (Bryan 73). In the following, certain concepts that were treated fully in that paper are referred to without definition.

As a second line of investigation, questions about the structure of the thesaurus can be posed in the framework of the model. The mechanical resolution of these problems may prove difficult because of the large size of the data base. Increased knowledge of abstract thesauri can be expected to lead to savings in the time and space required by these programs. Certain questions have answers wholly within the abstract system. It is easily shown, for example, that the "category-graph" of a thesaurus $\langle E,W,C\rangle$, in which vertices are categories and edges are pairs of categories (c_1,c_2) having the property that for some word, w, $c_1 \cap w \neq \emptyset$ and $c_2 \cap w \neq \emptyset$, shares properties concerning degree of connectivity with the corresponding "word-graph" of $\langle E,W,C\rangle$, and that it is therefore unnecessary to test certain hypotheses about one of those graphs if that information is already known about the other.

Finally, formal counterparts must be found to various semantic devices which the human user has at his disposal in using the thesaurus. Formost among these are measures on the sets E, W, and C, which assign "distances" to pairs of elements of those sets. Such formulated in terms of the lengths and/or number of "chains" connecting two elements of the set. It is clear that the length of "proper chains" connecting two entries has little to do with semantic relatedness since connecting any two homographs is a chain of length one. A major obstacle in efforts to find formal correlates to a number of semantic concepts is the fact that any link in a chain may be the semantically void link between homographs. The notion of "strong link" seems to go a long way

toward solving that problem. A category link, (c_1,c_2), is strong if $d(c_1,c_2) > 1$ where $d(c_1,c_2) = |r(c_1) \cap r(c_2)|$ and $r(c_1) = \{w \in W | c_1 \cap w \neq \emptyset\}$. Initial experiments have indicated that using a strong-link requirement as a filter very effectively weeds out semantically extraneous material from the output of "neighborhoods" of entries, words, and categories. This is important in light of the unmanagebly large outputs obtained when no restriction is placed on the expansion. A neighborhood expansion to radius four of the word "film", using as a data base the subthesaurus of R.I.T. consisting of all nouns which are one word in length, produced over 1000 words of output, much of which was completely unrelated semantically to the key word, and showed every indication of expanding to the whole subthesaurus in a relatively small number of steps. The same expansion allowed only along c-strong chains was completely free of semantically extraneous material and terminated in three steps. Most of the links allowed in the expansion were between clearly redundant categories. The exceptions were all between categories representing the literal and metaphorical use of a word.

Bryan, Robert, "Abstract Thesauri and Graph Theory Applications to Thesaurus Research" in S.Y. Sedelow, et al, Automated Language Analysis, 1972-1973, pp. 45-89.

SELECTED GRAPH THEORY APPLICATIONS TO A STUDY OF THE STRUCTURE OF ROGET'S THESAURUS: A DATA BASE FOR AUTOMATED LANGUAGE ANALYSIS

by Scott R. Taylor

TABLE OF CONTENTS

FORWARD

What follows are brief synopses of the chapters comprising this thesis. Each description consists of a brief discussion of the major points considered in each chapter. These synopses are in no way intended to summarize the research undertaken in conjunction with this thesis, but rather, they have been included in order to give the reader of this thesis an indication of the way in which it is organized and an indication of some of the major aspects of the research reported on.

Chapter I presents a very general introduction to the use of thesauri, and Roget's Thesaurus in particular, in automated language analysis. In addition, the questions concerning those aspects of the structure of Roget's Thesaurus which are considered in this thesis are briefly discussed, and the reasons for undertaking this study of thesaural structure are presented.

Chapter II consists of a discussion of Roget's Thesaurus in terms of its formal organization in the printed text. This discussion is intended to provide some insight into the characteristics of the structure of the Thesaurus, and specifically, the classification system used in the development of the printed version of the Thesaurus. Further, this chapter provides some background information necessary to understanding the way in which the graph representation of the Thesaurus is formulated.

Chapter III is concerned with a discussion of the theoretical foundations of this research and, more specifically, the graph theory concepts upon which the applications in this research are based. Included in this chapter are the formal definitions which are central to the concept of connectivity and the relationship of this concept to the graph structure proposed to represent the Thesaurus. The approach which is taken to the study of connectivity in this research effort is discussed in detail.

Chapter IV presents a discussion of possible alternative approaches to the study of thesaural connectivity in terms of changes in the way in which connectedness can be defined for the Thesaurus.

Chapter V deals with some of the semantic implications of the approach to the study of thesaural connectivity taken in this research effort. Further, attention is given to illustrating some of the many, and oftentimes complex, semantic relationships existing among words and groups of words in the Thesaurus. These semantic relationships necessarily must be considered when attempting to make conclusions concerning the structure of the Thesaurus.

Chapter VI is devoted to the presentation and discussion of the method developed for applying the graph theoretic techniques to the

study of the connectedness of the Thesaurus. The machine accessible version of the Thesaurus used in this research is briefly discussed, and the algorithm for the computer implementation of the method developed for analyzing the connectedness of the graph representation of the Thesaurus is presented. Included in this discussion is a look at the various programming considerations which were dealt with, including the problems which arose and the solutions to those problems.

Chapter VII presents some conclusions concerning the structure of the Thesaurus, and specifically, its connectedness characteristics. The experimental results generated by the computer program written to implement the method for analyzing the connectedness of the Thesaurus are presented, and some discussion is devoted to possible improvements in the method which was developed and its computer implementation. In addition, proposals for additional research and future extentions of this research are presented.

CHAPTER I

PURPOSES OF THIS STUDY

A recognition of the importance of and need for some means of determining the information content of written text has existed since the earliest attempts at utilizing the computer for language and information processing.[1] In such areas of research as automated document storage and retrieval, machine translation of texts, question-answering systems, and stylistic and content analysis of language and texts, the need for some "automatic" way of examining the meaning or content of natural language has been especially pronounced. An effective means of meeting this need has been the use of thesauri and dictionaries as stores of information from which sufficient semantic information can be extracted for successful content analysis. Thesauri, in particular, seem to be one of the most effective tools for use in analyzing content, and it is the "structure" of one such thesaurus, as affecting its use in information and language processing, with which this paper is concerned. Specifically, the thesaurus to be considered is Roget's International Thesaurus (1), which has been developed in computer-accessible form as a part of a project in automated language analysis.[2]

It is important at this point to explain in some detail what a thesaurus is, and how it can be used for the purposes of automatic information processing. As described by K. Sparck Jones, "a thesaurus in its widest sense is simply a classification of words by concepts, topics, or subjects. . ." (2). More simply stated in the introduction to Roget's Thesaurus, it consists of "words grouped by ideas." This grouping or classification of words into what Salton (3) describes as "concept classes" serves as the basis for what can be referred to as

---

[1] For a comprehensive review of efforts in automatic or computerized language and information processing refer to volumes of Carlos Cuadra's Annual Review of Information Science and Technology, which have appeared yearly since 1966.

[2] a) Throughout this paper Roget's International Thesaurus will often be referred to as Roget's Thesaurus or simply as the Thesaurus, which is distinguished from the generic usage of "thesaurus" by being capitalized and underlined.

b) This project, sponsored by the Office for Naval Research and under the direction of Dr. Sally Y. Sedelow, is being undertaken at the University of Kansas. A package of computer programs, referred to as VIA, has been developed for use in content analysis, and Roget's Thesaurus has been put into computer-accessible form to be used with these programs.

the structure of a thesaurus (in both a semantic and syntactic sense). It is the means by which these "concept classes" are constructed and arranged that largely determines the usefulness of a thesaurus to the human writer, as well as its usefulness as a base of information from which to process information mechanically. Oftentimes, a thesaurus is described as consisting of words grouped together on the basis of synonymy or near-synonymy.[3] As pointed out by K. Sparck Jones, this need not be the case, and although thesauri are often characterized in this fashion, I will avoid making such a distinction. Synonymy is a rather narrow way of describing the manner in which words are grouped in thesauri and synonymy, itself, is the center of considerable debate among linguists.[4] In addition, in defining relationships among words in thesauri, not only must synonymy be considered, but also such relationships as antonymy and homonymy. A more correct phrase for referring to the way in which words are grouped in thesauri would seem to be "conceptually related" or, prehaps more consisely stated, "semantically related." Although these terms may seem obscure to some, I will use them as a more general way of referring to relations of meaning among words. A thesaurus, then, through its structure and organization defines or establishes many of she semantic relationships existing among words in the thesaurus, with the existing "network" of relations determining in large part how a thesaurus can be used, as well as how effectively.

The use of a thesaurus, as a writer's aid, can be characterized as the process of taking a concept or idea and finding in the thesaurus the word which may be used to express the concept in written form. A distinction is often made between the use of "words as concepts" to locate other words to express the concept, or some particular aspect of the concept. Vickery (4) makes this distinction with the former words being classed as "idea-words" and the latter as "test-words," with the primary difference being in the function which the words serve when referencing a thesaurus. As this distinction relates to the use of Roget's Thesaurus, "idea-words" correspond to the terms which appear in the index to the Thesaurus and which are used in searching for semantically related words in the body of the Thesaurus, or the words corresponding to "text-words." Many of the words serve both purposes, the distinction being made with regard to the usage of a particular word. Also

---

[3] It should be explained at this point that I use "word" to refer to both words and phrases used in thesauri as basic "units of information." In addition, oftentimes when referring to these basic "units of information" the term "entry" or entries" will be used.

[4] For a thorough discussion of synonymy as it relates to the study of natural language, and as a basis for defining relations among words see K. Sparck Jones (2).

particular "text-words" may often be associated with several or many "idea-words," and vice versa. This situation is very important when considering the structure of a thesaurus, and specifically, Roget's Thesaurus.

The use of a thesaurus in information processing and retrieval systems can be characterized similarly. However, the method of using a thesaurus most often involves using specific words in a text or document to locate words in the thesaurus which will represent the idea more fully or completely. That is, the movement is from a specific word in a text to the concept or idea as expressed by a number of different words. In information retrieval systems, these words used in searching thesauri are usually referred to as "key-words" or indexing terms. Using these "key-words" it becomes possible to determine the concept being expressed by retrieving the semantically related words from the thesaurus and using these words in conjunction with words retrieved for other "key-words" in the text. Essentially, this process consists of determining the contextual use of words in the text by examining the overlap between the groups of words retrieved from the thesaurus to represent concepts. It is this one-to-many mapping of words in a text or document to words in a thesaurus which allows content analysis systems to determine effectively the thematic or conceptual content of the text or document.[5] The structure of a thesaurus, as evidenced in the relationships among words and word groupings within the thesaurus, therefore becomes an important consideration. Thus, structural analysis of thesauri is one way of gaining a greater insight into the way in which thesauri can be used effectively for content analysis. Likewise, the meaningfulness of the results obtained as a result of using thesauri should be more easily determined.

The primary thrust of this research is the application of certain graph theoretic techniques to an analysis of the structure of Roget's International Thesaurus. This thesis reports on the results of an investigation of the use of the techniques studied in determining certain specified characteristics of the structure of the Thesaurus when treated as an undirected graph. Specifically, the characteristics considered are those concerning the "connectedness" properties of the graph structure proposed to represent the Thesaurus. For the purposes of testing and demonstrating the computer implementation of the techniques studied, only a portion of the Thesaurus has been dealt with thus far. However, in discussing the application of these techniques and the implications of such structural analysis of the Thesaurus I will

---

[5] This method of using thesauri is discussed in depth by Salton (3), with an emphasis given to its applications in the SMART system. As it applies to the use of Roget's Thesaurus by the VIA programs refer to Automated Language Analysis, report on research, Sally Y. Sedelow, 1968, for a general description of its use and for further references. For a more detailed discussion of how the Thesaurus will be used with the VIA programs refer to reports on research for 1971 and 1972.

usually talk in terms of the entire Thesaurus.

A detailed and formal discussion of the structure of Roget's Thesaurus, both in the printed version and the proposed graph representation, will follow in later sections of this paper. Likewise, the means by which the graph theory concepts considered are applied to the graph representation of the Thesaurus are also explained in detail. For the moment, however, the "connectedness" of the Thesaurus, or some portion of it, can be defined in a very general way as follows. Starting with a given group of semantically related words, consider each word in this initial group individually and add to this grouping the "related" words found in other groups of semantically related words in the Thesaurus containing an occurrence of this word. Continue this process for all groups of words added to the initial grouping until, for each word in the final grouping, all occurrences in the Thesaurus of this word have been found and its "related" words added to the group. The question of connectivity to be considered can now be stated in the following manner. Will all of the words, or entries, in the Thesaurus be included in the final group of words found as described above? If such is the case, the Thesaurus, as represented as a graph, can be characterized as being "totally connected." Intuitively, this possibility may appear self-evident to some and unlikely to others, but no definitive statement has previously been made one way or the other. It is hoped that the results of this research will contribute to the making of such a statement.

In considering this question of "total connectedness" a number of other interesting questions arise which should be considered also. First of all, if it is the case that the Thesaurus is totally connected as described above, are there major groupings, or "clusters," of words or entries within the Thesaurus which can be considered as "central concepts" around which it is formed? And, if so, what are these "central concepts"? That is, do the words and concepts used seem to fall into identifiable, although interconnected, sub-groups? If such "clusters" are to be found, does there exist a correspondence between their arrangement and the classifications and/or the structural hierarchy imposed in the printed version of the Thesaurus? Secondly, if the Thesaurus is totally connected, are there some words or entries which occur so frequently as to serve as the primary "connecting elements," and which, if deleted, will cause the Thesaurus to become disconnected? That is, will the words or entries fall into or occur in only one of several (two or more) groups formed as described in the initial definition of connectivity. Likewise, supposing that frequently occurring words do not serve as primary "connecting elements," or appreciably affect the connectedness of the Thesaurus, can other groups of words or entries be identified that serve a similar binding function, and upon removal from the Thesaurus serve to disconnect it?

All of these questions have been raised in the context of the Thesaurus being totally connected. However, if it is the case that the Thesaurus is not totally connected, then again these questions can be raised with respect to the individual groupings, or totally connected

subsets, of entries found. Likewise, when considering only a portion of the Thesaurus, as was done for experimental purposes in this study, all of these questions are applicable to that portion of the Thesaurus considered. In fact, consideration of these questions of connectedness for those portions of the Thesaurus corresponding to formal divisions in its organization on an individual basis may reveal more completely the relationships between the structural characteristics of the Thesaurus in terms of connectedness and the conceptual organization imposed on it by the author(s). All of the questions concerning connectedness of the Thesaurus will receive a considerably more formal treatment in later sections of this paper.

The desirability and need for undertaking such an analysis of the structure of the Thesaurus is evidenced in the applications which the results of this research may have to its use in language analysis systems, and specifically, the VIA package of programs for which this thesaurus has been developed. Among the areas of general applications, as pointed out by Dillon and Wagner (5), are: 1) construction of thesauri, both manually and automatically, 2) modifications of existing thesauri, and 3) comparisons between thesauri.

The construction of thesauri by both manual and "automatic" methods is heavily dependent upon the human judgements of those involved in the construction. Decisions must be made as to what terms are to be included in or excluded from the thesaurus, how these terms should be grouped together, and how these groups should be arranged in order to provide the most meaningful organization, as well as to promote the most efficient use of the thesaurus. A wide range of criteria has been proposed for use in guiding the construction of thesauri (6,7,8,9). In most instances attention has been focused primarily on the criteria for compiling individual groups of conceptually or semantically related words, with little effort devoted to the overall organization of these groups within a structural framework. Structural analysis of Roget's Thesaurus in terms of its connectedness should help to provide insight into the way in which it has been organized, and thus, serve as a guide to those attempting to construct other thesauri. A number of methods for constructing thesauri for use with information processing systems have been proposed or implemented. Salton provides an extensive survey of these methods (3). The most promising methods for "automatically" constructing thesauri involve the use of "cluster analysis," whereby terms to be included in a thesaurus are clustered or grouped on the basis of the frequencies of word occurrences and the co-occurrences of words in a document collection in conjunction with the use of "similarity measures" devised to measure the degree of "relatedness" between words in a document. One such attempt at constructing a thesaurus in this manner is explained by Dattola and Murray (10). This method of thesaurus construction lends itself readily to analysis of connectivity characteristics, and, perhaps, the connectedness properties of Roget's Thesaurus can suggest improvements in the manner in which clusters are formed and organized within a thesaurus

constructed using this method.

In modifying existing thesauri it is important that any modifications undertaken be consistent with the overall structuring of the thesauri. This consideration is especially important when modifications are made to thesauri used in connection with automated systems for language and text analysis. Modifications must be consistent with the operation of the entire system in order that effective use can be made of the thesaurus. In general, treating a thesaurus as a graph structure and determining its properties of connectedness may suggest meaningful rearrangements, insertions, and deletions in the thesaurus. For example, some kinds of "conceptual gaps" may exist, whereby the inclusion of additional information may prove useful. Also, an unnecessary amount of redundancy may be evidenced, for which possible deletions might be desirable. Determination of the way in which words and phrases are grouped in a thesaurus and of the way in which these groups are connected or disconnected may suggest the means for incorporating these changes into the thesaurus, or at least suggest further research which might be useful in making such decisions. In addition, examining the properties of connectedness of the graph structure is one means of discovering the possible existence of some additional forms of structural and semantic biases which may affect the use of thesauri in general, and Roget's Thesaurus in particular. Another possibly helpful outcome of analysis of thesaural connectivity may be the assistance provided in developing techniques for optimizing the storage and search of thesauri and thesaurus-like structures used in automated language systems. One possible approach to the optimization of the storage and search of thesauri represented as graph structures is briefly discussed by Reisner (11). Likewise, some additional insight may be gained into the use and development of effective indexing schemes which may be constructed to exploit the connectivity characteristics of thesauri. With regard to the use of the Thesaurus by the VIA programs, totally connected subsets of entries represent groupings in the Thesaurus which, for some keyword, restrict the search for related entries to the particular totally connected subset of entries in which the keyword occurs. Thus, knowledge of the totally connected subsets of entries in the Thesaurus may suggest improvements in the way in which the Thesaurus is used by the VIA programs as well as suggest improvements in the Thesaurus itself. The existence of situations uncovered during the course of this study which may necessitate modifications in the Thesaurus will be discussed when the results of this research are presented in later sections of this report.

Properties of connectedness appear to be well suited criteria for making certain structural comparisons between thesauri. Such comparisons would be possible from both a semantic and syntactic viewpoint, and contribute to making decisions as to whether one thesaurus would be better suited than another for use with a particular language analysis system in light of the specific information needs of the system. For example, structural characteristics, as expressed in terms of thesaural connec-

tivity, may provide one means of comparing general-purpose thesauri, as represented by Roget's Thesaurus, with the use of special purpose or text specific thesauri. This means of comparing thesauri seems especially well suited for making comparisons between thesauri generated using methods of "cluster analysis" mentioned earlier.

CHAPTER II

THE FORMAL ORGANIZATION OF ROGET'S THESAURUS

As stated earlier, Roget's International Thesaurus consists of a classification of words and phrases into groups of semantically related terms. These groups are further organized by concept and may be referenced through the use of an elaborate alphabetical index to selected words and phrases occurring in the Thesaurus. Likewise, individual words and groups of words in the Thesaurus are linked to one another within the body of the text by the use of cross-references. The conceptual organization can be further characterized in terms of a hierarchy, presented in the "Synopsis of Categories" appearing prior to the body of the Thesaurus which attempts to define in part different levels of abstraction in the Thesaurus. This hierarchy is formed by grouping words and phrases which are semantically related, and in turn, grouping these groups of words in terms of the general concept to which they relate. These collections of groups are further organized in terms of more abstract concepts, and so on, until, at the highest level of abstraction, there exist eight general "concept classes." This hierarchy, and the cross-referencing capabilities provided by the index and by entries within the Thesaurus itself, serve to define in large part what can be termed the "semantic structure" of the Thesaurus.

This organization of the Thesaurus is also responsible for imposing a "syntactic structure" on the Thesaurus. In addition, further syntactic organization results from the grouping of words at a "lower" level of the Thesaurus by the part of speech which the words serve (e.g., as nouns, verbs, etc.). In summary, then, the grouping of words is governed by both semantic and syntactic considerations. These structural distinctions are important in a study of thesaural connectivity since they provide differing means of defining connectivity as relating to the Thesaurus and they also affect the interpretation of the results of such a study.

The formal hierarchy of the Thesaurus, as briefly described above, can be represented as a tree structure (see FIGURE 1), with the root of the tree denoting the Thesaurus as a whole, and each level of the tree

A Partial Tree Representation of the
Hierarchy of Roget's Thesaurus

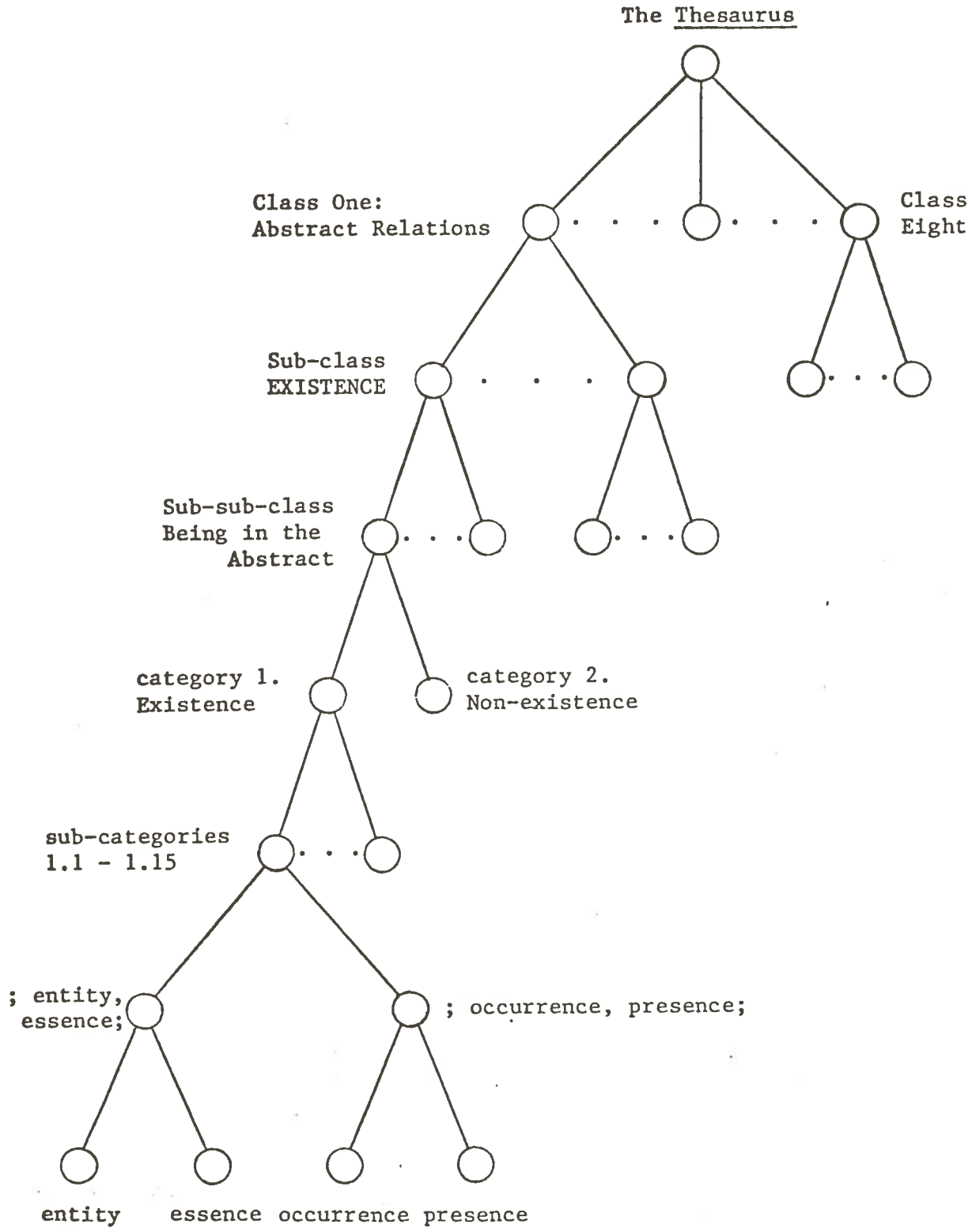


FIGURE 1

corresponding to one level in the hierarchy.[1] There are seven levels in the formal hierarchy. Each level of the hierarchy corresponds to a particular "conceptual partitioning" of the Thesaurus, with successively higher levels representing conceptually more inclusive divisions of the Thesaurus. The divisions may be identified, in the order of most inclusive to least inclusive, as follows: Classes, Sub-classes, Sub-sub-classes, Categories, Sub-categories, semi-colon delimited groups of entries,[2] and individual entries. Each of the eight general Classes are divided into sub-classes, which are in turn divided into sub-sub-classes, and so on. The first three levels of this hierarchy are not explicitly represented in the Thesaurus itself, but rather serve more as a guide to the way in which categories are arranged or organized. Each of the categories and sub-categories in the Thesaurus are labeled for identification.

There are 1,040 categories in the Thesaurus, and each category is numerically labeled consecutively as it appears in the Thesaurus. Likewise, the sub-categories are numerically labeled within the category in which they are placed. It is the use of these labels which provides access to the Thesaurus via the index, and allows for cross-referencing within the Thesaurus between categories and sub-categories. Using the example given in the tree representation in FIGURE 1, it can be seen that the entry "essence" is grouped with the entry "entity" to form a single semicolon group within sub-category 1.1, which falls within category 1. (Existence). This category is in turn classed in the "Synopsis of Categories" under the sub-sub-class "Being in the Abstract." Further, the sub-class containing this sub-sub-class is EXISTENCE, which is found in Class One: Abstract Relations. It is possible to define an additional level in the Thesaurus consisting of the groups within sub-categories as designated by the part of speech of the words in the sub-categories. For the purpose of this attempt at defining the structure of the Thesaurus this extra level provides little additional information necessary for understanding the hierarchical structure of the Thesaurus, and has been omitted. However, it is possible, as will be discussed later, that such an additional level designation may provide another means of looking at connectivity, and possibly influence the connectedness of the Thesaurus. In addition, Warfel (12) proposes that another

---

[1] For further discussion of the hierarchy of Roget's Thesaurus see Dillon and Wagner (5), and Warfel (12).

[2] If a group falls at the beginning or end of a sub-category, this way of referring to these word groupings is not totally correct, since they may also be delimited by a period. To be consistent, however, and to avoid confusion, I will always refer to these as semicolon delimited groups, or simply as semicolon groups.

level of the hierarchy may be defined, whereby categories are further organized within each sub-sub-class into groups which represent more closely semantically related categories within a sub-sub-class. It appears that in most cases these additional groups would consist of categories related in terms of negation or opposition. Warfel gives as an example four categories in the sub-sub-class RECURRENT TIME which are labeled with "Frequency," "Infrequency," "Regularity of Recurrence" and "Irregularity of Recurrence," and proposes that the former two categories are more closely related to one another semantically than to either of the other two, and thus, should be grouped together under one designation, whereas the latter two categories are closely related and likewise should be grouped together. Again, such an additional classification within the Thesaurus would provide still another means of defining the connectedness of the Thesaurus.

In considering the structure of the Thesaurus it is important to consider the role of cross-referencing between entries and groups of entries. Examination of the tree representation in FIGURE 1 shows that, with the addition of cross-references between entries in different categories and sub-categories, new arcs can be added to the tree. The additional arcs serve to create new ways of traversing the structure, and afford one way of gaining an understanding of the concept of connectivity as it relates to thesauri treated as graph structures. Although cross-referencing does not directly contribute to the study of connectivity undertaken in this research, its relation and application to this study will be firmly established later in this report. For a very comprehensive and detailed analysis of cross-referencing in thesauri and other indexing schemes see Kochen and Tagliacozzo (13). They define the "cross-reference structure" existing in thesauri and indexes, and model this structure as a graph, where the "see" and "see also" references often found in thesauri serve as arcs connecting terms designated as nodes of the graph. Among the characteristics of this structure which are considered are the degree of "connectedness" and "accessibility" of terms in the cross-reference structure. The degree of "connectedness" and "accessibility" are used primarily to describe statistical relationships among words and phrases in an index or thesaurus. The degree of "connectedness" refers to the "relative number of terms linked to other terms in the same index, as compared to the number of isolated index terms." Although this usage of connectedness differs from the way I define it in this study, the implications of this usage are similar. The degree of "accessibility" refers to "the average number of different paths (cross-references) which lead to an index term." This measure corresponds to the average number of arcs incident to a node of a graph. As a part of my research, and in an effort to determine those groups of entries which may serve as "primary connecting elements" (as described in Chapter I), the exact number of arcs incident to each node of the graph representation of the Thesaurus is computed. A formal discussion of this aspect of my research follows in Chapter III.

For the purposes of my research the structural relationships of primary interest are those existing at the sixth level of the hierarchy among the semicolon delimited groups of words and phrases. The properties of connectedness considered are those which exist at this level, and it is necessary that some discussion be devoted to this level of the Thesaurus. However, it should be stressed that the thesaural connectivity exhibited at this level necessarily must be viewed in the context of the total hierarchy. If the Thesaurus is totally connected at the sixth level it must necessarily be so at higher levels, although the reverse may not be true. Within categories in the Thesaurus words and phrases which are most closely related semantically are grouped together to form semicolon delimited groups. These semicolon groups are further grouped to form sub-categories. This process continues, as described earlier, to generate the hierarchical structure of the Thesaurus. As an example of this organization at the three lowest levels of the hierarchy consider the following two sub-categories, 1.1 and 1.2, which occur in category 1.

1. EXISTENCE
NOUNS 1. existence, subsistence, being; entity, essence; occurrence, presence; life 407.
2. reality, actuality, factuality; truth 515; authenticity 515.5; sober or grim reality, no joke, not a dream; ultimate reality, thing-in-itself (philos.).

It is supposed, for example, that the entries "existence," "subsistence," and "being" are most closely related to one another, and thus, grouped together to form one semicolon group. Likewise, the entries "entity," and "essence" are most closely related to each other, and are therefore grouped together, whereas the entries "being" and "entity," although related under the concept of "existence," are not as closely related to one another as they are to the entries in their respective semicolon groups. Thus, they appear in separate semicolon groups within the same sub-category. The entries "truth" and "essence," however, appear in separate sub-categories and, although related under the concept of "existence," they are not as closely related to each other as the entries in their respective sub-categories.

At this point a comment should be made about one aspect of the conceptual organization of the hierarchy exhibited by this example. Category 1. is labeled as EXISTENCE to designate that entries in the category are all related in some way under the broad concept of "existence." Note, however, that there is an entry "existence" in sub-category 1.1 which, as explained above, is grouped with the entries "subsistence" and "being" into a single semicolon group. The distinction between the use of the concept of "existence" to group all entries in the category and the seemingly more restricted use of the word "existence" in forming one semicolon group within the category may seem unclear. However, this situation, as discussed briefly in Chapter I, must be

recognized in terms of the use of "words as concepts" or "idea-words" and the use of words as "text-words" to represent concepts. Although possibly responsible for some confusion, this aspect of the conceptual organization of the Thesaurus is an important characteristic to be considered when discussing the structure of the Thesaurus. In any case, for the purposes of this research it is very important to have some understanding of the way in which semicolon groups have been formed and are organized in the Thesaurus in order to understand fully the way in which connectivity can be defined between semicolon groups. This definition will be explained in detail in the following section of this report.

Upon further examination of the above example some understanding should also be gained of how cross-referencing is performed within Roget's Thesaurus. Those numbers appearing within semicolon groups are the labels of categories and sub-categories which contain entries that are related to the entry or entries in the semicolon groups. For example, in sub-category 1.2 the semicolon group consisting of the entry "truth" is linked to category 515, which is labeled "TRUTH." Thus, an additional link in the thesaural structure is defined.

An understanding of the structural characteristics of the Thesaurus is necessary in order to grasp the concept of connectivity as it relates to this structure. The connectedness of the Thesaurus can be approached from a number of different viewpoints. Briefly, these approaches deal with the connectedness exhibited at each of the different levels of the formal hierarchy, and the effects of cross-referencing on the connectedness of the Thesaurus. In addition, a variety of restrictions can be imposed on the definition of connectivity to change the manner in which thesaural connectivity is viewed, and thus, the information which can be gained from the study of connectedness. Ways in which the definition of connectivity may be altered when studying the structure of the Thesaurus will be explained in Chapter IV. However, as mentioned earlier, this study deals with the connectedness between the semicolon delimited groups of entries comprising the Thesaurus, where these semicolon groups, and sets of semicolon groups, are interpreted as the nodes of a graph and the arcs between nodes represent the existence of identical entries in the semicolon groups. And, as a demonstration of the computer program written to implement the graph analysis techniques studied, the connectedness of only a small portion of the Thesaurus has been determined thus far. Specifically, the portion of the Thesaurus which has been dealt with is that portion corresponding to the first two sub-classes of Class One. Sub-classes I and II of Class One are made up of 28 of the 1,040 categories in the Thesaurus, and encompass 3,375 of the estimated 200,000 entries in the Thesaurus.

CHAPTER III

THE THEORY: THESAURAL CONNECTIVITY

Presented in this section are the primary theoretical foundations for this research. Much of the theoretical basis for this research is a direct result of the efforts of Robert Bryan at the University of Kansas to develop a model of abstract thesauri, of which Roget's Thesaurus serves as one instantiation (14). This model lends itself readily to the graph theory applications with which this research is concerned. Many attempts at automatic construction and structural analysis of thesauri and thesaurus-like structures have relied on representations of thesauri as graph structures. A quite extensive collection of papers dealing with selected areas of information science, many of which consider various graph theory applications to thesauri and thesaurus-like structures, has been assembled by Kochen (15). As stated earlier, my own research effort is concerned with the connectedness of Roget's Thesaurus when treated as an undirected graph. A number of definitions central to this concept of connectivity and the relationship of this concept to the graph structure proposed to represent the Thesaurus will be given in this chapter. All of these definitions and the related terminology are basic to most applications of graph theory and, except as relating to the Thesaurus, it is assumed that the reader of this section of the report has some familiarity with them. For more complete discussions of many of these definitions see (16,17,18).

As defined by Bryan, a "thesaurus" can be represented on an abstract level as a triple (E,W,C), where:

1) E is a non-null, finite set whose elements are called entries.
2) W and C are non-null partitions of E, where the elements of W are called words and the elements of C are called categories.
3) For any entry $e_i \in E$, $e_i \in w_k$ and $e_i \in c_j$ for some $w_k \in W$ and some category $c_j \in C$.
4) For a given word $w_k \in W$ and category $c_j \in C$, $o(w_k \cap c_j) \leq 1$, where $o(S)$ is the cardinality of the set S, or the number of elements in S.

Now, Roget's Thesaurus is a "thesaurus" in this sense when E, W, and C are interpreted as follows:

1) E is the set of all entries of the Thesaurus, where an entry is a single occurrence of some character string which is delimited by commas and/or a period or semicolon.[1]

---

[1]This definition of entry should be interpreted so as to allow for those special cases where an entry may contain "internal commas," but whose use is intended as a single unit of information.

2) W is the set of all words in the Thesaurus, where a word refers to the set of all occurrences of each distinct character string. That is, a word is the name given to each set of identical entries in the Thesaurus. For example, the character string "house" might appear in the Thesaurus as several different entries, but all of these entries comprise a set which is referred to as the word "house." $W = \{w_1, w_2, \ldots, w_n\}$, where the particular w_i representing the word "house" is a set comprised of all occurrences of the entry "house."
3) C is the set of all categories comprising the Thesaurus, where, for the purposes of this research, category refers to each semicolon delimited entry or group of entries.[2] That is, $C = \{c_1, c_2, \ldots, c_n\}$, and each c_i is the set of entries comprising a semicolon group. An example of a category which is taken from *sub-category* 697.17 of the Thesaurus, follows:

; harbor, haven, house, nestle;

In this example the character strings "harbor", "haven", "house" and "nestle" are all entries in the Thesaurus. The distinction between "word" and "entry" is similar to that distinction represented by the type-token relationship often used in language analysis, where a word corresponds to a type and each entry corresponding to this word is a token.

It is now possible to state that the entries comprising two categories are related, or connected, by defining a relation R on the set of categories C as follows:

Given two categories c_i, $c_j \in C$, c_i R c_j if $\exists$ a finite sequence $c_1, c_2, \ldots, c_k$ of elements of C. referred to as a chain of categories, such that:
1) $c_1 = c_i$ and $c_k = c_j$.
2) For all m, $1 \leq m < k$, there exists some $w_n \in W$ such that $c_m \cap w_n \neq \emptyset$ and $c_{m+1} \cap w_n \neq \emptyset$.
3) For all i and j, $1 \leq i \neq j < k$, $c_i \neq c_j$.

Essentially, then, two categories can be said to be related if there exists a chain of categories connected in such a way that for any two "adjacent" categories in the chain each eategory possesses at least one identical entry in common which serves to relate, or connect, the two categories "conceptually." For example, consider the following four categories from the Thesaurus.

---

[2] It is important that this usage of the word "category" not be confused with that of Chapter II, where "category" referred to one of the formal divisions within the hierarchy of the Thesaurus. To avoid confusion those references to divisions of the formal hierarchy of the Thesaurus will be *italicized* in the remainder of this paper.

c_1 ; asylum, haven, port, harbor;
c_2 ; asylum, home;
c_3 ; home, house, cabin, domicile, domiciliate;
c_4 ; harbor, haven, house, nestle;

It can be seen that for these categories c_1 R c_2, since the entry "asylum" is found in both categories c_1 and c_2. Likewise, c_1 R c_4, since the entries "harbor" and "haven" occur in both c_1 and c_4. The existence of the relation R between these categories establishes two sequences, or chains, of categories; c_1, c_2 and c_1, c_4. Likewise, there exists another chain of categories (c_1,c_2,c_3,c_4) also connecting c_1 and c_4, where the entry "asylum" appears in both c_1 and c_2, the entry "home" appears in both c_2 and c_3, and the entry "house" appears in both c_3 and c_4. Thus, the "conceptual" connection, or link, between categories c_1 and c_4 is established. Actually, then, what has been found is a chain of conceptually related entries; $asylum_1$, $asylum_2$, $home_1$, $home_2$, $house_1$, $house_2$. Or in terms of categories, c_1,c_2,c_3,c_4. In establishing the existence of such chains of categories an additional restriction is imposed, which is characterized in part three (3) of the definition of R given earlier; category chains must be "nonrepeating." An example of a "repeating" chain would be the chain of categories c_1,c_2,c_3,c_4,c_1,c_4. This chain is repeating since the categories c_1 and c_4 occur more than once in the sequence of categories comprising the chain. The reason for the restriction that chains be nonrepeating is that such situations introduce no new information to consider. Further, for any two categories c_i,$c_j \in C$, which are connected by a repeating chain of categories, there also exists a nonrepeating chain connecting c_i and c_j. Note, however, that this restriction applies only to repeating categories, and it is permissible for identical entries to occur any number of times in the chain of related entries corresponding to some category chain. That is, for the chain of categories c_1,c_2,c_3,c_4 constructed earlier, it could have been the case that the word "home" intersected all of the categories in the chain. In which case, one chain of entries corresponding to c_1,c_2,c_3,c_4 would have been $home_1$,$home_2$,$home_3$, $home_4$.

Further, the relation R defined above is an equivalence relation on the set of categories C. That is, the relation R satisfies the following three conditions:

1) For any category $c_i \in C$, c_i R c_i is true. Thus, the property of reflexivity exists for all categories of the Thesaurus.
2) For any two categories c_i and $c_j \in C$, if c_i R c_j is true for some chain c_i,c_{i+1},...,c_j, then c_j R c_i is true for the chain c_j, c_{j-1},..,c_i. Thus, the property of symmetry holds for this definition of R.
3) If c_i R c_j for some chain c_i, c_{i+1},...,c_j and c_j R c_k for the chain c_j,c_{j+1},...,c_k, then c_i R c_k is true for some chain c_i,...,c_j,...,c_k. Thus, the property of transitivity

holds true for R.[3]

Hence, in finding those chains of categories determined by an arbitrary category $c_i \in C$, an equivalence class of categories determined by c_i is established. An equivalence class determined by some $c_i \in C$ is, in effect, the set of categories which are connected to c_i by a chain. The set of all distinct equivalence classes determined by elements of C is a partition of C. The set of distinct equivalence classes is a partition of C since the union ($\cup$) of these equivalence classes is the set C, but any two distinct equivalence classes are disjoint. Formally, two equivalence classes are disjoint if, for any category c_i in one equivalence class and c_j in the other, the intersection of c_i and c_j is null (i.e., $c_i \cap c_j = \emptyset$).

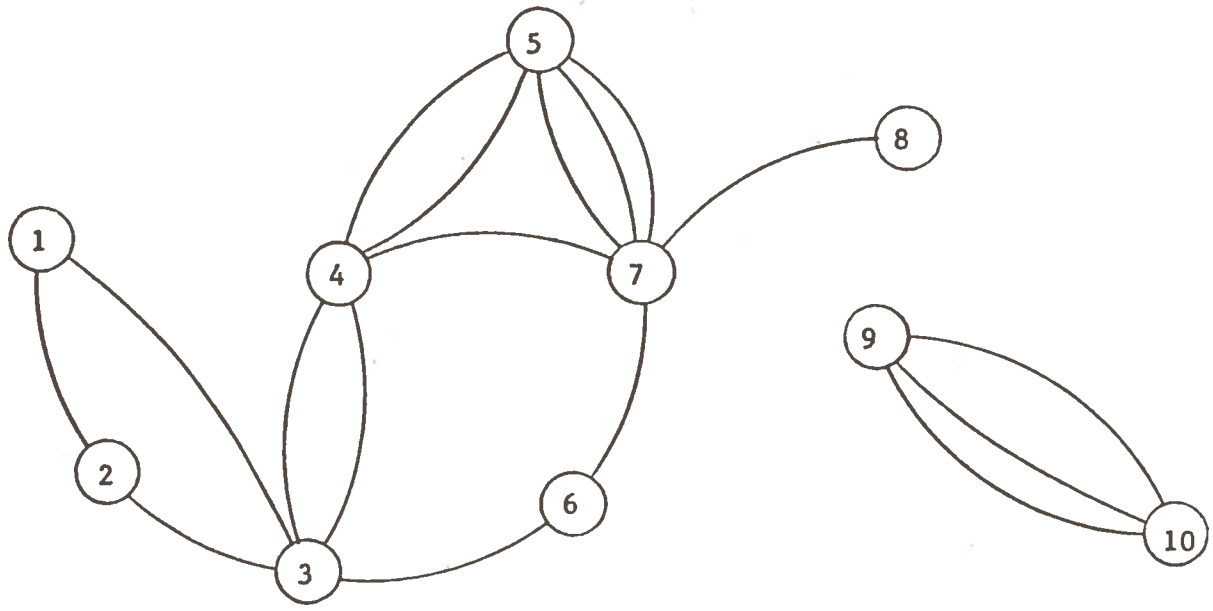
The Thesaurus can now be represented as an undirected graph, where the nodes of the graph correspond to the categories in the set C, and the non-null intersection of two categories (at least one identical entry occurring in each) is represented in the graph as an undirected arc between the two nodes of the graph representing the categories. As an example of how this representation relates to the Thesaurus refer to FIGURE 2. Each of the categories listed in FIGURE 2 is represented by a node in the graph. Note, however, that some of the categories have more than one entry in common, or occurring in each. As a result, the corresponding nodes of the graph have more than one arc connecting them.

Each equivalence class of categories corresponds to a subgraph of the graph representation of the Thesaurus. Further each subgraph corresponding to an equivalence class can be characterized as totally connected. That is, for any two categories c_i and c_j in each equivalence class c_i R c_j is true. Or, in other words, for any two categories in an equivalence class there exists a chain of categories which serves to connect the two categories. As this applies to the graph representation, a graph is totally connected if, for any two nodes in the graph, there exists a path (or sequence of nodes connected by arcs) which connects the two nodes. A distinct equivalence class of categories, then, corresponds to a maximal totally connected subgraph of the graph representation of the Thesaurus. A totally connected subgraph of a graph is maximal when it is a subgraph of no other totally connected subgraph. A maximal totally connected subgraph of a graph is often referred to as a component of the graph. As can be seen in FIGURE 2, this graph does not satisfy the definition of totally connected, although there are two subgraphs of the graph, comprised of nodes 1 through 8 and nodes 9 and 10, which are totally connected. Each of the sets of categories corresponding to these subgraphs is an

---

[3]Note, however, it may be the case that, although the two chains of categories, $c_i, c_{i+1}, \ldots, c_j$ and $c_j, c_{j+1}, \ldots, c_k$, are nonrepeating, the chain $c_i, \ldots, c_j, \ldots, c_k$ is repeating. However, as stated previously, for any repeating chain connecting two categories it is possible to construct a nonrepeating chain connecting the same two categories.

An Example of the Graph Structure Proposed to Represent the Relations Among Selected Categories from Roget's Thesaurus



node	corresponding category	
1	c_1	port, seaport
2	c_2	port, porthole
3	c_3	asylum, haven, port, harbor
4	c_4	harbor, haven, house, nestle
5	c_5	house, domicle, domiciliate, hovel
6	c_6	asylum, home
7	c_7	home, house, cabin, domicile, domiciliate
8	c_8	cabin, stateroom
9	c_9	homelike, homish, homey, homely
10	c_{10}	homelike, homey, homely, lived-in

FIGURE 2

equivalence class as it has been defined. Further, if the categories listed with FIGURE 2 were taken to define a thesaurus, then it can be seen that the thesaurus would be comprised of two distinct equivalence classes. Thus, the thesaurus would not be totally connected.

Associated with each node of a graph is what is referred to as the degree of the node. The degree of a node is the number of arcs incident with the node, or the number of arcs connecting the node and other nodes in the graph. For example, node 3 of the graph of FIGURE 2 has a degree of five (5). This measure of degree is similar to the "accessibility" of a node mentioned in Chapter II, and may provide some indication as to which nodes may cause a graph to become disconnected upon their deletion from the graph. That is, it may be the case that nodes with a relatively large number of arcs incident with them serve as the "primary connecting elements." Formally, those nodes in a connected graph which cause the graph to become disconnected upon their deletion comprise what is referred to as a cut set of nodes of the graph. It may be found that there exists a correspondence between nodes with a relatively large degree and those nodes comprising a cut set of nodes of the graph representation of the Thesaurus, although this need not be the case. In any case, nodes with relatively large degrees should give an indication of possible clusters of nodes in the graph, which correspond to groups of categories in the Thesaurus serving as central concepts (as mentioned in Chapter I), around which the Thesaurus is organized. Such information should prove useful when using the Thesaurus for automated language analysis.

In attempting to analyze the structure of the Thesaurus by automatic means, working with a graph representation in which each node of the graph corresponds to a single category, or semicolon group, involves a considerable amount of searching of the categories in order to determine the existence of connections. This process involves working with character strings which are of variable length. And, making comparisons between character strings, which may be quite long, is a time consuming procedure. Likewise, these character strings may require large amounts of storage. The process of constructing and working with a graph representation of the Thesaurus can be simplified, however, if some means of minimizing the search and storage of categories and entries can be found. Given the format of the machine accessible version of the Thesaurus used in this study, one method of simplifying construction and analysis of a graph representation of the Thesaurus is to construct what Harary (18) calls an "intersection graph." As will be described in Chapter VI, each entry of the Thesaurus can be identified by a unique numerical code assigned to the category in which it occurs. As a result, it is possible to construct a graph representation of the Thesaurus using the identifying codes of the categories.

An intersection graph can be defined as follows: given some set S, and a set F of distinct subsets of S, $F = \{s_1, s_2, \ldots, s_n\}$, whose union is the set S, each of these subsets can be interpreted to be a node of the graph. Further, for any s_i and s_j in S whose intersection is non-null

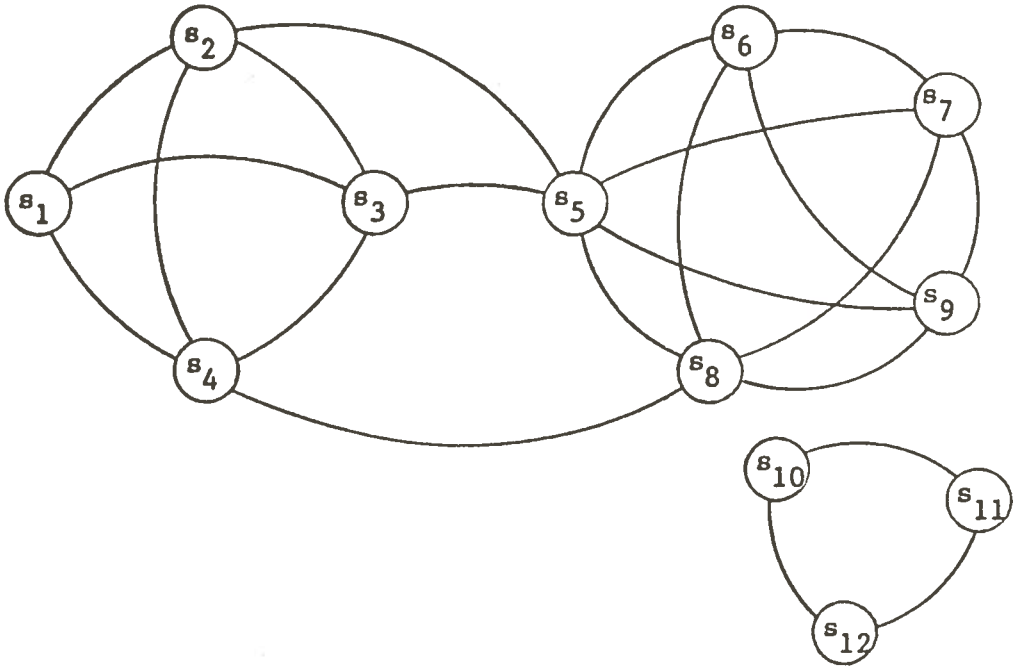
(i.e., $s_i \cap s_j \neq \emptyset$) an arc is established between the corresponding nodes of the graph. The graph representation of the Thesaurus, as illustrated in FIGURE 2, is, itself, an intersection graph, whereby the set S is interpreted as the set of entries and the set F as the set of categories. However, a new intersection graph can be constructed which will retain the connectedness characteristics of the original graph, but, at the same time, be more easily worked with. Specifically, this new intersection graph is based on an interpretation of the set S as the set of categories in the Thesaurus, each identified by a unique numerical code, and the set F as the set of subsets of categories corresponding to totally connected subgraphs of the original graph representation, as given in FIGURE 2. For the purposes of this study, the totally connected subgraphs used in constructing this intersection graph are those corresponding to the subsets of intersecting categories in the Thesaurus "induced" by each word, w_i, for which one of the following two restrictions is satisfied:

1) at least two entries belong to the set w_i (i.e., $o(w_i) > 1$).
2) if the set w_i consists of only one entry (i.e., $o(w_i) = 1$), then each of the entries in the c_i which intersects w_i must belong to a w_j containing only one entry.

Given the set of entries comprising some $w_i \in W$, the subset of all categories which contain an occurrence of one of the entries belonging to w_i corresponds to a totally connected subgraph of the graph representation of the Thesaurus in FIGURE 2. Further, each of these totally connected subgraphs can be characterized as complete. An undirected graph, or subgraph, is complete when there exists an arc connecting every pair of nodes in the graph. Isolated nodes, or nodes not connected to any other nodes of the graph, are assumed to be complete also. An isolated node, which is characterized in the second restriction above, corresponds to a category in which each of the entries, e_i, in the category is the only element of its corresponding w_i. That is, for each entry in the category there exist no identical entries elsewhere in the Thesaurus, and, as a result, no connections with other categories will exist. The first restriction given above eliminates consideration of entries which occur only once in the Thesaurus, but do not occur in categories that are isolated. Such entries do not affect the connectedness of the Thesaurus,as defined for this study, since they result in no connections between categories. Referring to the graph in FIGURE 2, nodes 4, 5 and 7 comprise a complete subgraph induced by the word "house." Likewise, nodes 1, 2 and 3 comprise a complete subgraph induced by the word "port." In fact, every pair of nodes in the graph which are connected by an arc comprise a complete subgraph, although not necessarily the complete subgraph induced by some w_i in W.

The intersection graph which can be constructed (as described in the preceding paragraph) for the set of categories given in FIGURE 2 is shown in FIGURE 3. Likewise, the subsets of categories corresponding to the complete subgraphs of the graph of FIGURE 2 induced by the set of words are listed in FIGURE 3. Note, that some of the subsets of categories have

Intersection Graph Constructed from the Graph of FIGURE 2



subsets of categories	word
$s_1 = \{c_1, c_2, c_3\}$	port
$s_2 = \{c_3, c_4\}$	haven
$s_3 = \{c_3, c_4\}$	harbor
$s_4 = \{c_3, c_6\}$	asylum
$s_5 = \{c_4, c_5, c_7\}$	house
$s_6 = \{c_5, c_7\}$	domicile
$s_7 = \{c_5, c_7\}$	domiciliate
$s_8 = \{c_6, c_7\}$	home
$s_9 = \{c_7, c_8\}$	cabin
$s_{10} = \{c_9, c_{10}\}$	homelike
$s_{11} = \{c_9, c_{10}\}$	homey
$s_{12} = \{c_9, c_{10}\}$	homely

FIGURE 3

more than one category in common, or occurring in each. As a result, there are multiple arcs connecting the corresponding nodes of the intersection graph which is constructed from these subsets of categories. These additional arcs, however, do not affect the connectedness of the graph as defined here, and, for the purposes of illustration, multiple arcs between nodes have been replaced by a single arc. Such a graph representation is referred to as a "reduced" graph. Upon further examination of the categories given in FIGURE 2 it can be seen that some categories have occurrences of more than one word in common. For example, categories c_9 and c_{10} in FIGURE 2 are connected as a result of three words ("homelike," "homely" and "homey") intersecting the categories. As a result, assuming that the categories in FIGURE 2 define a thesaurus, the subsets of intersecting categories induced by the words in this thesaurus may not always be distinct. This situation results in a larger number of nodes in the intersection graph which is constructed, as shown in FIGURE 3, than would be necessary. In order to remedy this situation it would be necessary to consider only the distinct subsets of categories induced by the words in the thesaurus. Referring to FIGURE 2, the categories comprising these distinct subsets correspond to the nodes of the maximal complete subgraphs of the graph. And, those complete subgraphs which are not maximal need not be represented as separate nodes of the intersection graph which is initially constructed. The subsets of categories corresponding to the nodes of the graph of FIGURE 2 which form maximal complete subgraphs are as follows:

$$s_1 = \{c_1, c_2, c_3\}$$
$$s_2 = \{c_3, c_6\}$$
$$s_3 = \{c_3, c_4\}$$
$$s_4 = \{c_4, c_5, c_7\}$$
$$s_5 = \{c_6, c_7\}$$
$$s_6 = \{c_7, c_8\}$$
$$s_7 = \{c_9, c_{10}\}$$

The graph of FIGURE 2 is shown again in FIGURE 4.a, with each of the maximal complete subgraphs circled with a broken line. The intersection graph which can be constructed from this graph is shown in FIGURE 4.b. Each of the nodes of the graph of FIGURE 4.b is labeled with the s_i, as given above, for the subset of categories which it represents.

The complete subgraphs induced by the words of the Thesaurus oftentimes will be maximal, as with the complete subgraph induced by the word "port" in categories c_1, c_2 and c_3 of FIGURE 2. This is not always the case, however, as with the complete subgraphs induced by the words "homely," "homey" and "homelike" in categories c_9 and c_{10}. Although the intersection graph shown in FIGURE 3 is larger and more complicated than the graph of FIGURE 2, this should not be interpreted as particularly indicative of the size and complexity of the intersection graph which is constructed for the Thesaurus, since some of the categories considered in these examples have a relatively large number of words in common, thus

Construction of the Intersection Graph from the Maximal Complete Subgraphs of the Graph of FIGURE 2

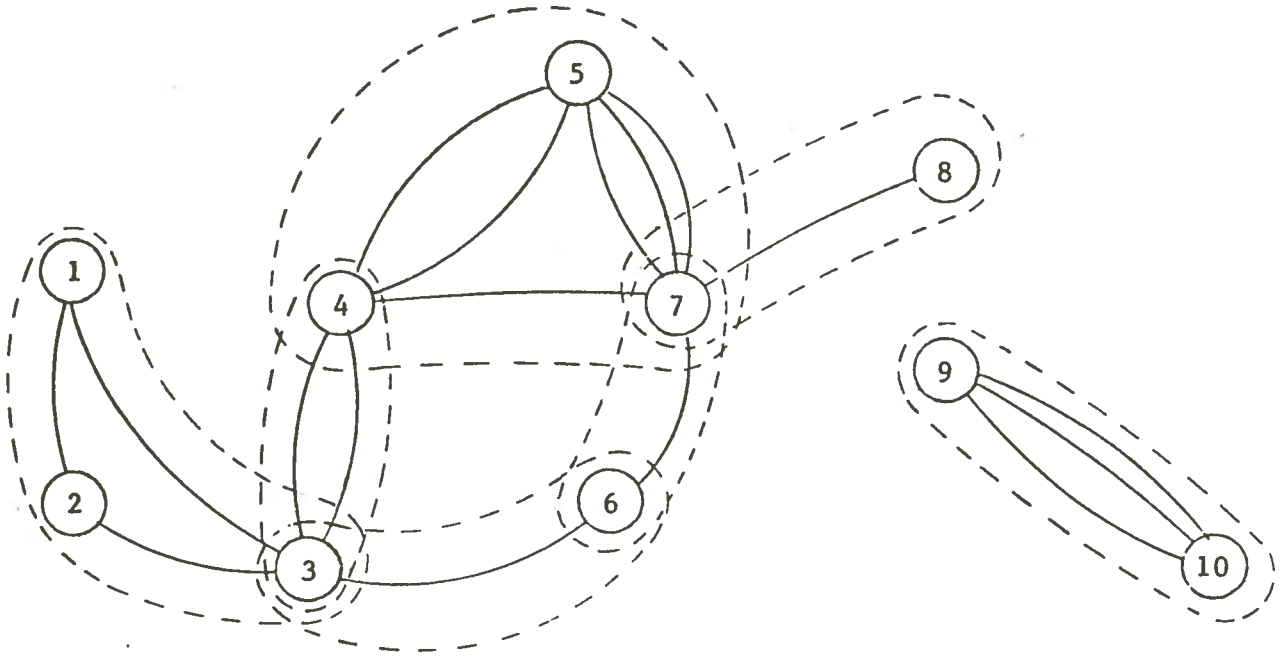


FIGURE 4.a

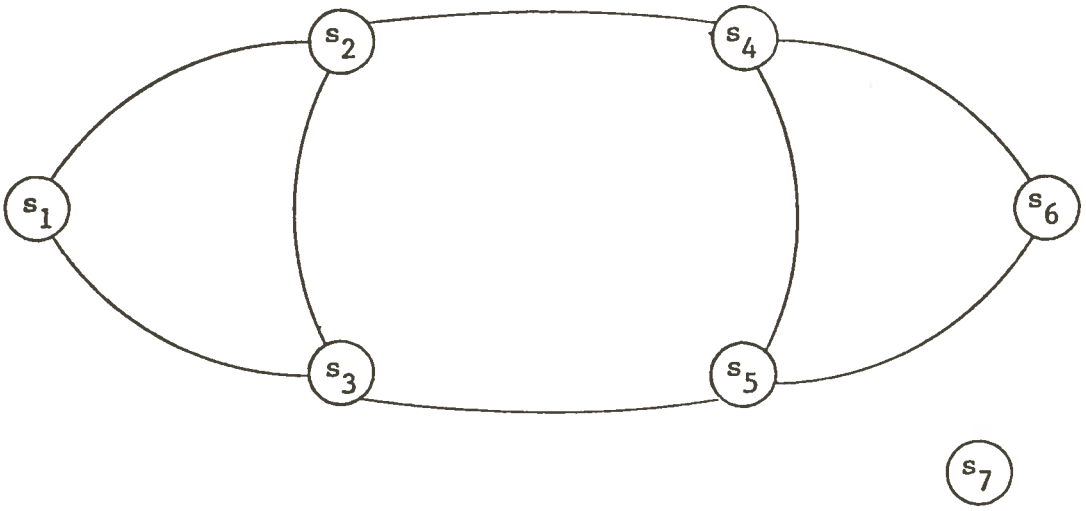


FIGURE 4.b

resulting in a larger number of nodes in the graph of FIGURE 3. An extension of this research effort, as will be discussed in Chapter VII, should involve devising an efficient means of determining the distinct subsets of categories induced by the words of the Thesaurus. The machine accessible version of the Thesaurus used in this study and the means by which a graph representation of it is constructed will be discussed in greater detail in Chapter VI.

Aside from allowing the initial graph representation of the Thesaurus to be easily constructed, the use of intersection graphs greatly facilitates subsequent analysis of the graph structure proposed to represent the Thesaurus. Specifically, as connected nodes of the initial intersection graph are determined a new intersection graph can be constructed, whereby the union of each of the subsets of categories corresponding to nodes of totally connected subgraphs generates a new set, F, of subsets of categories. These new subsets of categories can then be interpreted as the nodes of a new intersection graph. Each time an intersection graph is constructed in this manner the connectedness properties of the graph are preserved, while the size of the graph is optimized. When attempting to analyze the structure of a large graph, such as that proposed to represent the Thesaurus, the use of intersection graphs can greatly facilitate the analysis of structural characteristics. The specific details of the use of intersection graphs in this research are given in Chapter VI.

Thus far, the graph structures which have been proposed as representations of the relations among categories and groups of categories in the Thesaurus have been represented pictorially. However, for the purposes of automatically analyzing properties of a graph, such graph structures are usually represented by an adjacency matrix, sometimes called a connectivity matrix. An adjacency matrix M can be defined for a graph representation of the Thesaurus as follows: the labels for the rows and columns of the matrix correspond to the nodes of the graph structure, with n rows and n columns for an n-node graph. In the case where each node of the graph corresponds to a single category (as in FIGURE 2 - page), for every pair of categories c_i and c_j in the Thesaurus, or some subset of the Thesaurus, whose intersection is non-null (i.e., at least one identical entry occurs in both categories) the corresponding element, M_{ij}, of the matrix (row-column intersection) is set equal to one (1), thus signifying the existence of an arc connecting the two nodes in the graph. Otherwise, the element M_{ij} of the matrix is set equal to zero (0). Referring to the definition of R, which establishes the existence of a connection between two categories, each M_{ij} of the matrix M is defined as follows:

1) $M_{ij} = 1$, if $c_i \cap c_j \neq \emptyset$ is true for the categories c_i and c_j in the set C.
2) Otherwise, $M_{ij} = 0$.

The adjacency matrix for the graph depicted in FIGURE 2 is shown in FIGURE 5. The adjacency matrix for the intersection graph which is con-

The Adjacency Matrix Corresponding to
the Graph of FIGURE 2

	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	0	0	0	0	0	0
2	1	1	1	0	0	0	0	0	0	0
3	1	1	1	1	0	1	0	0	0	0
4	0	0	1	1	1	0	1	0	0	0
5	0	0	0	1	1	0	1	0	0	0
6	0	0	1	0	0	1	1	0	0	0
7	0	0	0	1	1	1	1	1	0	0
8	0	0	0	0	0	0	1	1	0	0
9	0	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	0	1	1

FIGURE 5

structed for subsets of categories, as described in preceding paragraphs, is similarly constructed, except that the nodes of the graph represent subsets of connected categories. Formally, each M_{ij} of the matrix M is defined as follows:

1) $M_{ij} = 1$, if $s_i \cap s_j \neq \emptyset$ is true for the subsets of connected categories s_i and s_j in the set F, $F = \{s_1, s_2, \ldots, s_n\}$.
2) Otherwise, $M_{ij} = 0$.

The adjacency matrix representation of a graph provides a compact and straightforward means of representing a graph structure, and greatly facilitates analysis of the graph using the computer. An important characteristic of the adjacency matrix, which can be seen in FIGURE 5, is that the matrix is symmetrical. That is, for each M_{ij} set equal to one, likewise so is M_{ji}. This characteristic of the matrix is the result of the property of symmetry existing for the relation R used in defining the existence of connections between categories in the Thesaurus. The symmetry of the matrix is an important comsideration when the attempt is made to manipulate the matrix algorithmically. The specific use made of the matrix form of graph representation, as well as the implications of the symmetry property of the matrix, will be discussed in detail in Chapter VI.

Attempts to make similar applications of graph theory to the representation and analysis of thesauri have been undertaken. Abraham (19,20) represents a thesaurus as a directed graph, where nodes of the graph correspond to terms in the thesaurus and arcs represent the existence of a "semantic association" between terms. For the purposes of his research "semantic association" is defined in terms of both near-synonymy (see K. Sparck Jones (2)) and hierarchical relationships between terms.[4] In this graph representation, near-synonymy between terms results in a bi-directional arc between the nodes corresponding to these terms, which essentially is equivalent to an undirected arc between the nodes. Hierarchical relationships are represented by uni-directional arcs, where each arc is directed from the node corresponding to the less conceptually inclusive term to the node corresponding to its more conceptually inclusive counterpart. Amoug the characteristics considered by Abraham is that of connectedness. He analyzes the resultant graph structure in terms of subgraphs which he calls "leaf graphs," within which any two nodes are "cycle-connected." The cycle-connectivity of the nodes in a graph defines

---

[4]These hierarchical relationships are similar to those discussed in Chapter II, in which terms are grouped under successively broader classifications. Although Abraham does not explain his usage of the phrase "hierarchically related" in detail, it can be ascertained that two terms would be hierarchically related when one term is conceptually more "inclusive" than the other.

an equivalence relation which results in a graph being partitioned into mutually exclusive subsets of nodes, or distinct equivalence classes. My approach is similar to that taken by Abraham, and the intent of my research and its implications are basically the same.

It is now possible to make a more precise and meaningful statement of the major thrust of this research effort. The structural analysis described here of Roget's Thesaurus is directed toward determining the set of distinct equivalence classes existing for the set of categories (semicolon groups) comprising the Thesaurus. Each distinct equivalence class, as discussed earlier, represents a maximal totally connected subset of the categories of the Thesaurus, with no connections existing between any of the distinct equivalence classes constructed. The existence of only one distinct equivalence class containing all categories, or semicolon groups, in the Thesaurus, or some subset of it, would allow it to be characterized as totally connected. The existence of several (two or more) distinct equivalence classes precludes such a characterization being made.

CHAPTER IV

ALTERNATIVE APPROACHES TO THESAURAL CONNECTIVITY

As mentioned briefly in Chapter I, thesaural connectivity can be approached in a number of different ways. With regard to Roget's Thesaurus, each approach is determined primarily by altering the way in which the definition of an abstract thesaurus given at the beginning of Chapter III is interpreted as it relates to the Thesaurus. Specifically, altering the interpretation of the sets E, W, and C will necessarily alter the way in which thesaural connectivity can be defined. Likewise, changing the definition of relatedness (the relation R) will also alter the interpretation of thesaural connectivity. Altering the interpretation of the definition of thesaurus and changes in the definition of the relation R allow connectivity to be studied at different levels of the hierarchy of the Thesaurus, as well as to allow the examination of different aspects of connectivity at each particular level of the hierarchy.

The primary means of changing the interpretation of the definition of thesaurus is to alter the interpretation of the set of categories, **C**. For the purposes of this research the Thesaurus can be represented as a set of categories C, of which each c_i belonging to this set is a single semicolon group occurring in the Thesaurus. It is possible, utilizing the hierarchical structure of the Thesaurus, to define each category $c_i \in C$ in "broader" terms. For example, each category could be interpreted to be a *sub-category* (as defined in Chapter II), which in turn is comprised of some number of semicolon groups. Thesaural connectivity will then be defined for the next higher level of the hierarchy. Likewise, it would be possible to define C as the set of all *categories*, and so on, with each reinterpretation resulting in the connectivity of the

Thesaurus being defined for a different level of the hierarchy of the Thesaurus. As a result of varying the interpretation of the set C, it should be possible to determine any differences in connectedness existing between levels of the Thesaurus. It can be shown quite easily that the Thesaurus is totally connected at level one (1) of the hierarchy (when the set of categories is interpreted as the set of *classes*). Likewise, it is probable that the Thesaurus is totally connected at level two (2) of the hierarchy (when the set of categories is interpreted as the set of *subclasses*). In all of the cases outlined above, the interpretation of the categories comprising the set C is more inclusive than that which is made for this research, in that the number of entries comprising each category increases as the interpretation of each is made for successively higher levels of the hierarchy. Likewise, it is possible to investigate connectivity as existing for the set W of words in the Thesaurus, whereby, each word would be represented by a single node in the graph representation.

In addition to changing the interpretation of an abstract thesaurus as relating to Roget's Thesaurus, it is also possible to alter the connectedness of the Thesaurus by changing the definition of the relation R. As the relation R is defined for this research, two categories are said to be related, or connected, if the categories have at least one identical entry in common (occurring in each). The relation R could be redefined so that two categories are related only when at least two identical entries occur in each, and so on. The manner in which categories may be related using such a redefinition of R can be seen in the categories of FIGURE 2 (page). For example, if the relation R is defined to exist only for categories with at least two entries in common, only categories c_3 and c_4, c_5 and c_7, c_9 and c_{10} would be related. Likewise, if there must be three entries in common before two categories are said to be related, only categories c_5 and c_7, and c_9 and c_{10} of FIGURE 2 would be related. Further, the definition of the relation R could be altered to allow entries with the same stem to relate categories. For example, using this definition of R the entries "home" and "homely" in the categories corresponding to nodes 6 and 9, respectively, of the graph in FIGURE 2 would serve to connect the two categories. And thus, an arc would exist between nodes 6 and 9. Whereas the changes in R outlined above serve to "restrict" the definition of connectivity, in that fewer connections would exist between thesaural categories, allowing entries with identical stems to establish a connection serves to "expand" the definition, or make it more inclusive, with more connections being evidenced among categories. An additional change in the definition of R which could be made would be to allow for cross-references to establish the relation R between two categories, whereby, the existence of a cross-reference between two categories, or between a category and a set of categories, would serve to relate, or connect, them, and thus, result in the addition of an arc between the corresponding nodes of the graph representation of the Thesaurus, Oftentimes it is the case that a cross-reference merely refers to another semicolon group in the Thesaurus where an identical entry can be found. If this is the case, then the cross-reference will not affect the connectedness of two semicolon groups, as they are already connected by virtue of

containing an identical entry. However, a cross-reference in the Thesaurus normally references a *category* or *sub-category* comprised of several semicolon groups, of which one contains an entry identical to the entry found in the semicolon group possessing the cross-reference. If the existence of the cross-reference is taken to relate all of the semicolon groups in the referenced *category* or *sub-category* to the semicolon group in which the reference appears, then significant changes in the connectedness of the Thesaurus may result. Likewise, if the entry possessing the cross-reference serves to label a referenced *category* (as with the example of cross-referencing given in Chapter II), but does not appear within the *category*, then again, significant changes in the connectedness of the Thesaurus may result.

The changes in the definitions and their interpretations outlined in this chapter are in no way intended to be an exhaustive list of possible ways to alter the way in which thesaural connectivity may be studied. In fact, the changes discussed above may be combined in various ways to suggest additional approaches, and, as more research is done in the area of thesaural connectivity, possibly many more changes in these basic definitions and their interpretations will be suggested. This brief discussion has been intended as a demonstration of some of the many structural approaches to thesaural connectivity which may be considered.

CHAPTER V

SOME SEMANTIC IMPLICATIONS OF THE APPROACH TO THESAURAL CONNECTIVITY IN THIS STUDY

The changes in the definition of relatedness between categories in the Thesaurus and alterations in the interpretation of an abstract thesaurus as it relates to Roget's Thesaurus can be viewed in both semantic and syntactic terms. Specifically, the connectedness of a thesaurus, and Roget's Thesaurus in particular, can be interpreted as one characteristic of syntactic structure, which I use broadly to refer to those characteristics of a thesaurus which affect the way in which information in the theaaurus may be accessed. Likewise, structural characteristics which determine syntactic structure are responsible for defining certain aspects of the semantic structure of a thesaurus, as reflected in the organization of concepts and words in the thesaurus. As a result, connectedness of a thesaurus reflects certain aspects of both the syntactic and semantic organization of the thesaurus. And, since the relationship between the syntactic and semantic structure of a thesaurus must necessarily be a mutually dependent relationship, studying thesaural connectivity is one means of clarifying properties of both the syntactic and semantic structure of a thesaurus. However, the approach to thesaural connectivity taken in this research effort has its greatest implications in regard to the syntactic structure of the Thesaurus, primarily as a result of treat-

ing words in terms of character strings, whereby much of the semantic information directly associated with a word in the Thesaurus is largely ignored. Likewise, much of what can be called the context of a word occurrence in the Thesaurus is ignored, thus, further restricting the amount of semantic information associated with word occurrences. As a result, the kinds of inferences which can be made concerning semantic properties of the Thesaurus must necessarily be somewhat restricted, although, as mentioned earlier, some semantic properties of the Thesaurus may be clarified, as a result of the interdependence between its syntactic and semantic organization.

There are a number of semantic implications arising from the approach to thesaural connectivity taken in this research. As the set of words, W, is defined, a word is taken to be the set of all occurrences of some comma delimited character string. This interpretation restricts the amount of semantic information which can be attached to each word. That is, treating words as character strings ignores characteristics of words such as us ge, spelling, and affixing. As a result, many entries in the Thesaurus which are in fact semantically related will not be treated as such. Likewise, many entries which are not semantically related will be treated as if they were. Some examples of the way in which thesaural connectivity, as studied in this research, is affected by this interpretation of words follow.

First of all, this interpretation of word ignores differences in meaning resulting from different forms of usage of the word. For example, no distinction is made between the occurrence of the word "jump" as a noun and the occurrence of the word "jump" as a verb, although semantic differences arise as a result of each particular usage.[1] Further, associated with many word occurrences, or entries, in the Thesaurus is some form of qualifying information contained in brackets or parentheses, which often serves to "expand" or "restrict" the maaning of a particular entry. For a discussion of the kinds of information contained in brackets refer to Taylor (21). However, as an example of how such qualifying information may affect the meaning of particular word occurrences, in one case an entry may be designated as being a dialectal usage of some word and grouped with entries having similar meanings, but in another, nondialectal case,

---

[1] As briefly mentioned in Chapter II, *sub-categories* in the Thesaurus are grouped in terms of the part of speech which words in the *sub-category* may be used. Thus, this information concerning word usage can be extracted from the Thesaurus, and it is possible, then, to use this information when determining the semantic relatedness between entries and groups of entries in the Thesaurus. As a result, changes in the approach to thesaural connectivity could be made fairly easily to reflect the distinctions between this aspect of word usage.

the entry may be grouped with other entries having an entirely different meaning from the former group. Thus, some degree of semantic ambiguity results. In both of the above examples it can be argued that the word occurrences retain some degree of semantic relatedness, and thus, linking them together can serve to provide some kind of semantic information. For the purposes of this research this is assumed to be the case.

Secondly, defining semantic relatedness in terms of character strings ignores the relation existing between a particular word and the spelling variants of that word, except when they occur in the same or related categories. Whereas these words should be treated as having the same meaning, for the purposes of this research they are not. The entries "rhyme" and "rime" are examples of spelling variations of a word. There are a number of difficulties attached to the problem of determining spelling variations of a word, except where marked in the Thesaurus, and the handling of this problem is beyond the scope of this research effort.

Thirdly, word occurrences having identical stems, but different affixes, will not be considered as related, again, except as they occur in the same or related categories. For example, singular and plural forms of a word will not serve to relate two categories, and are not considered to be related, using the definition of R given in Chapter III. There are many more cases of words with identical stems that are, in fact, tenuously related and result in a distant or weak semantic connection, and thus should not be considered related. However, as in cases of singular and plural forms of a word, the relation may be very strong. As a result, distinguishing between these cases by automatic means presents many difficulties. And, as with spelling variations, the handling of this problem is beyond the scope of this research. In any case, even with singular and plural forms of words, it can be argued that there exists a semantic basis for distinguishing between words possessing the same stem but different affixes.

In addition to the problems associated with treating words simply as character strings, as discussed in the preceding paragraphs, there exist additional problems which are, perhaps, more serious, and certainly more difficult to solve. Just as contextual information is important in determining the meaning of word occurrences in documents and texts, likewise, it is important when determining the relationships existing among word occurrences in a thesaurus. That is, the meaning of an entry in a thesaurus is largely dependent upon its placement within the conceptual organization of the thesaurus, or in other words, its context. The context of an entry in a thesaurus is determined by three major considerations:

1) the entries associated with it within a semicolon group,
2) the division within the formal hierarchy of the thesaurus in which it occurs, and
3) any qualifying information, such as that information appearing in brackets or parentheses, directly associated with the entry.

Ignoring these considerations when attempting to determine the existence of semantic relations among entries in the Thesaurus is responsible for introducing some degree of ambiguity, not only between entries, but also between categories in which there are occurrences of the entries. Examples of the kinds of information associated with entries, as specified in the third consideration above, have been mentioned in preceding paragraphs and will not be discussed further here. However, as an example of the contextual information conveyed from the word associations established within semicolon groups, and as a result of the classifications represented in the formal hierarchy, consider the following two semicolon groups found in Roget's Thesaurus:

; home, house, cabin, domicile, domiciliate;
; cabin, stateroom;

The occurrence of the entry "cabin" in each semicolon group serves to relate, or connect, these two groups, although the entries occurring with "cabin" in each group cause two different meanings of the word, "cabin," to be implicitly defined. In the former semicolon group, the meaning of "cabin," as created by the entries "house" and "home," appears to be that of some form of dwelling (perhaps a log cabin or vacation cabin), whereas the meaning of "cabin" in the second semicolon group, when its associated entry, "stateroom," is considered, seems to be that of a room or compartment on a ship. As a result, some degree of ambiguity arises. However, taking into consideration the *categories* of the Thesaurus in which these semicolon groups occur, this semantic ambiguity is somewhat resolved, and the nature of the connection is made clearer. The first group occurs in *category* 187, which is labeled "Habitation," and the second group occurs in *category* 191, which is labeled "Room." In any case, the nature of the semantic relationship, or connection, between these two semicolon groups is not considered when basing the relatedness of these groups solely on the occurrence of the character string "cabin" in each.

The discussions of the preceding paragraphs were intended to provide some understanding of the many, and oftentimes complex, semantic relationships existing among entries and groups of entries in the Thesaurus. Also, it is hoped that these discussions have served to illustrate some of the problems associated with studying characteristics of thesaural structure. A study of the structure of the Thesaurus, as undertaken in this research effort, can be very useful in uncovering and clarifying certain semantic properties of the Thesaurus. However, any conclusions or inferences concerning the semantic characteristics of the Thesaurus resulting from such a study must take into consideration, and be interpreted in light of, the kinds of semantic relations existing among entries, as described above, and the existence of ambiguities which are not easily recognizable or resolved when attempting to analyze these relationships automatically.

Problems of word ambiguity are well-recognized among those involved

with computerized language analysis, and considerable study has been devoted to means of minimizing or eliminating it. Such ambiguities in word meanings necessarily affect the way categories, or semicolon groups, may be connected. One study dealing with the semantic "disambiguation" of the categories comprising Roget's Thesaurus has been undertaken by Dillon (22).

The usual approach -- and it is followed by Dillon -- to devising means of disambiguating word meanings by automatic methods is based on developing "measures of semantic relatedness" or "measures of association" between words and groups of words. These measures are usually concerned with frequency of word occurrences and co-occurrences, whether it be within and between thesaurus categories or within and between documents being analyzed. Also, measures of semantic relatedness may take into consideration the "distance" between words, where "distance" may refer to the spatial distance between words in a text or, in the case of categories in a thesaurus, the number of intervening categories in the chain of categories which serves to relate two categories. Gotlieb and Kumar (23) provide a detailed discussion of an experiment which they have undertaken to develop an effective measure of association to be used in reorganizing "indexing vocabularies." Their use of semantic measures of association, although not dependent upon frequencies of term occurrences, involves consideration of the distance between terms. The approach taken by Gotlieb and Kumar was to treat an "index system" as a graph structure, and represent this graph structure by an adjacency matrix, using a measure of association based on the distance between terms to determine clusters of closely semantically related terms within the index. There are obvious applications of these techniques to the study of thesaural connectivity, since connectedness, as explained earlier, is determined by the existence of category chains serving to connect categories in a thesaurus. Incorporating these considerations into the definition of relatedness between categories would provide still another approach to analyzing thesaural connectivity.

The majority of attempts to minimize the effect of word ambiguities in thesauri are incorporated either during the construction of a thesaurus for use with a particular language analysis system (e.g., when using cluster analysis) or when making use of an existing thesaurus. Reisner, however, has studied an interesting alternative, whereby the particular user of a thesaurus constructs his or her own semantic relations between terms in the thesaurus (24). The construction of associative thesauri can also be interpreted as an attempt to minimize the problem of semantic ambiguities between thesaurus entries for some group of users.

In conclusion, when considering the definition of an abstract thesaurus and the interpretations of this definition, it is apparent that a large number of approaches to studying thesaural connectivity can be taken. In addition, there are, as discussed in this chapter, many semantic and syntactic characteristics of thesauri which can or should

be taken into account when analyzing connectedness defined in terms of the relations among entries and groups of entries. Likewise, in attempting to deal with connectivity by automatic means there are a similar number of algorithmic approaches to be considered. For the approach taken in this research effort there are a number of other considerations related to the use of the computer which must be dealt with also. These considerations, and the methodology chosen for implementation of the algorithm selected, will be discussed in Chapter VI.

CHAPTER VI

THE METHOD

As previously discussed, the approach to the study of thesaural connectivity taken in this research effort is based on the representation of the Thesaurus as a graph structure. An important part of this research effort has been the development of a method for explicating certain of the connectedness characteristics of the graph structure proposed to represent the Thesaurus. Treating the Thesaurus as a graph structure poses several problems related to use of the computer in analyzing this structure. That is, there exist some constraints on how this graph structure can be represented and manipulated efficiently by computer. Specifically, due to the large size of the Thesaurus and its graph representation, the search and storage of this graph representation are important factors to be considered when developing and implementing an algorithm for manipulating this graph representation.[1] In addition, time constraints on the use of the computer for this task become an important consideration. This chapter discusses the approaches used for dealing with these factors. The algorithm used, and its implementation as a computer program, are also presented and discussed. But first, some discussion will be devoted to the machine accessible version of the Thesaurus which has been used, and thus, some understanding of its use should be gained.

The machine accessible version of Roget's Thesaurus used for this study is the result of essentially three phases of editing and reformating of the printed version of the Thesaurus. Briefly, the purpose of this editing has been to develop a version of the Thesaurus which maintains the semantic structure and information content of the printed version, yet

---

[1] The Thesaurus is comprised of approximately 200,000 entries, which may be organized into as many as 60,000 categories, or semicolon groups. As a result, a graph representation corresponding to the set of all categories in the Thesaurus would have 60,000 nodes, if the estimate of the number of categories is correct.

will allow an automated language analysis system to access and use this information to the fullest extent possible. Much of the information content of the printed version of the Thesaurus can be utilized only with a considerable amount of prior knowledge about many of the semantic relationships implicitly existing among entries and groups of entries in the Thesaurus. In addition, there exist in the printed version a number of syntactic, or structural, conventions regulating use of the Thesaurus which had to be reformatted in order to make it usable by an automated system. For a detailed discussion of the editing and reformating refer to Harris (25,26).

The machine accessible version of the Thesaurus has evolved from an exact copy of the printed version through the various stages of editing and reformating to the "parsed" version currently existing. This parsed version consists of the entries of the Thesaurus separated out along with all of the semantic and syntactic information directly associated with each entry in the printed text of the Thesaurus. As mentioned in Chapter III, categories in the parsed version of the Thesaurus are identified by unique numerical codes which represent the location of the categories (in terms of the ***category***, *sub-category* and the sequence of the semicolon group within the *sub-category*) in the formal hierarchy of the printed version of the Thesaurus. During the parsing of the Thesaurus each entry is assigned the identifying code for the category in which it occurs. Further, these codes are used in this research to represent the categories of the Thesaurus and, in this chapter, when I refer to a set of connected categories I am referring, in actuality, to a set of identifying codes representing categories that are connected. These identifying codes serve as the data to the process for constructing a graph representation of the Thesaurus and the process for analyzing the structure of this graph representation.

For the purposes of this study, the entries (with the numerical codes identifying the categories in which they occur) comprising that portion of the Thesaurus used in testing and demonstrating the computer program which has been written were sorted alphabetically, and all identical entries appear "adjacent" to one another in the file which was created. As a result, the occurrences of identical entries and the information concerning their location in the formal hierarchy of the Thesaurus are easily accessed. This format of the machine-accessible version of the Thesaurus greatly facilitates the process of extracting information to be used in determining the sets of connected categories induced by the words of the Thesaurus, as discussed in Chapter III. The next phase of the editing and reformating of the Thesaurus will involve the sorting of the entries for the entire thesaurus, and the method for determining the connectedness of the Thesaurus was developed in anticipation of using this sorted version.

The selection of the method to be used in determining the connectedness of the Thesaurus was based on several primary considerations: 1) the size of the graph structure representing the Thesaurus, 2) the

organization of the data, and 3) the capability for reorganization as "new" data is generated. As briefly mentioned earlier in this chapter, the size of the Thesaurus and its graph representation is an important consideration when attempting to develop a method for manipulating this structure using a computer. The amount of storage and the time necessary for searching a graph representation of the Thesaurus are prohibitive when working with a large graph. As a result, some means of representing this graph structure in the most efficient manner possible had to be found. One means of increasing storage efficiency is to represent the graph as an adjacency matrix, as described in Chapter III. In addition, to further increase storage efficiency the particular form of adjacency matrix chosen was that of a "packed" matrix, whereby each element, or row-column intersection, of the matrix is represented in one bit of a computer word, as opposed to an entire computer word for an "unpacked" matrix. Therefore, for a graph with N nodes, the N x N matrix corresponding to the graph can be stored in a packed form, with each row of the matrix requiring [N/W] + 1 computer words of storage, where W is the size, in bits, of each computer word.[2] Hence, the storage required for the entire matrix would be N x ([N/W] + 1) computer words. However, since the matrix representing the graph structure is symmetrical in nature (refer to Chapter III), only that portion of the matrix above or below the main diagonal is necessary to represent the graph. As a result, a savings in storage of nearly 50 percent can be realized if this form of the matrix is used. Specifically, the number of computer words necessary for the storage of an N x N matrix is approximately (N/2) x ([N/W] + 1). The exact number of computer words can be computed using the following expression:

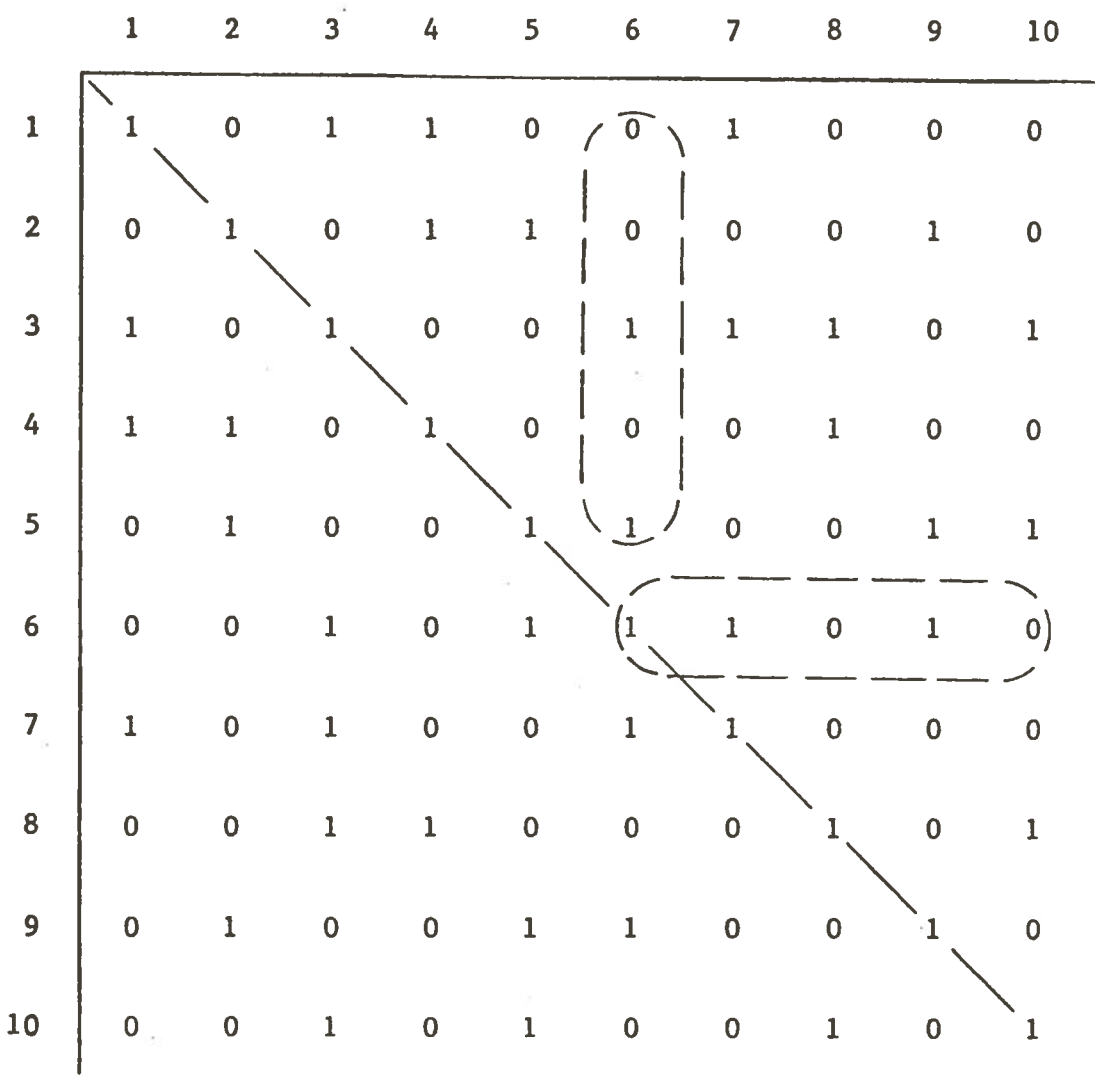
$$\sum_{i=1}^{N} (([N/W] + 1) - [(i - 1)/36])$$

In an effort to capitalize on the reduced amount of storage necessary to store only "one half" of a symmetric matrix, such a matrix representation has been used in this research. The method of using an adjacency matrix consisting of only that portion above the main diagonal is illustrated in FIGURE 6. It can be seen that the elements of row six (6) of the matrix given may be determined by examining those portions of column six and row six above the main diagonal (those portions circled in the matri of FIGURE 6). Although the determination of some row of the matrix in this

---

[2]a)For some value M, [M] is evaluated as the largest integer less than or equal to M. Thus, any fractional part is ignored. Since the use of some fractional part of a computer word requires that the entire word be used, one (1) is added to [N/W] for the purposes of allocating storage in order to account for the fractional part of the value of N/W.
b)The word size of the Honeywell 635, on which this research has been conducted, is 36 bits. Therefore, the value of W would be 36.

An Example of the Use of a Symmetric
Adjacency Matrix for Some Undirected Graph



	1	2	3	4	5	6	7	8	9	10
1	1	0	1	1	0	0	1	0	0	0
2	0	1	0	1	1	0	0	0	1	0
3	1	0	1	0	0	1	1	1	0	1
4	1	1	0	1	0	0	0	1	0	0
5	0	1	0	0	1	1	0	0	1	1
6	0	0	1	0	1	1	1	0	1	0
7	1	0	1	0	0	1	1	0	0	0
8	0	0	1	1	0	0	0	1	0	1
9	0	1	0	0	1	1	0	0	1	0
10	0	0	1	0	1	0	0	1	0	1

FIGURE 6

manner is a straightforward process, its implementation, when using a packed adjacency matrix, is fairly complicated. The means of extracting that information contained in one bit of a computer word requires a considerable amount of computation and manipulation of the computer words comprising the matrix. Specifically, when working with only "one half" of the matrix it is necessary to store the matrix as a one-dimensional array or vector in order to efficiently work with the matrix. As a result, different elements of a particular column of the matrix may be stored in different computer word and bit configurations for each row of the matrix. For example, if it is assumed that the matrix in FIGURE 6 is stored in three bit computer words, then the element of column six in row one would be stored in the third bit of the second word of the row. However, the element of column six in row five would be stored in the second bit of the first word of the row.

Some examples of the amounts of storage necessary for the matrices corresponding to some selected N-node graph sizes may be found in Appendix A. As can be seen from the chart in Appendix A, even when using "one half" of a packed adjacency matrix, the amount of storage required for a relatively small matrix can be quite large. As a result, some way of organizing the data (thesaural categories) had to be formulated which would allow for a practical means of handling the graph structure as represented in matrix form.

The most direct, and most easily carried out, means of making the task of working with the Thesaurus a manageable undertaking is to work with the Thesaurus in "parts." That is, the approach taken was to consider initially only a portion of the set of thesaural categories at any one time, and analyze the graph corresponding to the subset of categories considered. Thus, initially the graph representation of the Thesaurus is not considered in total, but rather, sets of connected categories are generated in a series of steps. In addition, as these sets of connected categories are generated they are subsequently treated as single nodes in the graph structure, as described in Chapter III. As a result, the graph representation of the Thesaurus is a dynamically evolving structure. Some nodes of the graph structure which evolves will represent sets of connected categories. The size of this graph, then, is optimized, while the information content of the graph is retained.

The process of working with the Thesaurus in parts, and the graph representation which evolves, as briefly described above, requires a considerable amount of reorganization of the data, or thesaural categories (represented by their identifying codes), during the process of generating the sets of connected categories. Essentially, as sets of connected categories are generated and new categories are added to them the storage of these sets must be reorganized. Likewise, when an arc is found to exist between two nodes of the graph representation the sets of categories

corresponding to these nodes must be merged to form a new set of connected categories. All of these changes must be reflected in the storage arrangemany of the categories serving as data.

Basically, the process for determining the distinct equivalence classes of categories involves generating an initial set of subsets of connected categories, where each subset of categories in this initial set is comprised of the categories in the Thesaurus which contain an occurrence of some specified word. That is, using the sorted-parsed version of the Thesaurus, wherein identical entries are grouped together, the set of those categories in which a word (set of identical entries) occurs is determined for all of the words comprising the Thesaurus, or that portion being studied. The categories comprising each set which is constructed in this manner are connected, and correspond to the nodes of a complete subgraph of the graph representation in which each category is represented by a single node. That is, each of the categories in each set are related, or connected, as a result of the occurrence of an identical entry in each category. These sets of categories are interpreted as the nodes of the initial intersection graph which is constructed, as discussed in Chapter III. Examples of the words, or sets of identical entries, comprising the sorted-parsed version of that portion of the Thesaurus for which the above procedure was performed (*sub-classes I and II* of *Class One*) are found in Appendix B. Likewise, the sets of categories (each category represented by its unique identifying code) constructed from these sets of identical entries are found in Appendix C. In addition, associated with each of the categories in these sets of categories is the number of connections between the category and the other categories in the set. This number corresponds to the number of categories in the set minus one (1) and, as sets of connected categories are merged, the number of connections for each category is used in determining what was referred to in Chapter III as the degree of a node. Specifically, upon completion of the process of determining the distinct equivalence classes of categories the number associated with each category will equal the degree of the corresponding node in the graph representation in which each category is a single node. Further, as can be seen in these examples, there are some entries for which there are no other identical entries occurring in that portion of the Thesaurus considered. As a result, each of these unique entries result in the construction of a set of categories containing only one category. These sets of categories correspond to either: 1) a category which is not connected to any other category (an isolated node), or 2) a category in which there is an entry for which no identical entries occur elsewhere, yet the category does not represent an isolated node. As briefly discussed in Chapter III, those sets of categories which correspond to the second case above do not affect the connectedness of the Thesaurus, since the words which induced the sets of categories result in no connections between categories, and they need not be considered when determining connectedness. In those cases in which a category corresponds to an isolated node the connectedness of the graph representation of the Thesaurus is affected, and these sets of categories must be accounted for

when determining connectedness. However, since it is necessarily true that sets of categories containing only one category will correspond to one or the other of the above cases, they need not be included in the data to the computer program which determines distinct equivalence classes of categories. Such sets of categories can be handled more efficiently by considering them separately. In the case of isolated nodes, they contribute to the sparseness of the matrix representation which is constructed and result in unnecessary search and storage costs. The sparseness of an adjacency matrix, as constructed in this study, refers to the relative number of zero elements, as compared to the number of nonzero elements (those set equal to one). The larger the number of zero elements of the matrix, the more sparse it is. The experimental results which have been obtained seem to indicate that the sparseness of the matrix which is constructed contributes significantly to a higher cost of the analysis of the matrix. This aspect of the results of this research will be discussed in greater detail in the next chapter.

Briefly, once the initial sets of connected categories have been generated, as described earlier, these sets of categories serve as the initial data to the computer program written to determine the distinct equivalence classes of categories. As connections between sets of categories are determined these sets are merged to form new sets of connected categories. The existence of a connection between sets of categories is established by the occurrence of at least one identical category (identifying code) in each set of categories. The occurrence of some category in two or more sets of categories signifies the existence of a chain or chains of categories connecting all categories comprising the intersecting sets. And thus, these sets of categories can be merged to form a new set of connected categories. The process of merging intersecting sets of categories continues until no new connections are found to exist, in which case, for the categories being considered, the set of distinct equivalence classes of categories has been determined. The process of determining distinct equivalence classes of categories is outlined below. A flowchart representation of this process is shown in FIGURE 7.

Step 1) Starting with the set of entries comprising the Thesaurus, for each word (set of identical entries) determine the set of categories, or semicolon groups, containing an occurrence of the word. These sets of categories correspond to the initial nodes of the graph representation.

Step 2) Select some number of the sets of categories (nodes) and, if at least one of the sets is connected to the first one selected, construct the adjacency matrix representation of the graph structure corresponding to these sets of categories. If no nodes are connected to the first one selected, then disregard it and select another node for consideration. If no connections can be found go to Step 5.

Step 3) Determine the maximal totally connected subgraphs (M.T.C.S's) of the graph represented by the matrix constructed in Step 2.

Flowchart of the Algorithn for Finding the Distinct Equivalence Classes of Categories

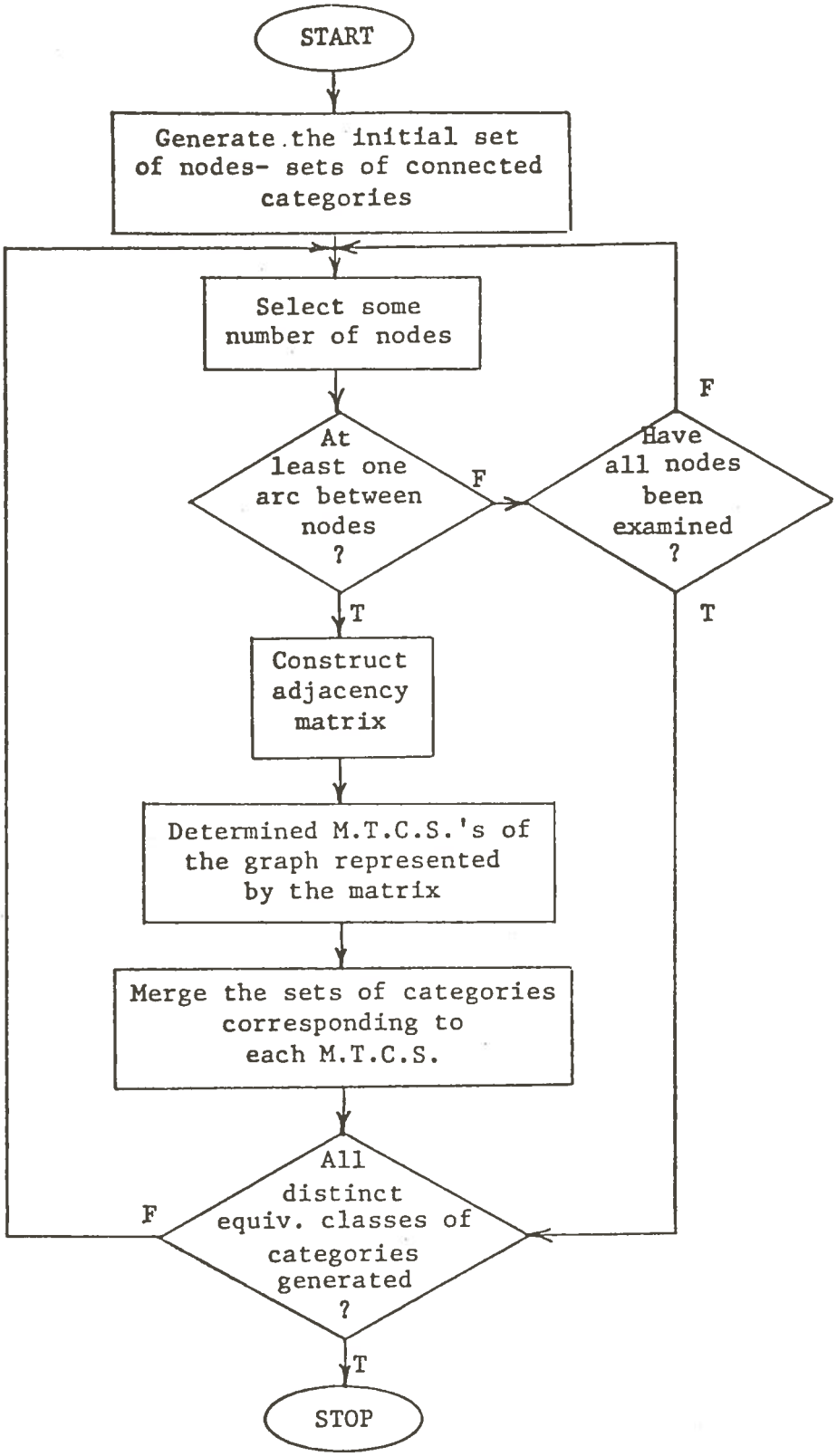


FIGURE 7

Step 4) Merge the sets of categories which were found to be connected in Step 3 (M.T.C.S's) into new sets of connected categories (new nodes of the graph). Reorganize the storage of the categories to reflect the existence of these new sets. Go to Step 2.

Step 5) If all distinct equivalence classes have been generated, then STOP. Otherwise, determine what connections still exist and return to Step 2.

Some additional explanation of the individual steps in this process, particularly of the step in which connections between sets of categories are determined (Step 3), is necessary in order to understand fully how this process has been implemented. Initially, the data to the computer program written to implement this process consists of the sets of connected categories generated in Step 1 above. Then, some number of these sets of categories, which correspond to the nodes of the graph, are input and the adjacency matrix representation of the graph structure corresponding to these sets of categories is constructed. The construction of the adjacency matrix proceeds as follows: the first of the sets of categories (a node of the graph) is selected for examination, and each succeeding set of categories input is searched for an occurrence of a category also contained in the initially selected set of categories. If one is found, then the corresponding row-column intersection of the matrix is set equal to one (1). Otherwise, the row-column intersection is set equal to zero (0). This process is continued until all of the sets of categories have been searched, after which the first row of the matrix will have been initialized. Then, the second set of categories is selected for examination and the process of searching the remaining sets of categories is repeated, after which, the second row of the matrix will have been initialized. The third set of categories is then used to begin the search in order to initialize the third row of the matrix, and so on. This means of searching the sets of categories corresponding to the nodes of the graph results in consideration of only the relations among the sets of categories which would be represented above the main diagonal of a symmetric adjacency matrix. At this point, it should be stated that the amount of searching required for setting up the matrix is considerably less than that which would be required for a procedure for determining all of the connections between nodes based entirely on searching the sets of categories. Such an approach would necessitate that all sets of categories be searched in order to find the connections with any one of the sets of categories, whereas, the amount of searching which is required for setting up "one half" of the adjacency matrix decreases as succeeding rows of the matrix are initialized.

After all of the rows of the matrix have been initialized the process of determining connections, and thus, totally connected subgraphs is begun. The algorithm used to determine the existence of arcs between nodes, and thus, determine the totally connected subgraphs, is an adaptation of an algorithm proposed by Ramamoorthy (27) for partitioning a given directed graph into its "unconnected" subgraphs. The so-called "unconnected" subgraphs generated by Ramamoorthy's algorithm correspond, in actuality, to

the set of maximal totally connected subgraphs of a graph, as defined in Chapter III. Although no significant changes were made to the basic algorithm, one important adaptation was necessary. The algorithm proposed by Ramamoorthy was designed for use with directed graphs, and thus, it was necessary to alter it somewhat in order to use it for an undirected graph. Specifically, this adaptation concerned the use of only that portion of the adjacency matrix above the main diagonal. Ramamoorthy's algorithm was intended to make use of the entire matrix.

In addition to the adjacency matrix of a graph, the algorithm makes use of what Ramamoorthy calls a "reachability vector," which is used to keep track of the nodes which are "reachable" from, or connected to, the node being considered. Also, a set of vectors, S, is defined, which is used to store the "nodes" comprising each totally connected subgraph that is found. Instead of a set of vectors, for which the storage allocation of each vector would be static, I have defined a linked list, EQCLA, to perform the function of storing connected nodes.[3] The basic algorithm for finding the totally connected subgraphs, or sets of categories, is given below with a flowchart representation of this algorithm given in FIGURE 8. Note, this algorithm represents the specific steps involved in Step 3 of the algorithm given on page 66, and the steps of the algorithm given below are labeled so as to indicate this.

Step 3a) Given the upper portion of an adjacency matrix M for an undirected graph, pick any node k.

Step 3b) Set the reachability vector R equal to those portions of the kth column and kth row above the main diagonal of M.

Step 3c) Examine the elements of R. Let R_i denote the ith element of R, which corresponds to the ith column or row of M. If all R_i, except that element lying on the diagonal, are zero, then store k in EQCLA and go to Step 3e. Otherwise, select some R_i which is nonzero and has not been examined previously, and update R by performing a logical addition of R with the ith "row" of M. Repeat this step for all nonzero elements of R until no changes occur in R.

Step 3d) Examine all nonzero elements of R. These elements correspond to the set of nodes connected to node k, and represent one totally connected subgraph. Store this set of nodes in EQCLA.

---

[3] Use of a linked list results in a considerable savings of storage, since only that amount of storage required for the number of nodes of the largest possible connected subgraph (which would be the graph itself) must be reserved. The partitioning of this storage is accomplished dynamically, as totally connected subgraphs are found. Hence, the amount of storage for EQCLA is that which would be required for one of the vectors in the set S.

Flowchart of the Algorithm to
Determine Maximal Totally Connected Subgraphs

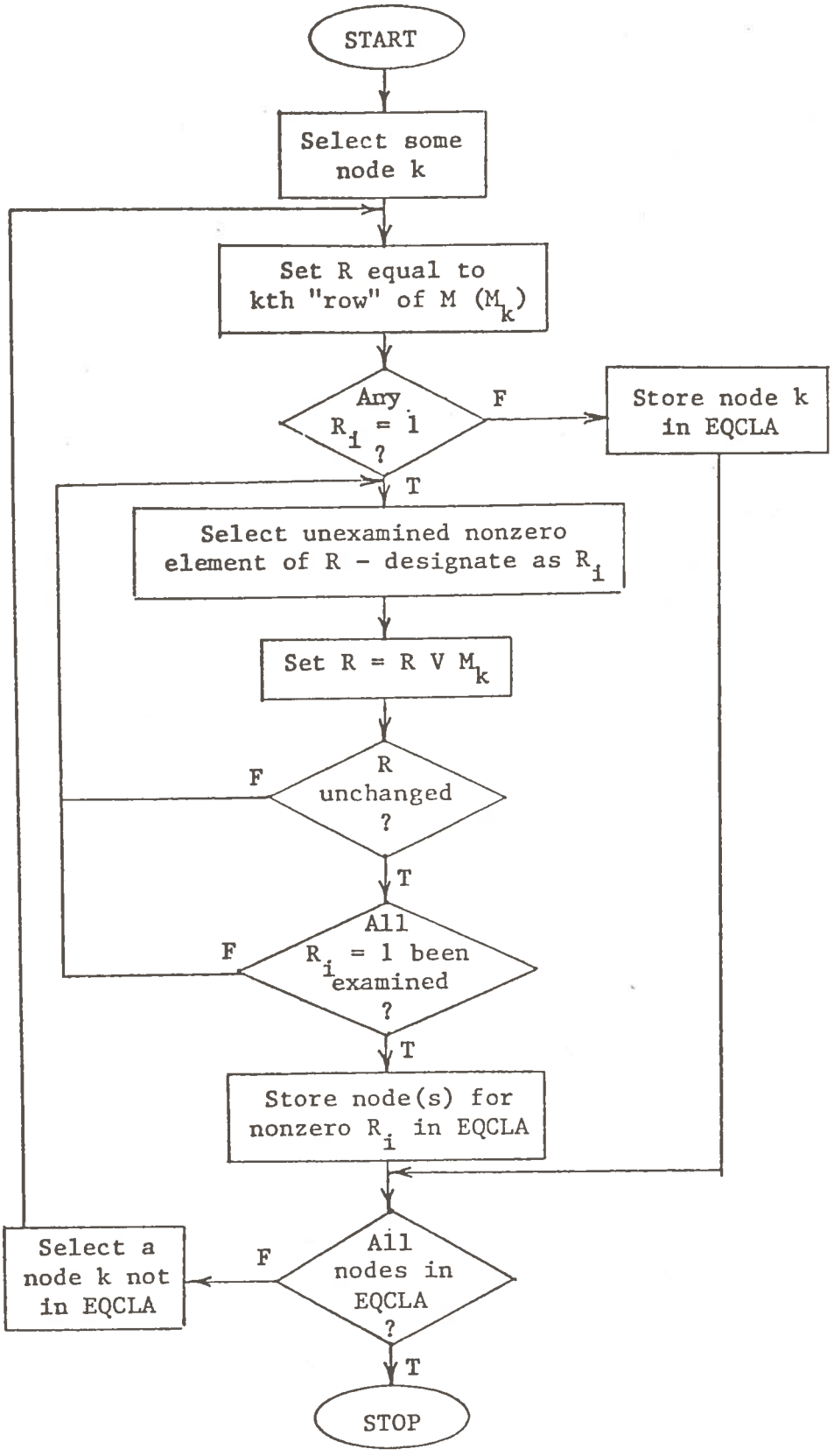


FIGURE 8

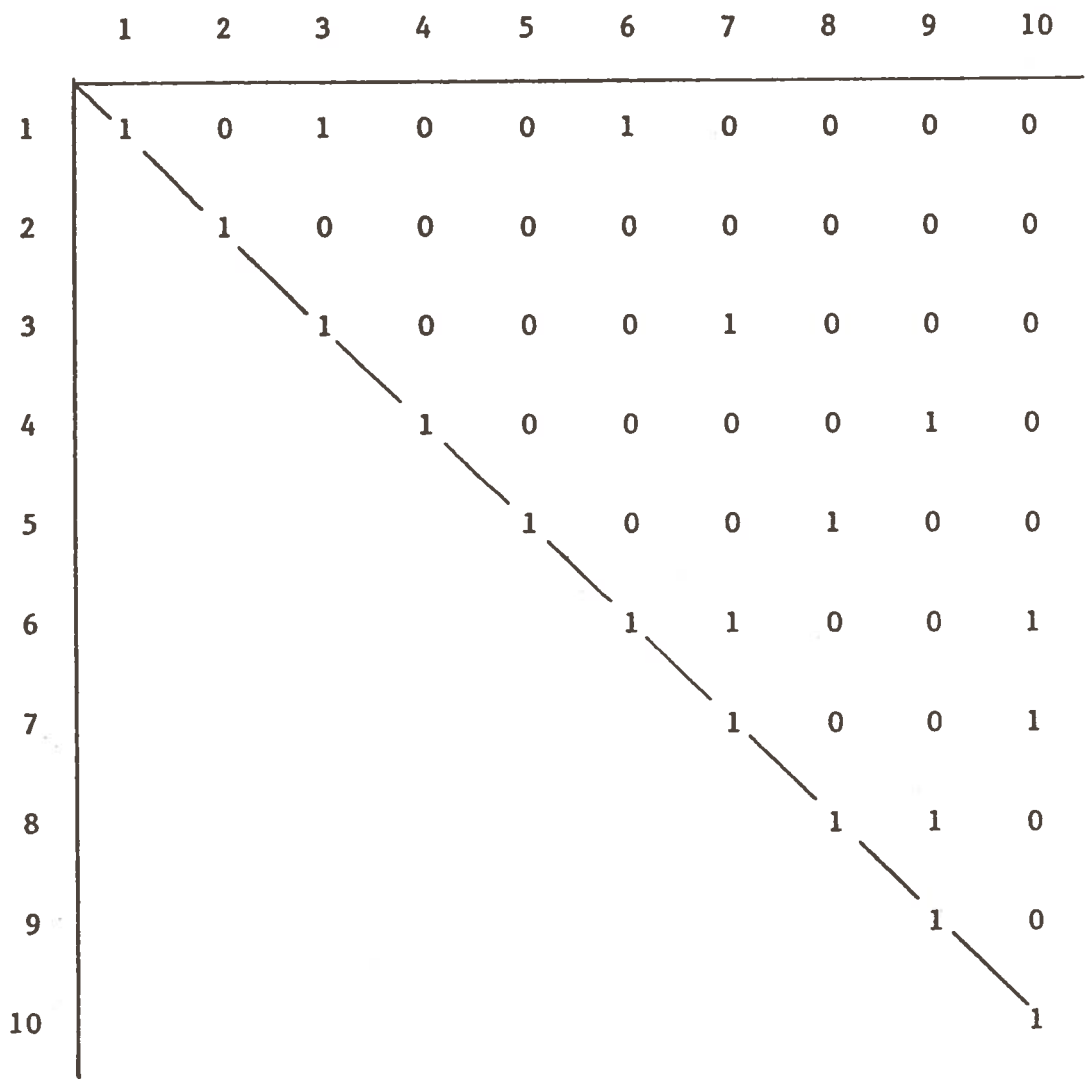
Step 3e) If all nodes have been assigned to some partition of EQCLA, then STOP. Otherwise, pick a node from the remaining unexamined nodes, and repeat steps 3b, 3c and 3d.

Upon completion of the above process each subset of nodes in EQCLA will correspond to a totally connected subgraph. Then, as was described earlier, the sets of categories corresponding to these connected nodes are merged together. The sets of categories formed in this manner then serve as "new" data to the process of generating the distinct equivalence classes of categories.

What follows is a brief example of the process outlined in the algorithm given above. Assuming that the matrix, given in FIGURE 9 and designated here as M, has been set up for ten sets of categories, the procedure for determining the maximal totally connected subgraphs of the graph represented by the matrix is as follows:

1) Set R equal to row one of the matrix (designated as M_1).
   R = 1 0 1 0 0 1 0 0 0 0
   The nonzero R_i (aside from that R_i corresponding to a diagonal element of M) are R_3 and R_6.
2) Logically add (V) R to the "rows" (M_i) of the matrix corresponding to these nonzero R_i.
   R = R V M_3 = 1 0 1 0 0 1 1 0 0 0
   R = R V M_6 = 1 0 1 0 0 1 1 0 0 1
   New nonzero R_i: R_7 and R_{10}. Hence, continue logically adding to R.
   R = R V M_7 = 1 0 1 0 0 1 1 0 0 1
   R = R V M_{10} = 1 0 1 0 0 1 1 0 0 1
   At this point, no additional changes in R have been generated, and all of the "rows" of the matrix corresponding to the nonzero R_i have been added (V) to R. As a result, the nonzero R_i correspond to the nodes which comprise a maximal totally connected subgraph of the graph represented by the matrix. This set of nodes is {1,3,6,7,10}.
3) Select a new, unexamined node; node 2. Set R = M_2.
   R = 0 1 0 0 0 0 0 0 0 0
   All elements of R (except that corresponding to a diagonal element of M) are zero. Therefore, this node is isolated, and is stored as the set {2}.
4) Select a new, unexamined node; node 4. Set R = M_4.
   R = 0 0 0 1 0 0 0 0 1 0
   Nonzero R_i: R_9.
5) Logically add R to the M_i corresponding to nonzero R_i.
   R = R V M_9 = 0 0 0 1 0 0 0 1 1 0
   New nonzero R_i: R_8.
   R = R V M_8 = 0 0 0 1 1 0 0 1 1 0
   New nonzero R_i: R_5.
   R = R V M_5 = 0 0 0 1 1 0 0 1 1 0
   At this point, there are no additional changes in R,

Sample Adjacency Matrix for Determining
Maximal Totally Connected Subgraphs



	1	2	3	4	5	6	7	8	9	10
1	1	0	1	0	0	1	0	0	0	0
2		1	0	0	0	0	0	0	0	0
3			1	0	0	0	1	0	0	0
4				1	0	0	0	0	1	0
5					1	0	0	1	0	0
6						1	1	0	0	1
7							1	0	0	1
8								1	1	0
9									1	0
10										1

FIGURE 9

and all nonzero elements have been examined. Therefore, the set of nodes corresponding to another maximal totally connected subgraph of the graph being considered is {4,5,8,9}.

6) All nodes have been examined, and are included in some set of nodes corresponding to a maximal totally connected subgraph. Therefore, the procedure has been completed and the maximal totally connected subgraphs of the graph represented by the matrix have been determined.

Each of the sets of nodes generated above are stored in EQCLA. Then, as mentioned earlier, the sets of categories corresponding to the connected nodes of the graph are merged to form a new set of connected categories, as described in Step 4 of the algorithm on page .

In conclusion, the discussions in this chapter were intended to provide some understanding of the method in which the graph theory techniques have been applied to this study of the structure of Roget's Thesaurus. Many of the specific details of the computer implementation of these techniques have been ignored, since it was felt that many of the details of this implementation do not contribute significantly to obtaining a general understanding of the methods used in working with the Thesaurus and its graph representation. In the case of the algorithm which was presented and discussed in this chapter the functional and logical correspondence between the algorithm and its implementation are, for the most part, preserved. Again, however, in order to promote greater understanding of the algorithm some aspects of its implementation are not completely represented.

CHAPTER VII

CONCLUSIONS

There are many approaches, as discussed in this thesis, to the development of a graph representation of the Thesaurus and to the study of the connectedness of the Thesaurus. Briefly, these approaches concern:

1) the connectedness exhibited at different levels of the hierarchy of the Thesaurus.
2) the effects of cross-referencing on the connectedness of the Thesaurus.
3) consideration of the semantic information directly associated with entries -- the part of speech of entries and the qualifying information appearing in brackets and parentheses with entries.

4) alterations in the definition of relatedness between categories.
5) the use of semantic measures, such as the "distance" between connected categories, for determining the strength of connections.

Each of these approaches determines to some extent the particular structure of the graph representation which can be developed for representing relations among words and groups of words in the Thesaurus. Likewise, each approach affects the amount and kinds of semantic information which may be associated with the connectedness of the Thesaurus. However, there are a number of characteristics of any graph structure proposed to represent the Thesaurus which may be considered when analyzing connectedness:

1) the maximal complete subgraphs of the graph.
2) the cut set of nodes for totally connected subgraphs.
3) the cycles (or circuits) of totally connected subgraphs.[1]

The maximal complete subgraphs of the graph representation of the Thesaurus in which each category corresponds to a single node represent "closely connected" groups of categories, or semicolon groups. And, as briefly discussed in Chapter III, the development of an efficient method for determining the maximal complete subgraphs should greatly facilitate working with the Thesaurus as a graph, since an intersection graph can be constructed in which each node represents a maximal complete subgraph of the original graph representation. Using the maximal complete subgraphs as the nodes of the initial graph representation constructed in this research would considerably reduce the size of the initial graph representation. In addition, knowledge of these maximal complete subgraphs may be quite useful in determining "central concepts," around which the sets of categories corresponding to totally connected subgraphs cluster.

The cut set of nodes and the cycles of a totally connected graph or subgraph should provide additional information concerning the accessibility of information in the Thesaurus, beyond that which is provided

---

[1] A cycle exists when there is a path (sequence of nodes connected by arcs) connecting nodes of a graph such that, for any node in the cycle, the path may be traversed so that it begins and ends with the node considered. A set of nodes comprising a cycle is referred to as a cycle set of nodes. In terms of a chain of categories a cycle exists when the initial category in the chain is also the terminal category in the chain. As an example, consider the chain $c_1, c_2, \ldots, c_k, c_1$.

by knowledge of connectedness alone. A number of methods for determining the cycles of undirected graphs have been proposed (28,29,30). Further, Paton (31) has proposed what appears to be a good algorithm for finding cut sets of nodes. Each of these methods have been, or can be, implemented using the matrix representation of a graph, and the logical extension of this research effort should be an investigation of these methods and the adaptations of them which would be required in order to utilize the techniques that I have implemented as a part of this research effort.

The method for examining the connectedness of the Thesaurus which has been developed during the course of this research seems to provide the most efficient means of working with a large graph, such as that representing the Thesaurus. In terms of the search and storage required for constructing and analyzing the matrix representation of the graph, the method which I have implemented for analysis of thesaural connectivity seems to provide the optimum approach. That is, in constructing and analyzing the adjacency matrix, of which only that portion above the diagonal is actually constructed, a considerable savings in search time over that which would otherwise be required is realized, as discussed in Chapter VI. In addition, the matrix form of representing a graph seems to be the most compact and most easily manipulated representation of a graph. Further, a packed adjacency matrix increases the storage efficiency of the matrix form of representation, although manipulation of the packed form of matrix is somewhat more complicated than that of an unpacked matrix. As briefly disucssed in Chapter VI, the sparseness of the adjacency matrix which is constructed seems to have considerable effect on the amount of search time, and thus, the cost of analyzing the matrix. During the process of testing the computer program which implements the process of setting up and analyzing the adjacency matrix there seemed to be a direct correspondence between higher costs of running the program and an increase in the sparseness of the matrix. As a result, it was decided that all "isolated" categories (categories not connected to other categories) should be removed prior to running the computer program against the Thesaurus. As discussed in Chapter VI, isolated categories may be easily removed from consideration. When this is done, the costs involved in executing the computer program are considerably reduced, and the reduction in costs far outweighs the costs of special handling of isolated categories.

An important characteristic of the method developed for determining the connectedness of the graph representation of the Thesaurus is that the method is not dependent upon the particular approach which is taken in defining the relatedness between categories, or semicolon groups. Changes in the definition of the relation R used in determining the existence of connections between categories affects the manner in which the initial sets of connected categories can be constructed. Once these initial sets of connected categories are constructed, however, the method developed for analyzing the resultant

graph structure is not dependent upon the particular relation R which was used. This was an important consideration in the development of the method and its computer implementation. As a result, and in anticipation of future research in thesaural connectivity, the computer program which has been written to implement this method can be used for studying connectedness defined in a number of different ways, some of which have been listed earlier in this chapter.

The approach to thesaural connectivity in this study has been concerned primarily with the connectedness of Roget's Thesaurus defined in terms of relations among the semicolon delimited groups of entries comprising the Thesaurus. Specifically, two semicolon groups have been defined as being related, or connected, when each semicolon group contains at least one entry identical to an entry occurring in the other. Using this definition of relatedness, the computer program written to implement the method developed for determining the connectedness of a graph representation of the Thesaurus has been run for the semicolon groups comprising *sub-classes I and II* of *Class One* of the Thesaurus. Although the connectedness characteristics of *sub-classes I and II* of *Class One* may or may not be indicative of the connectedness of the Thesaurus as a whole, there are several characteristics which I feel are likely to hold true for the entire Thesaurus.

1) there is a small number of relatively large distinct equivalence classes of categories.
2) there is a relatively large number of distinct equivalence classes consisting of only one category, or in other words, a large number of categories which are not connected to any other category.
3) there is a fairly strong correspondence between the formal organization of the Thesaurus, as represented by divisions within its hierarchy, and clusters of connected categories.

In the subclasses I and II data, there exists one equivalence class of categories which is much larger than any of the others, and the vast majority of the equivalence classes are quite small relative to the number of categories in that portion of the Thesaurus considered. Further, there is a relatively large number of isolated categories (491). One reason for this situation can be traced to the fact that the proportion of unique entries in respect to the total number of entries in *sub-classes I and II* of *Class One* is very large. However, it appears likely that the number of isolated categories will remain relatively large when treating the Thesaurus in its entirety. The reason for making such a statement concerns a decision which was made during the parsing of the Thesaurus. Specifically, there are a large

number of *sub-categories* in the Thesaurus consisting of lists of entries (e.g., lists of capitols of the world, lists of countries, etc.). During the parsing of the Thesaurus each entry of a list was interpreted as one semicolon group. As a result, given the very specialized and unusual nature of many lists, it is likely that many entries in lists appear only once in the Thesaurus. And thus, each of these entries, interpreted as comprising a semicolon group, will result in an isolated node of the graph representation of the Thesaurus.

From an examination of the equivalence classes of categories which have been determined thus far, it appears that there exists a correspondence between groups of connected categories and the formal divisions within the organization of the Thesaurus. For *sub-classes I and II* of *Class One* this correspondence is most pronounced with respect to groups of connected semicolon groups and the *categories* comprising this portion of the Thesaurus. Specifically, the majority of the equivalence classes which have been determined thus far seem to be formed around semicolon groups contained in some one of the *categories* in the *sub-classes* of *Class One*. That is, for some equivalence class of categories, or semicolon groups, it appears to be the case that quite often the majority of these semicolon groups occur in the same *category* of the Thesaurus. In most instances very few connections, or chains of connected categories, exist between the categories, or semicolon groups, comprising two *categories* that possess labels which are antonyms of one another (e.g., the *categories* labeled "Uniformity" and "Ununiformity"). All of these results, then, would seem to indicate that the formal divisions in the organization of the Thesaurus do correspond to some extent to the totally connected groups of categories in the Thesaurus, at least at the *sub-class* level of the hierarchy of the Thesaurus. And, as a result, the conceptual organization imposed on the Thesaurus by the author(s) may in fact correspond closely to the conceptual organization reflected in its connectedness.

Considering the small portion of the Thesaurus for which connectedness has been determined thus far, these conclusions must be taken as tentative. However, it is my considered opinion that the Thesaurus, as it currently exists in machine-accessible form, is not totally connected. A definite statement concerning the connectedness of the Thesaurus, however, must wait until the techniques and methods for analyzing connectedness which were presented in this thesis are applied to the Thesaurus in its entirety. This should be the next step in the efforts which have been directed towards explicating the connectedness properties of the structure of the Thesaurus. And, considerably more research will be necessary before the structure of the Thesaurus, as evidenced in its properties of connectedness, can be explicated fully.

REFERENCES

(1) Berry, Lester V. (Ed.), Roget's International Thesaurus, Thomas Y. Crowell Co., New York, 1962.

(2) Jones, Karen Sparck, Synonymy and Semantic Classifications (Published Ph.D. Thesis), Cambridge Language Research Unit, Cambridge, England, 1964.

(3) Salton, Gerard, Automatic Information Organization and Retrieval, McGraw-Hill, Inc., New York, 1968.

(4) Vickery, B.C., "Thesaurus - A New Word in Documentation, " J. DOC., vol. 16, no. 4, December 1960.

(5) Dillon, Martin, and David J. Wagner, "Models of Thesauri and Their Applications: in Automated Analysis of Language Style and Structure in Technical and Other Documents, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1971.

(6) Blagden, J.F., "Thesaurus Compilation Methods: A Literature Review," ASLIB Proceedings, vol. 20, no. 8, August 1968.

(7) Neufeld, Margaret L., "Linguistic Approaches to the Construction and Use of Thesauri: A Review," Drexel Library Quarterly, vol. 8, no. 2, April 1972.

(8) Rogers, V.G., "Thesaurus Construction: An Introduction," Drexel Library Quarterly, vol. 8, no. 2, April 1972.

(9) Rolling, Loll N., "Compilation of Thesauri for Use in Computer Systems," Information Storage and Retrieval, vol. 6, no. 4, Oct. 1970.

(10) Dattola, R.T., and D.M. Murray, "An Experiment in Automatic Thesaurus Construction," Scientific Report No. ISR-13 to NSF, Cornell University Department of Computer Science, Gerard Salton, Dec. 1967.

(11) Reisner, Phyllis, "A Note on Minimizing Search and Storage in a Thesaurus Network by Structural Reorganization of the Net" in Some Problems in Information Science, M. Kochen, Scarecrow Press, Inc., New York, 1965.

(12) Warfel, Samuel, "The Value of a Thesaurus for Prefix Identification" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1972.

(13) Kochen, M., and R. Tagliacozzo, "A Study of Cross-referencing," J. Doc., vol. 24, no. 3, Sept. 1968.

(14) Bryan, Robert, "Abstract Thesauri and Graph Theory Applications to Thesaurus Research" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1973.

(15) Kochen, M. (Ed.), Some Problems in Information Science, Scarecrow Press, Inc., New York, 1965.

(16) Tutte, W.T., Connectivity in Graphs, University of Toronto Press, Toronto, Canada, 1966.

(17) Maxwell, Lee M., and Myril B. Reed, The Theory of Graphs: A Basis for Network Theory, Pergamon Press, Inc., New York, 1971.

(18) Harary, Frank, Graph Theory, Addison-Wesley Publishing Co., Reading, Massachusetts, 1969.

(19) Abraham, C.T., "Techniques for Thesaurus Organization and Evaluation" in Some Problems in Information Science, M. Kochen, Scarecrow Press, Inc., New York, 1965.

(20) _____, "Graph Theoretic Techniques for the Organization of Linked Data" in Some Problems in Information Science, M. Kochen, Scarecrow Press, Inc., New York, 1965.

(21) Taylor, Scott R., "Handling of Bracketed Information" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1973.

(22) Dillon, Martin, "Automated Disambiguation of Thesaural Categories," an unpublished report, Department of Computer Science, School of Library Science, University of North Carolina, 1973.

(23) Gotlieb, C.C., and S. Kumar, "Semantic Clustering of Index Terms," J. ACM, vol. 18, no. 4, Oct. 1968.

(24) Reisner, Phyllis, "Semantic Diversity and a 'Growing' Man-Machine Thesaurus" in Some Problems in Information Science, M. Kochen, Scarecrow Press, Inc., New York, 1965.

(25) Harris, Herbert R., "Further Editing of Roget's Thesaurus Tape and Some Observations of Further Studies of the Thesaurus" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1972.

(26) _____, "The Conversion of Roget's International Thesaurus to an Automated Data Base" in Automated Language Analysis, Sally Yeates Sedelow, University of Kansas, Lawrence, Kansas, 1973.

(27) Ramamoorthy, C.V., "Analysis of Graphs by Connectivity Considerations," J. ACM, vol. 13, no. 2, April 1966.

(28) Welch, John T., Jr., "A Mechanical Analysis of the Cyclic Structure of Undirected Graphs," J. ACM, vol. 13, no. 2, April 1966.

(29) Gotlieb, C.C., and D.G. Corneil, "Algorithms for Finding a Fundamental Set of Cycles for an Undirected Linear Graph," Comm. ACM, vol. 10, no. 12, Dec. 1967.

(30) Paton, Keith, "An Algorithm for Finding a Fundamental Set of Cycles of a Graph," Comm. ACM, vol. 12, no. 9, Sept. 1969.

(31) _____, "An Algorithm for the Blocks and Cutnodes of a Graph," Comm. ACM, vol. 14, no. 7, July 1971.

APPENDIX A

AMOUNT OF COMPUTER STORAGE NECESSARY FOR SELECTED N-NODE GRAPH SIZES

No. of Nodes	Max. No. of Words/Row	No. of Words in Matrix
200	6	740
400	12	2776
600	17	5496
800	23	9908
1000	28	14608
1200	34	21396
1400	39	28076
1600	45	37240
1800	51	47700
2000	56	57440
2200	62	70276
2400	67	81996
2600	73	97208
2800	78	110908
3000	84	128496

APPENDIX B

SAMPLE OF THE SORTED ENTRIES AND THE CODES FOR THE CATEGORIES IN WHICH THEY OCCUR

9002003	6APPLICABILITY
26005002	6APPLICABILITY
9008001	6APPLICABLE
26018001	6APPLICABLE
9002003	6APPLICATION
9004002	6APPLY
9003001	6APPLY TO
9008001	6APPLYING TO
9008001	6APPOSITE
26018001	6APPOSITE
9009002	6APPOSITELY
9002002	6APPOSITENESS
26005002	6APPOSITENESS
9002002	6APPOSITION
26005002	6APPOSITION
3004003	6APPRECIABLE
20001004	6APPROACH
20007006	6APPROACH
26019002	6APPROPRIATE
26005001	6APPROPRIATENESS
9007001	6APPROXIMATE
20007006	6APPROXIMATE
20008001	6APPROXIMATE
20014001	6APPROXIMATE
9007001	6APPROXIMATING
20014001	6APPROXIMATING
9001005	6APPROXIMATION
20001004	6APPROXIMATION
9007001	6APPROXIMATIVE
20014001	6APPROXIMATIVE
6002003	6APPURTENANCE
9008001	6APPURTENANT
9005004	6APPURTENANT TO
9008001	6APROPOS
26018003	6APROPOS
9010001	6APROPOS OF
26018001	6APT
26005002	6APTITUDE

APPENDIX C

THE SETS OF CATEGORIES CONSTRUCTED FROM THE ENTRIES OF APPENDIX B

150	2									
	9002003	1	26005002	1		0	1		0	1
151	2									
	9008001	1	26018001	1		0	1		0	1
152	1									
	9002003	0		0	0		0	0		0	0
153	1									
	9004002	0		0	0		0	0		0	0
154	1									
	9003001	0		0	0		0	0		0	0
155	1									
	9008001	0		0	0		0	0		0	0
156	2									
	9008001	1	26018001	1		0	1		0	1
157	1									
	9009002	0		0	0		0	0		0	0
158	2									
	9002002	1	26005002	1		0	1		0	1
159	2									
	9002002	1	26005002	1		0	1		0	1
160	1									
	3004003	0		0	0		0	0		0	0
161	2									
	20001004	1	20007006	1		0	1		0	1
162	1									
	26019002	0		0	0		0	0		0	0
163	1									
	26005001	0		0	0		0	0		0	0
164	4									
	9007001	3	20007006	3	20008001		3	20014001		3
165	2									
	9007001	1	20014001	1		0	1		0	1
166	2									
	9001005	1	20001004	1		0	1		0	1
167	2									
	9007001	1	20014001	1		0	1		0	1
168	1									
	6002003	0		0	0		0	0		0	0
169	1									
	9008001	0		0	0		0	0		0	0
170	1									
	9005004	0		0	0		0	0		0	0
171	2									
	9008001	1	26018003	1		0	1		0	1
172	1									
	9010001	0		0	0		0	0		0	0
173	1									
	26018001	0		0	0		0	0		0	0
174	1									
	26005002	0		0	0		0	0		0	0

Professional Activities of Project Personnel

Sally Yeates Sedelow

Publications

Automated Language Analysis, Report on Research for the Period September 1, 1972--August 31, 1973, Contract N00014-70-A-0357-001, Office of Naval Research, University of Kansas, 303 pp.

"The Use of the Computer for Stylistic Studies of Shakespeare," Computer Studies, Vol. IV, No. 1 (June, 1973), pp. 33-36.

"Literary Text Processing" (Abstract), IFIP Conference Proceedings, Vol. 42, (1973), p.8.

Lectures

"Literature, Fine Arts and the Machine," Phillips Fund Lecture, Haverford College, April, 1974.

"Natural Language Patterning," SUNY, Albany, January 1974.

"One Culture," Modern Language Association Annual Meeting, December, 1973.

"Patterns in Natural Language," Psychology Pro-Seminar, University of Kansas, December, 1973.

Activities

Senior Fulbright-Hays Program, Advisory Screening Committee in Computer Science, 1973--.

Research Review Panel, National Endowment for the Humanities, 1973--.

Co-Editor, Computer Studies in the Humanities and Verbal Behavior, 1966--.

Advisory Committee for Computing Activities, National Science Foundation, 1972-1973.

Committee on Information Technology, American Council of Learned Societies, 1970-1973.

Co-Principal Investigator, NSF Study re Possible National Center/Network for Computational Research on Language, 1971-73.

Advisory Committee, Computer Application Section, Midwest Modern Language Association, 1973--.

Invited Participant, Conference on Data Dissemination for Carcinogenesis, National Cancer Institute, Jan. 29-Feb. 1, 1974.

Chairman, Panel on the Use of Technology in the Delivery of Educational Services, Productivity Planning Conference, National Institute of Education, Feb. 5-7, 1974.

Chairman, Computer Application Section, Midwest Modern Language Association, 1973-1974.

Education Committee, American Federation of Information Processing Societies.

Member, Modern Language Association ad hoc planning committee for National Archive for Literary Texts, 1973--.

Reviewer, IEEE Transactions.

Reviewer of Papers, IFIP Congress, 1974.

Reviewer of Papers, National Computer Conference, 1973--.

Proposal Evaluation, Canada Council.

Proposal Evaluation, National Endowment for the Humanities.

Proposal Evaluation, Special Projects Program, NSF.

Proposal Evaluation, Division of Social Systems and Human Resources, NSF.

Faculty Senate Committee on Scholarly Publication, Kansas University, 1971--1973.

Graduate Council, University of Kansas, 1973--1974.

Graduate Area IV Committee, University of Kansas, 1971-1974

Affirmative Action Committee for Vice-Chancellor for Academic Affairs, University of Kansas, 1973--.

University Council, University of Kansas, 1974--.

WALTER A. SEDELOW, JR.

PUBLICATIONS:

Current Trends in Language Research and the Computer, contributing author and (With Sally Yeates Sedelow) editor, (contracted for by Mouton & Co., The Hague, The Netherlands).
"Bibliography for a Science of Language," Computer Studies in the Humanities and Verbal Behavior. (In Press, 1974).
"A Moment of Cultural Shifting" (essay review), The Virginia Quarterly Review, 50, (4), 1974, pp. 627-631.

PAPERS/SEMINARS/ADDRESSES/etc:

Panelist, "Are the Arts Necessary?: The Arts, the Community, and You," Lawrence Art Guild, Lawrence, Kansas, February, 1974.
"Rhetoric and History," Phi Alpha Theta, Annual Nebraska Regional Meeting, Nebraska Wesleyan University, Lincoln, April, 1974.
"Literature and Society," English Department Dissertation Seminar, University of Kansas, October, 1974.
Panelist, "Dress Style/Life Style," University of Kansas Museum of Art, October, 1974.

ACTIVITIES:

Member, Committee for the Program in the History and Philosophy of Science, K.U.
Referee, National Science Foundation, 1974
Referee, National Endowment for the Humanities, 1974.
Referee, U.S. Technical Program Committee for IFIP Congress 74, 1974.
Chairman, University of Kansas Chapter, AAUP Committee on University Government, 1972-74; Vice President, University of Kansas Chapter, 1974-75.
Member (& Panel Chairman), Ad Hoc Faculty Interview Committee for Rhodes, Marshall, & Danforth Fellowships, University of Kansas, 1974.
Member, Faculty Advisory Committee, African Studies, University of Kansas, 1973--.
Member, Advisory Committee, Alternative Approaches to the Management and Financing of University Computer Centers Project (NSF), Denver Research Institute, University of Denver, Colorado, 1971-1974.
Board of Editors, Computer Studies in the Humanities and Verbal Behavior, 1966--.
Advisory Board, Historical Abstracts, 1973-81.
Member, Advisory Council, Computer Studies Institute, International Academy at Santa Barbara, California, 1972--.

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) University of Kansas Lawrence, Kansas 66045	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

Automated Language Analysis

4. DESCRIPTIVE NOTES (*Type of report and, inclusive dates*)

5. AUTHOR(S) (*First name, middle initial, last name*)

Sedelow, Sally Yeates

6. REPORT DATE 31 December 1974	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) (*Any other numbers that may be assigned this report*)	

10. DISTRIBUTION STATEMENT

Distribution of this report is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY

13. ABSTRACT

This report includes (i) a summary of research pursued under this contract; (ii) a description of the editing of a computer-accessible version of Roget's International Thesaurus; (iii) a discussion of mathematical approaches to the modelling of Thesauri, with Roget's serving as an instantiation; (iv) a Master's Thesis exploring graph theory applications to the study of the structure of Roget's. Articles include "Brief Overview of Research Under This Contract" by Sally Yeates Sedelow, "The Automated Version of Roget's International Thesaurus: A Description with Suggestions for Future Editing" by Herbert R. Harris, "Further Discussion of the Use of Brackets and Parentheses in Roget's Thesaurus" by Scott Taylor, "Modeling in Thesaurus Research" by Robert Bryan, and "Selected Graph Theory Applications to a Study of the Structure of Roget's Thesaurus: A Data Base for Automated Language Analysis" by Scott Taylor.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Automated Language Analysis Stylistic Analysis Thesauri Prefixing Content Analysis						

DD FORM 1 NOV 65 1473 (BACK)

S/N 0101-807-6821