AUTOMATED LANGUAGE ANALYSIS

1972-1973

Report on research for the period

September 1, 1972 - August 31, 1973

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

Approved for public release; distribution unlimited

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

AUTOMATED LANGUAGE ANALYSIS

1972-1973

Report on research for the period

September 1, 1972 - August 31, 1973

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

Approved for public release; distribution unlimited

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

AUTOMATED LANGUAGE ANALYSIS

1972-1973

Report on research for the period

September 1, 1972 - August 31, 1973

Sally Y. Sedelow, Principal Investigator
Walter Sedelow, Consultant
Robert Bryan
Herbert Harris
Peggy Lewis
Scott Taylor
Sam Warfel

The views, conclusions, or recommendations expressed in this document do not necessarily reflect the official views or policies of agencies of the United States Government.

TABLE OF CONTENTS

I. Introduction 3

II. A. The Conversion of _Roget's International Thesaurus_ to an Automated Data Base 5

B. Handling of Bracketed Information 29

C. Etc. in _Roget's International Thesaurus_ 35

D. Abstract Thesauri and Graph Theory Applications to Thesaurus Research 45

III. Studies in the Semantics of English Prefixation 91

IV. VIA User's Manual 173

V. Professional Activities of Project Personnel 299

ABSTRACT

This report covers the following topics: the editing of Roget's International Thesaurus, the mathematical modelling of thesauri, and a user's guide to the VIA content analysis programs as implemented at The University of Kansas. Articles in the report concerned with the practical and theoretical issues raised by the effort to edit the Thesaurus include, "The Conversion of Roget's International Thesaurus to an Automated Data Base," by Herbert Harris, "Handling of Bracketed Information," by Scott Taylor, and "Etc. in Roget's International Thesaurus," by Sally Yeates Sedelow. "Abstract Thesauri and Graph Theory Applications to Thesaurus Research," by Robert Bryan explores the mathematical modelling of thesauri. Robert Bryan and Peggy Lewis have provided the user's guide to the VIA programs.

PREFACE

I would like to express my appreciation to the Information Systems staff, Office of Naval Research, for their administrative support. Thanks are also due Mr. Mark Katz, University of Kansas Computation Center, who has been unfailingly helpful to all of us associated with this research effort. The support of the Departments of Computer Science and Linguistics should also be acknowledged. Finally, special thanks are owing Ms. Melanie Mitchell who has served both as secretary and occasional research assistant on this project.

INTRODUCTION

This project has been concerned with the comprehensive identification of characteristics of style in spoken and written language. Style has been identified so as to include content, and one major thrust of this project has been the development of programs to get at the content, or semantic, aspects of language generation. The development of these programs (which, taken together, are called the VIA programs) has also resulted in research on thesauri; this past year that aspect of the project was given greatest emphasis. Section II of this document reports on that effort. The first three articles--by Harris, Taylor, and Sedelow--describe both practical and theoretical problems and issues raised by our effort to get a specific thesaurus--Roget's International Thesaurus--into computer-accessible form. The fourth paper in that section--"Abstract Thesauri and Graph Theory Applications to Thesaurus Research"--by Robert Bryan, addresses the general problem of modelling a thesaurus, for which Roget's would serve as an instantiation. This research has been undertaken in order to attempt to define the structure of Roget's, so as to understand its biases and, thus, ascertain desirable modifications to that thesaurus as it stands as well as to ascertain desirable characteristics in an "ideal" thesaurus.

Since a thesaurus provides networks of semantically-related words, the utility of this work, and of the thesaurus which will emerge from it, for computer-based language processing of many different types is considerable. Thesauri are defined across time and across the usage of a culture and thus provide outlines of language usage appropriate for applications ranging from computer-assisted instruction to artificial intelligence cognitive model approaches.

The paper by Warfel, Studies in the Semantics of English Prefixation, grows out of our interest in developing theory to guide and perhaps buttress ad hoc decisions we and others have made as to when initial strings of characters in words are serving as prefixes and when they are not. We have a need in our language-analytic programs, as is the case for many other research scientists and scholars, to pull together words having common roots. In order to do so, it is necessary to separate character strings serving as roots from those which serve as prefixes or suffixes. When we determined to deal with prefixes and searched for relevant theory to aid us in our decision-making, we discovered relevant theory to be almost non-existent. Warfel's paper--his doctoral dissertation--represents an effort to explore some aspects of the theory of prefixation using a generative semantics approach. We hope that Warfel's research and that of others who may have an interest in prefixation will provide some leads toward the development of theory which can then be drawn upon for the development of algorithmic approaches to the identification of prefixes.

The fourth section of this report consists of a User's Manual for the VIA programs as they have been implemented on the Honeywell 635 at the University of Kansas. We feel that this Manual will considerably ease access to our programs by the uninitiated user; we have also provided examples of output so that the user will know what to expect at major points in the operation of the programs.

II.A. The Conversion of Roget's International Thesaurus to an Automated Data Base

by Herbert Harris

The editing of Roget's International Thesaurus* during the past year has been directed toward the goal of making the Thesaurus usable in an automated system.

This goal has required two types of editorial change: One type of change is to reformat the text to make the entries in the Thesaurus and their locations readily accessible. The aim of this reformatting is a parsed version of the Thesaurus separating out each entry together with information about the type of entry and its location in the Thesaurus. The other type of editorial change needed to make the Thesaurus more compatible with an automated system has involved making the implicit distinctions in the Thesaurus explicit. Most human beings using the Thesaurus have available an implicit system or systems of interlocking sub-systems which allow cross-checking of entries against dictionary entries, against parsing routines, and against trial output vis-a-vis memory of encountered input. But for limited automated systems, there can be no reliance on these implicit systems. Examples of distinctions existing only implicitly in the Thesaurus are:

1. The simple collapsing, using an _or_ connector, of several entries--for example, "put _or_ get one's Irish up." The two entries are

* Roget's International Thesaurus, 1962, ed. by Lester V. Berry, New York: Thomas Y. Crowell Company.

"put one's Irish up" and "get one's Irish up." In sixty percent of the cases using an <u>or</u>, the parsing of the multiple entries is like the above example, with one word (get) being replaced by another (put) to form another entry. But in forty percent of the <u>or</u> entries it is necessary to call upon auxiliary systems, such as the dictionary. The entry "Throw away <u>or</u> waste the opportunity" does not satisfy the previous 'replacement' algorithm. In order to isolate the proper entries one must know from a dictionary that "throw away" is a two-word verb. The two entries for "two shakes of a dead sheep's tail or brass monkey's tail" should be "two shakes of a brass monkey's tail" and "two shakes of a dead sheep's tail." Even though this entry is not algorithmic in the way stated above, a speaker of English would probably correctly sort out both entries. A knowledge of the idioms of the language might allow recognition of one entry. The second could then be patterned after the first. This example also shows that even if the idioms were unknown to a person looking at the *Thesaurus* it would be possible to separate out the two entries by parsing, since one noun phrase is being substituted for another.

2. Another implicit distinction entails the use of the hyphen at the ends of lines in the *Thesaurus*. For example:

```
....................per-
iod.....................
..................self-
sacrificing............
```

The hyphen in self-sacrificing must stay and the hyphen in period must go. A knowledge of spelling conventions is required to differentiate these two cases.

Four levels of Thesaurus editing have been defined and are being generated. The first is an exact copy of the text in computer-accessible form. All the problems outlined above and those described in last year's report (Herbert Harris, "Further Editing of Roget's Thesaurus Tape and Some Observations on Further Studies of the Thesaurus," in Sally Yeates Sedelow, *et al*, *Automated Language Analysis*, Lawrence, Kansas: The University of Kansas, Department of Computer Science, 1972, pp. 19-30) must be faced by anyone using this version.

For the second level, extensive editing has taken place. The text is still in paragraph format and no text has been deleted. But (to be described below) paragraphs and individual entries are clearly delimited and some inconsistencies have been eliminated.

The third version is a parsed version of the *Thesaurus*. Each entry of the *Thesaurus* appears on one line together with all the necessary information about the type of entry and its location in the *Thesaurus*; all of the bookkeeping information necessary to use the *Thesaurus* appears with the entry.

The fourth level, an index, is a sorted and collapsed version of level three. In this version, all unique character strings appear only once. Associated with each string is information about all its occurrences in the text.

The editing to establish level one was completed in 1972. The editing that has taken place to produce level two began with the elimination of all the hyphens at the ends of lines. About 16,000 lines (out of a total of about 77,000 lines) ended in a hyphen. These hyphenated lines were all examined individually. Those for which decisions could not be reached immediately were looked up in *Webster's Third International Dictionary*. If the entries could not be found there, the *Oxford English Dictionary* was consulted. There were, however, thirty-eight cases that could not be found in either reference work. These were resolved by 'fiat'. The lefthand column contains the entries as they appear in the text and the right hand column contains the entries as they appear on the edited tape. If there is more than one hyphen in the entry, the underlined hyphen is the case in question.

	Text	Tape
1.	stark-staring	stark-staring
2.	take-away (subtraction)	take-away
3.	sub-algebra	subalgebra
4.	seven-out (dice)	seven-out
5.	dragged-out	dragged-out
6.	non-habitable	non-habitable
7.	wizen-faced	wizen-faced
8.	starved-looking	starved-looking
9.	lath-legged	lath-legged
10.	bead-like	bead-like
11.	air-woman	air-woman
12.	brolly-hop (parachute jump, Eng.)	brolly-hop
13.	constant-chord-rotor helicopter	constant-chord-rotor helicopter
14.	tag-tail	tag-tail
15.	rightabout-face	rightabout-face
16.	inward-bound	inward-bound
17.	outward-bound	outward-bound
18.	vol-au-vent	vol-au-vent

19.	acid-base	acid-base
20.	IDA (Integro-Differential...	IDA (Integro-Differential...
21.	ultra-masculinity	ultra-masculinity
22.	non-understanding	non-understanding
23.	lack-brained	lack-brained
24.	thick-pated	thick-pated
25.	stark-mad	stark-mad
26.	after-mirage	after-mirage
27.	sky-aspiring	sky-aspiring
28.	so-so-ish	so-so-ish
29.	wee-wows	wee-wows
30.	fango-therapy	fango-therapy
31.	senatus-consult	senatus-consult
32.	bark-bound	bark-bound
33.	ix-nay	ixnay
34.	peel-house	peel-house
35.	wag-wit	wag-wit
36.	same-someness	same-someness
37.	scare-sinner	scare-sinner
38.	marrow-bones	marrow-bones
39.	Dad-blame	Dad-blame
40.	love-pot	lovepot
41.	abba-comes	abbacomes

Hyphen usage is not consistent from one reference work to another. Therefore, in text processing it would appear that in comparing an incoming text against a reference data base it will be necessary to make systems 'aware' that three possibilities involving hyphens can occur. Two words may be either two separate words, two words combined with a hyphen, or a single compound word. Sometimes these three alternatives may indicate (1) a difference in meaning, (2) a difference in syntactic construction, or (3) simply no difference at all except in rhetorical style. Some general statements can be made about the use of the hyphen, but these reflect only general tendencies and are not consistently applied by the users of the language. Some examples of the differences between reference works will illustrate. *Roget's* has hyphens in the following two constructions: 'shell-shock' and 'rock-bottom', but *Webster's Third International* has entries for each of the constructions as two separate words. The dictionary actually has two listings for the above words. The noun entry for *rock bottom* appears as two words, but for the adjective entry the two words are hyphenated. For *shell-shock* the noun is two words and the verb is hyphenated. As a general tendency, either a noun entry or a verb entry used as preposed adjective (*e.g.*, run-down as in 'a run-down shack') will be hyphenated in the dictionary. Two-word verbs used as nouns often have a hyphen, *e.g.*, in *Webster's* 'sell off' goes to 'sell-off' but, in contrast to this general pattern, 'sell out' goes to 'sellout'. It may be that the more frequent the word's usage the more likely it is to be spelled without a hyphen. An example of this possibility is the list of words in the *Thesaurus* with 'pan-' in front of them. All have a hyphen (*i.e.*, pan-American, etc.) except for 'panhellenic' which does not. Hyphens may also appear in constructions that will never appear in any reference work, *e.g.*, "a never-to-be-forgotten moment."

These variant uses of the hyphen indicate that for most economical processing, the reference work should be made to conform as closely as possible to the conventions of the incoming text. But some account will have to be taken of hyphens as a possible cause for a lack of matching.

The hyphen also causes another problem in handling text material in automated processing. In reference works, hyphenated forms are listed after the same words without hyphens. Also, words with hyphens and those having the same first word but with a suffix are interspersed, ignoring the hyphen as a character. For example:

rundlet
run down
run-down
rundown
rune

Therefore, the hyphen and the blank cannot simply be ignored in alphabetizing. The aim in sorting the Thesaurus entries will be to use a traditional sort order so as not to put people off by forcing them to adjust to an 'odd' alphabetizing order.

The second step in editing the Thesaurus (for level 2) was to reformat lists. These are collections of terms that comprise generic categories.

205.8 fibers,threads

acetate rayon
Acrilon
Acralac
.
.
.

near-silk (chiefly coll.)
nylon
.
.
.

The lists in the book were formatted as double columns with entries alphabetized from top to bottom of the first column and then the alphabetical listing continued from top to bottom in the second column. Since one line from the book was key-punched on one card image, this caused the two entries in these lists on a card image to be related in a complex way. If an entry had to be continued for more than one line, then subsequent parts were indented on succeeding lines. These lists have been reformatted so that a single alphabetical entry appears on a single card image. The indented material has been raised so that all of one entry appears on one card image. Hyphenated words also occurred in these lists; these hyphens, too, were eliminated.

The next step was to eliminate the *or*'s from the *Thesaurus*. As mentioned earlier, *or*'s were used as a shorthand notation for multiple entries. For example, 'sober *or* grim reality' is really two entries, 'sober reality' and 'grim reality.' The *or* used in connection with a tilde was a shorthand for three or more entries. 'nice , fine , delicate *or* subtle distinction,' is really four entries, 'nice distinction, fine distinction, delicate distinction, subtle distinction.' Most of the time, as in the above examples, words immediately to the right or left of the *or* are substituted in some larger expression. The items around the *or* might occur at the left of a larger expression as in the above examples, or they might appear at the right of a larger expression as in 'the other side of the picture *or* coin.' Or they might be embedded in the middle of a larger expression as in 'all to stick *or* sticks and staves' which I take it should be expanded as 'all to stick and staves' and 'all to sticks and staves.' It is also the case that the *or* might occur with different forms of one word, as in 'transcendence *or* transcendency.'

The *or* entries also interact with other elements of the *Thesaurus*. For example, they can be followed by brackets as in 'break to smithers *or* smithereens [coll.].' In these cases, most of the time the bracket has to be distributed to each of the entries formed out of the *or* entry. However, an exception is 'apropos $*or* á propos# [F.].' (@ before a character string indicates that the following characters are in boldface and $ indicates italics.)

The *or* entries also interact with the operators that indicate the typeface used in the *Thesaurus*. '@bear on# $or# @upon#' is an example where both sides of the *or* will be boldface, resulting in '@bear on#' and '@bear upon#.' However, it is not always the case that all entries formed from the *or* will have the same type face. In the example above, with 'apropos,' one entry will be in a normal typeface and the other will be in italics. An example with boldface is '@come to nothing# $*or*# naught' which will generate one entry in boldface and the other not.

The program to eliminate the *or*'s associated with brackets and typeface operators would have been complicated, but possible, if the use of the *or* had been consistent throughout the *Thesaurus*; however, this is not the case. Some examples will suffice to demonstrate the difficulties.

1. 'break to flinderation $or# all to flinderation [slang].' Simply interchanging the word around the *or* will give 'break to flinderation to flinderation [slang]' and 'break to all to flinderation [slang],' when it should be 'break to flinderation [slang]' and 'break all to flinderation [slang].'

2. 'Tote $or# tote up' will give 'tote up' and 'tote up' when it should be 'tote' and 'tote up.'

3. '@apropos# $or# á propos# [F.]' (assuming you could resolve the type conflict) will give 'apropos propos [F.]' and 'á propos [F.]' instead of simply deleting the *or*. Since there are entries where the *or* has simply to be deleted and a following bracket distributed such as 'nunks $or# nunky [slang],' the *apropos* case is also an exception to the protocol for bracket distribution.

The *or* can be viewed as simply a shorthand notation (albeit non-algorithmic) for multiple entries with the same repeating partial strings. But, in looking over the entries with *or*'s, the impression is that in a lot of cases the *or*, while saving space, also has a semantic function. However, this semantic function is not consistently carried out.

Before discussing the semantic function of *or*, some understanding of the structure of the *Thesaurus* is necessary. With reference to paragraph level subdivisions, there are two formats in the *Thesaurus*. One is a paragraph of items such as:

Nouns 1. @existence,# subsistence, @being;# entity, essence; @occurrence,# presence; @life#/406.

The other is a list of items.

7. Fasteners

anchor
band
bar
barrette
bellyband
belt
bind
binding
etc.

The paragraph is divided into sections with semicolons. The entries inside the semicolons may be substituted for each other to form paraphrase equivalents. For example, using the paragraph above, one might write:

1a. The occurrence of a boat at that time of night seemed strange.
1b. The presence of a boat at that time of night seemed strange.

But, the words in that paragraph cannot be completely synonymous since they cannot be substituted for each other everywhere, _e.g._,

2a. His presence was duly noted.
??2b. His occurrence was duly noted.

These entries then may be nearly synonymous in some contexts, but do not have the same distributional patterns. Thus, they are not usually completely synonymous. But what if two entries _do_ share the same privilege of occurrence? How would this be indicated? This indication is at least part of the use of the _or_.

What follows is a taxonomy of these semantic uses of the _or_.

1. Spelling variation, 'aught _or_ ought.' Two entries for a single phonetic referent and, therefore, a single semantic referent would result when the spelling conventions for some word are not firmly established or there are different spelling conventions competing with each other.

2. There are spelling variants that seem to imply a difference in pronunciation. Such a difference could result from regional or social pronunciation differences reflected in the spelling system.

Interpreting the semantics of a lexical item to be included in the *Thesaurus* very broadly to include potentially all information conveyed by the use of that item, the use of one pronunciation variant as opposed to another will convey some special social dialect information about the person using that lexical item. (It should be noted that this implied definition for a pronunciation variant is different from that actually adopted in the edited form of the *Thesaurus*. This distinction implies a real pronunciation distinction. There may, however, be spelling variants which do not in fact represent pronunciation variation. These differences in written form are used by authors to signal social or geographical distinctions about characters using the forms. Since the forms imply the same distinctions as actually occurring distinctions, these strictly literary forms are also labeled pronunciation variants in the edited form of the *Thesaurus*.) The denotation of such a pair, however, remains the same, thus justifying the use of the *or*. Such a pair is 'kilter *or* kelter [dial. & coll.].'

3. Translation equivalents like @apropos# $or á propos# [F.].

4. Paraphrastic equivalents like 'remotely *or* distantly related.' One type of these equivalents results from the substitution of a synonym in some phrase as above. Another type occurs when the image of a metaphor is changed but the same meaning is intended, *e.g.*, 'stick to like a barnacle *or* leech.'

5. Morphological variants that do not have readily ascertainable differences, *e.g.*, 'transcendence *or* transcendency,' 'divy *or* divvy up [slang].'

6. Alternate names for the same item or referent like 'April Fool *or* All Fools' day,' 'ground-hog *or* woodchuck day [U.S.].'

7. Closely related pairs of words like 'foster brother *or* sister.' Notice that these last are not closely related semantically. A foster brother is not a foster sister, although they are both offspring. The relationship between them is generic and not substitutable, whereas all of the former are more or less substitutable.

In order to make what is an entry in the *Thesaurus* more explicit, the *or*'s had to be edited out. This was especially the case since their use was not algorithmic syntactically or semantically. But, in order not to lose some of the relationships between words that they connected, it was necessary to introduce some additional classificational categories. Three types of completely synonymous expressions have been explicitly indicated in the edited version of the *Thesaurus*. These are: (1) words that are spelling variants of each other, (2) words that are pronunciation variants of each other, and finally (3) entries where one of the entries is an abbreviation of another.

One example of spelling variation was given earlier; another is 'rhyme *or* rime.'

Pronunciation variation, of course, always implies a spelling variation in the *Thesaurus* entries. (There, of course, may be pronunciation variation for a word with a consistent spelling.) The attempt, in editing the *Thesaurus*, is to indicate which words are used in print to indicate pronunciation variation. The word, in fact, may not have the pronunciation variation implied by the letter sequence used. For example, the words 'highfalutin' and 'highfaluting' are indicated as being pronunciation variants of each other. This word is probably never pronounced in the way the presence of the *g* implies. But, writers using these two forms probably mean to imply some 'folksy' as opposed to standard pronunciation for these two forms.

The pronunciation variation probably will imply one of two distinctions. It will be used to indicate either (1) a geographical dialect for some speaker or (2) will be used to indicate some social dialect. The example of 'highfaluting' might be used as an indication of such a social dialect. On the other hand, the word, itself, might be a regionalism. To be complete, the presence of an indication that two words are pronunciation variants of each other should have an indication of what the distinction implies. For geographical dialects, a region should be indicated and for social dialects, a social level indicated. With these distinctions present, it would be possible in text processing to ascertain the presence of these dialects in the text and, if a system is able to ascribe speech segments to specific characters, to make social or dialect judgments about those characters in machine processing.

In the unedited *Thesaurus*, there are social or usage distinctions indicated in the labels, [slang], [coll.], [dial.] , etc. There is, however, no indication of what the explicit difference between these designations is. With the appearance of the *Dictionary of American Regional English* in 1976 there is the possibility that words can be properly tagged and, perhaps, the Thesaurus can be expanded to include regionalisms that appear in print. The hope is that the inclusion of the distinction of pronunciation and spelling variation will give, at some later date, an entreé to the addition of distinctions that may be social or geographical and further expand the usefulness of the Thesaurus in text processing.

The indication that some entries are abbreviation variants of other entries has not been implemented. Some of the cases are juxtaposed with an or between them, but not all of them occur this way. These will turn up when the parsing out of every separate entry in the Thesaurus begins, and they will then be distinguished explicitly.

The distinction of a spelling or pronunciation variation has not been used with respect to multi-word entries. These might well have been included, especially in the case where the difference between two entries is a pair of words that have been previously viewed as spelling variants of each other. There are also cases where alternate names for the same referent occur in multi-word entries, for example, 'hard of hearing, dull of hearing, thick of hearing.' These also could have been indicated as being spelling variants of each other. But this seemed either to be unnecessary or to involve a gratuitous assumption. First, it is not necessary to indicate phrases as being spelling variants of each other if they only differ by words that have previously been indicated as spelling variants of each other. Second, to indicate variants that are near paraphrases as spelling variants would be to assume gratuitously that the paraphrases are completely synonymous.

Therefore, these multi-word entries were left as separate, but closely related, entries.

The actual editing of the *or*'s has progressed in a number of stages. First, a search of the entire text was made for all the *or*'s, getting at enough of the context to be able to make judgments about the *or* entry. These entries were then examined to identify those that did not conform to the majority algorithm (see pages 12-13). These nonconforming *or*'s were then edited with a time-sharing text-editing system here at KU.

The time-shared editing of the *or*'s was performed by two people, and it seemed advisable to develop some rules of thumb for the editing process to eliminate as much inconsistency as possible. The rules of thumb were as follows:

1. As indicated above, no idiom (multi-word entry) is a spelling or pronunciation variant of any other.

2. Words that vary only by *i* and *e* as the first letter are spelling variants of each other, *e.g.*, entrench vs. intrench.

3. Words that vary only by having a capital as the first letter of one of the entries are spelling variants of each other.

4. Words that are identical except for a final *e* in one and a final *a* in the other are two entries.

5. Two entries where the first has spaces between the words and the second is written as a compound word are spelling variants of each other.

6. Entries differing by a known derivational suffix are two separate entries.

7. Words that differ by the use of s in one and z in the other are spelling variants of each other.

8. Cases where both entries are not in italics and one of them has accent marks and the second does not are marked as spelling variants of each other.

9. Unknown cases are looked up in the Oxford English Dictionary or Webster's Third International Dictionary.

10. Cases that are undecidable are made two separate entries.

Ultimately, any inconsistencies in the editing can be eliminated when the index of the entire Thesaurus is created, since all entries with the same character string will fall together; these can then be checked to see whether all entries with a certain character string have cross reference numbers of the same type in all locations and, if not, whether the missing locations correspond to locations where a spelling variant does not and should not occur. Spelling, pronunciation, and abbreviation variants are indicated by giving a cross reference number to the location of the variant in a given section of the Thesaurus. This limitation to a given section of the Thesaurus is necessary since the variation may be applicable only to one meaning of the word. This would be especially the case with homographs. Bure is a Scottish variant of bore, but this variant must not be confused with bure, a Fijian temple, or bure, the color of yellowish-brown. This section limitation is also necessary for abbreviations, since again there may be homographs for some abbreviated word and the abbreviation must not be interpreted as meaning what the homograph means. I do not know of an actual case, but as an illustration suppose there is an entry "gram," meaning "grandmother"; 'gr.' must not be interpreted as an abbreviation for this entry. It would not be so interpreted insofar as it is an abbreviation only for entries that are explicitly stated. These entries are indicated by co-occurrence in the same section with each other, or by hand processing.

Following the time-shared editing of the non-conforming _or_'s, a program was written to reformat the algorithmic occurrences of the _or_'s. Next, the reformatted entries for both non-algorithmic and algorithmic _or_'s were reinserted in the body of the _Thesaurus_. To give some idea of the size of the task, it might be noted that there were originally about 5800 separate _or_'s in the text. These _or_'s were embedded in a file of about 9000 card images. After editing, this file was expanded to about 15000 card images. These reformatted entries were then reinserted in the complete text.

Programs have been written for a number of editing tasks, but running the programs has awaited the completion of the editing of the _or_'s.

A change of typeface in the _Thesaurus_ as originally keypunched was indicated by inserting a special character, to act as an operator, at the beginning of a character string. For example, all the entries in the following between the @ and the # would be in boldface type.

@subsist, stand,# obtain, hold, prevail,

In order to parse out the individual entries in the Thesaurus, these operators will have to be distributed to each individual entry. A program has been written by Scott Taylor to perform this distribution.

Another editing task entails the distribution to individual entries of bracketed information at the end of a multiple-entry list. For a discussion of this task, see the article by Scott Taylor in this volume.

The uses of parentheses in the Thesaurus have not been examined in great detail and there seems to be some inconsistency between the use of brackets and parentheses. But until the parentheses can be examined more extensively, a file is being made of them and, in the present implementation of the Thesaurus, their presence will be indicated by a number arbitrarily assigned to each unique parenthesis. Three specific uses of the parentheses are being edited out and will be discussed below, but to give some idea of the range of usage of the parentheses, I provide the following examples:

Sometimes, an example is provided in a parenthesis following an entry, 'brass tacks (as, to get down to $brass tacks# [coll.]),' 'near- (as $near#-silk [coll.]).' Sometimes, grammatical information is provided as in 'antipodes (pl. used as sing.).' Sometimes a techniqual term is translated into non-techniqual language, 'lycopodiaceae (clubmosses).'

Three types of occurrences of the parenthesis are being dealt with at the present time. First, there are entries with a parenthesis enclosing a possible suffix, resulting in an entry with nearly the same meaning, 'pot(ful).' This parenthesis is really a shorthand for two entries, 'pot' and 'potful.' This type of entry also occurs with _al_ as in 'allotropic(al),' with _ic_ as in 'animist(ic),' or 'Bolshevist(ic),' and with _ed_ as in 'sway-back(ed).' A program has been written to expand this shorthand notation into explicit entries. Notice also that this notation is an alternate form for the _or_'s especially serving the same semantic purpose as the entries where there is a morphological variation, 'excrescence _or_ excrescency.' This is an example of the inconsistent use of notation to delimit semantic distinctions.

Second, as was mentioned earlier, there are two formats in the _Thesaurus_, one indicating that items are substitutable and the second indicating a generic relationship. But, apparently for aesthetic reasons, when a list of items was too short, these items were formatted as a paragraph. The paragraph was preceded by a parenthesis which contains what would be the heading of a list. An example is:

> .00823.12 (famous thieves) Robin Hood, Jesse James, Claude Duval, Bill Sikes, Jack Sheppard, Robert Macaire, Dick Turpin, Jonathan Wild, Autolycus, Macheath, Nevison, Thief of Bagdad.

These paragraph entries will be reformatted into a list format. The distinction between what paragraphs should be put into list format and what should go in a paragraph format is not as clearcut as I have made it appear here. There are many borderline cases, in part because the distinction is not used consistently. For example:

> 12.2 in-laws (coll.), @relatives-in-law,#
> wrecking crew (joc.); brother-in-law,
> sister-in-law, father-in-law.

The first part of this paragraph satisfies the criterion of substitutability while the last three entries do not. The last three entries are generically related examples of the first set of entries. Only those paragraphs that are obviously lists and are headed by a parenthesis will be reformatted.

Finally, there are paragraphs that are collections of terms that are either all slang terms or all colloquialisms. Most of the time, such entries are designated as such by a following bracket, with the designation [slang] or [coll.]. But these paragraphs have the designating term at the beginning of the paragraph in a parenthesis, for example

> 34.5 (slang) @oodles,# oodlins, gob, @gobs,#
> slather, @slathers, scads,# swad, lump, smear,
> @whole smear,# fat lot, dead loads, quite a
> shucks (U.S.).

These entries will be reformatted in the prevailing form.

A program has also been written by Scott Taylor to make the Thesaurus format more accessible to machine processing. Already, category numbers have been moved in front of each paragraph number. These numbers always have a leading zero. In the Thesaurus, a part of speech designation may precede the paragraph number. This part of speech is the part of speech for all succeeding paragraphs until another part of speech is encountered. This program will distribute the part of speech to each individual paragraph and then will invert the part of speech and the number of the paragraph. The result will be, for example, 01039.12 // nouns##. This inversion will cause every paragraph to begin with a zero and the first entry, or a parenthesis (if the paragraph has one), will begin after the space following the double pound sign. The rule that the part of speech is to be distributed does not hold universally. A list may follow any part of speech designation but such lists appear to be only nouns; they will be labeled as such.

Another editorial task concerns the various uses of _etc._ For a discussion of the editing and use of _etc._ see the article by Sally Yeates Sedelow in this volume.

When those tasks for which either programs have been written or work is underway are completed, level B of the edited _Thesaurus_ will have been achieved. This level will comprise a tape of the _Thesaurus_ that is accessible to machine processing and in which all the original textual material is present. Beyond this point, processing the _Thesaurus_ will involve replacing and drastically reformatting some of the text to make its use in an automated environment more efficient. This work involves three major tasks: (1) first, the _Thesaurus_ will be parsed, (2) second, the parsed _Thesaurus_ will be sorted, and (3) finally, a complete index of the entries will be created.

The parsing of the _Thesaurus_ involves isolating each entry in the

Thesaurus and associating with it all the information about what type of entry it is and what its location is in the Thesaurus. Enough bookkeeping information will also be included to facilitate running programs against this version. Each entry will appear with the above information on a single line. By rough estimate, there are about 310,000 entries.

Next, these entries will be sorted so that all occurrences of the same character string will occur together. The sort order is currently being investigated. As mentioned earlier, the attempt will be to maintain a more or less traditional order.

Finally, all entries with the same character string will be compressed onto one line. This will result in an index in which fixed field location designations will occur first on a line, one for each of the locations of the character string in the Thesaurus. These would be followed by a variable length field with the entry itself.

B. Handling of Bracketed Information

by Scott Taylor

During the process of editing the Thesaurus in preparation for constructing the parsed tape version, a considerable amount of attention has been focused on the problem of handling that information which appears bracketed in the Thesaurus following individual entries, as well as groups of entries. There are 17,653 occurrences of bracketed words and phrases in the printed version of the Thesaurus associated with various entries which serve as an important source of additional information pertaining to the Thesaurus entries. Of these bracketed words and phrases there are essentially five major kinds of information. These are listed below along with examples of each taken from the Thesaurus.

1) The origin and/or usage of words and phrases.

Gestapo [+Ger.]
patella [archaeol.]
inwards [coll.]
polyandrium [Gr. antiq.]

2) Geographic information pertaining to entries, primarily in terms of the countries, locales within countries, or parts of the world with which an entry is most commonly associated or in which it is most often used.

chauki [India]
vega [S. Amer.]
dead stand [So. U.S.]

3) The point in history or time associated with the usage of a particular word or phrase.

fair trade [18th-century euphemism]
flapper [1920's]

4) Translations of foreign phrases and words.

en règle [F., according to rule]
semper eadem [L., always the same]

5) Sources of quotes, primarily in terms of the authors.

"sharp-toothed unkindness" [Shakespeare]
"evening's calm and holy hour" [S. G. Bulfinch]

In addition, there are also cases of other kinds of information appearing in brackets, but it appears that the majority of these cases are errors resulting in some confusion as to the use of brackets as opposed to parentheses.

The information enclosed in brackets oftentimes consists of a combination of words and phrases (representing the kinds of information exemplified above). When such is the case, the words or phrases are separated by commas, and in some cases by semi-colons. As a result, there are many different variations and combinations of information appearing in bracketed form. Two examples are:

get one's comings [slang, U.S.]
landlubber [naut.; derog.]

In addition, this situation is compounded by the use of different forms of the same word or phrase, usually involving the use of abbreviations. For example, the word "colloquial" appears in brackets, as well as do the abbreviations "coll." and "colloq.". This lack of consistency, which can also be found to occur for other words as well, complicates the means of dealing effectively with bracketed information. Likewise, there are some differences in the "syntactic usage" of bracketed information in the *Thesaurus*. That is, whereas the majority of the instances of bracketed information are used in conjunction with a single entry, some bracketed information is designated as referring to two entries or all the entries within a single semi-colon group. These designations are made by specifying BOTH or ALL within the bracket, where the former is used to specify the two immediately prior entries and the latter to specify all entries in the semi-colon group in which the bracket information appears. In all cases thus far examined, these designations appear with the last entry of a semi-colon group.

There are a total of 700 occurrences of bracketed information modified with an ALL and 636 occurrences of bracketed information modified with BOTH. Examples of the use of each follows.

; apple, biscuit, nubbin [all slang];
; belly-buster, belly-whopper [both dial.];

Using this convention it can be seen that it would not always be possible to extract an entry from the Thesaurus and obtain with it related bracketed information. As a result, and in anticipation of the need for this capability when ready to parse the Thesaurus in terms of its entries, a program has been written and debugged that "distributes" the bracketed information which appears modified by an ALL or BOTH. For the examples given above this program will produce the following.

; apple [slang], biscuit [slang], nubbin [slang];
; belly-buster [dial.], belly-whopper [dial.];

Another implication of these changes in the form of presentation of bracketed information is that the number of bracketed words and phrases occurring in the Thesaurus will undoubtedly increase significantly.

As a part of the editing of the Thesaurus, a tape file of all bracketed words and phrases was constructed along with their respective locations in the Thesaurus in terms of the category and paragraph in which they occur. It was from a listing of this file that the number of occurrences of bracketed information which was given earlier (17,653) was obtained. Following the "distribution" of brackets, as described above, this number will be incorrect. However, from this initial listing of bracketed information it has been possible to construct an additional file consisting only of the unique or distinct bracketed words or phrases occurring in the Thesaurus and assign a unique numeric identifying code to each. When generating this file, the occurrences in brackets of ALL and BOTH were ignored, thus permitting all bracketed information modified in this way to be accounted for in determining the unique words and phrases. During this process the total number of times each distinct bracketed piece of information occurred in the Thesaurus, as well as the total number of characters comprising the occurrences of bracketed information was computed. It was found that there exist 1,561 distinct brackets, with a total number of characters comprising all occurrences of bracketed information in the Thesaurus of 125, 173. Again, this figure is based on the number of brackets in the original file (17,653). And since this number will increase as a result of the "distribution" of the cases of brackets utilizing ALL and BOTH, the eventual total "character count" will also increase, probably significantly. The concern for the number of characters comprising all bracketed information in the Thesaurus arises from the desire to estimate the amount of space or storage required for the bracketed information and thus estimate the marginal savings in storage which might be realized by creating an index to all of this bracketed information and replacing it with numeric keys to this index. Eventually, it is anticipated that the file of distinct brackets will be used to generate a file of the unique words and phrases comprising the bracketed information in the Thesaurus. That is, as noted earlier, much of the bracketed information in the Thesaurus consists of various combinations of words and phrases. As a result, it should be possible to factor, or break up, the bracketed information into a much smaller set of distinct words and phrases used in constructing these "entries". In fact, it is estimated that from the 1,561 distinct brackets found in the Thesaurus, there may exist at most 500 distinct words and phrases comprising these brackets. Eventually, an index to these distinct words and phrases will be constructed and each occurrence of a word or phrase from this index will be replaced in the Thesaurus by a unique identifying code associated with it in the index.

Before this index can be constructed, however, it is necessary to do a considerable amount of editing work on the bracketed information in the Thesaurus. In addition to the inconsistencies noted earlier involving the use of abbreviations, there are a number of inconsistencies in the use of punctuation. As a result, in order to proceed more quickly with the parsing of the Thesaurus, it has been necessary to delay the construction of such an index for the time being and find some other means of dealing with the bracketed information. Since a file of the unique brackets currently exists, it has been decided to simply replace each occurrence of bracketed information in the Thesaurus with the corresponding numeric code already assigned for each of the 1,561 brackets in the presently existing index. When the file was created, each bracket was assigned a number corresponding to its placement in the file, and thus, the numeric codes range in value from 1 to 1,561. Each occurrence of bracketed information in the Thesaurus has a corresponding "entry" in this file or index, and during the parsing of the Thesaurus each occurrence of bracketed information will be replaced by the proper four digit code, thus enabling easy reference to the index. At a future date, this index can be further refined in the manner proposed above. Upon completion of the "distribution" of bracketed information described earlier in this report, there will probably be somewhere around 20,000 occurrences of bracketed information in the Thesaurus. If each is replaced with a four (4) digit code, then a total of 80,000 characters will be involved. This in itself should result in an estimated savings in storage required for bracketed information of at least 60,000 characters. Note, this estimate is based on the total number of characters computed for the original file of 17,653 brackets and may very well be a conservative estimate.

C. *Etc. in Roget's International Thesaurus*

by Sally Yeates Sedelow

Like the *or*, and the ~ used as an or, *etc.* functions as a shorthand device in *Roget's International Thesaurus*. Because it is the most complicated of all the shorthand devices we have encountered in the *Thesaurus*, it is fortunate that it occurs just 203 times. Here are some examples of its use:

4 Pan-American, Pan-Pacific, Panhellenic, etc.

12 (multiply by five, etc.)

22 hourly, daily, etc.; every hour, every day, etc.; hour by hour, day by day, etc.

29 (contents of a container) cup, cupful, etc.

33 art center, medical center, shopping center, shipping center, railroad center, etc.

44 fishing fleet, whaling fleet, etc.

51 threefold, fourfold, etc.

55 diabetic diet, allergy diet, etc.

66 alarm fire, two-alarm ~, three-alarm, etc., fire

72 temperature, pressure, flow, liquid level, humidity, weight, color, etc.

74 light as air, ~ a feather, etc.

88 navy-blue, baby-blue, etc.

101 hayfield, corn field, wheat field, etc.

117 fast, slow, etc. passages

121 square inch, foot, etc.

131 split, blow, blow the gaff (naut.), sell out, rat, stool, fink, nark, put the finger on, snitch on, squeal on, etc.

160 master craftsman, master workman, master carpenter, etc.

163 marksmanship, seamanship, airmanship, horsemanship, etc.

186 peso (Argentina, Mexico, etc.)

An examination of the entries above indicates that *etc.* is used to indicate the possible extension of a list characterized by the entries preceding the *etc.* One approach to examining the uses of *etc.*, with a view to determining how to deal with them both in the editing and processing of the Thesaurus, is to see whether the list implied by the etc. falls into any special groupings.

First of all, it could be argued that all the lists are to some degree generic--that is, that at least one common thread of relationship runs through all the items on a given list. For example, in entry 33 (above) such diverse activities as those implied by the words art, medical, shopping, shipping, and railroad are all held together by the fact that in each case a center for one of these activities is being described. There are some cases--as, for example, in marksmanship, seamanship, airmanship, horsemanship, etc.--in which the linking thread has been used in such contexts so long and often as to convey little meaning. Nonetheless, in the example, "-manship" clearly indicates that the activities listed are those performable by a human being. Thus, as we consider how to deal with the etc. shorthand, the recognition that the lists are generic implies a need to identify and define the common thread or threads linking the items on the list and those implied by the etc.

Some of the lists fall into larger categories based upon structural or functional characteristics. For example, a number of the lists might be described as having common 'semantic affixes.' Examples of lists for which each has a common 'semantic suffix' are the following:

33 art center, medical center, shopping center, shipping center, railroad center, etc.
88 navy-blue, baby-blue, etc.
101 hayfield, corn field, wheat field, etc.

Examples of lists for which each has a common 'semantic prefix' are:

4 Pan-American, Pan-Pacific, Panhellenic, etc.
160 master craftsman, master workman, master carpenter, etc.

About eighty of the 203 uses of etc. are associated with what I have labeled 'semantic suffixes' and about thirty with 'semantic prefixes.' Awareness of the existence of both these types of lists is important because recognition of other strings which might possibly be included in such lists will be, at least in part, dependent upon straightforward string matching. Other characteristics of the English language also become apparent when looking at the distribution of lists having either common 'prefixes' or 'suffixes.' One aspect of this distribution is described below in relation to lists referring to a form of quantity.

Another type of list which might be thought of as structural or functional comprises those lists used as idioms or slang. Examples are:

31 bean-pole, etc. (tall, thin person)
74 light as air, ~ a feather, etc.
131 split, blow, blow the gaff (naut.), sell out, rat, stool, fink, nark, put the finger on, snitch on, squeal on, etc.

Items in the first list would ordinarily be described as metaphors, those in the second list as similes, and those in the third list would most probably function as metaphors. Also, since the philosophy of this research project tends toward defining idioms in terms of multiple word strings occurring in excess of a certain frequency threshold, there are other lists which might be included in this category.

Examples are:

40 rig for diving, etc.
41 stand by, stand by to weigh anchor, stand by the main sheet, etc.
203 administer a sacrament, administer the eucharist, etc.

Recognition that there is a structural category comprising slang and idioms is important because this category poses special problems for both editing and processing. About 33 of the etc.'s are associated with lists of this type.

In addition to the structural or functional categories identified above, a number of the lists have at least one semantic 'plane' in common. For example, seventeen lists provide groupings based on color. Examples are:

81 rose, flesh, etc.
84 lemon, saffron, etc.

Approximately forty lists are concerned in some way with measurement. Of these forty, 29 contain an integer in some form. Examples are:

12 (multiply by five, etc.)
51 threefold, fourfold, etc.
66 two-alarm ~, three-alarm, etc., fire

Six others describe quantity but do not use any form of integer. Examples are:

29 (contents of a container) cup, cupful, etc.
121 square inch, foot, etc.

It is interesting to note that of the 29 lists containing some form of integer, 19 have, in each case, a common 'semantic suffix' and four have a common 'semantic prefix.' In contrast, of the six lists relating to quantity, at least five, and possibly all six, have a common 'semantic prefix.' The questionable list in this category is number 29 (see above) for which the examples both begin with *cup*, but the additional items implied by the *etc.* presumably might not.

Another group of lists which should be mentioned comprises those naming a form of money and then a list of the countries in which that form of money is used. Examples are:

183 cent (China, Netherlands, etc.); centavo (Portugal, Argentina, Mexico, etc.)
186 peso (Argentina, Mexico, etc.)

The survey above of types of lists involving the use of *etc.* is directed toward identifying and defining the types of extensions (and possible limits) which might replace the *etc.* In general, three types of approaches to determining whether a given word or word string is implied by the *etc.* seem to be suggested. These are:

I. A structural test, involving orthography or syntax, or both;

II. A test to see whether the word or word string falls within some conceptual/semantic plane. In some cases, the *etc.* implies a logical extension of that plane and in other cases, especially in some idiomatic and slang expressions, the extension seems in some respects illogical;

III. A test to see whether a given word or word string falls into a semantic plane which is functionally related to a word or word string (often in parentheses) which heads the list. An example of such a list in relationship to a word in parentheses would be a list of countries in which a particular form of money is used.

For many of the lists which fall under category I, it should be possible to use the etc. as a switch to a computer program which can apply either a syntactic and orthographic or simply an orthographic test to the word or word string in question. For example, in the following list--complex, inferiority complex, superiority complex, Oedipus or nuclear complex, Electra complex, persecution complex; castration complex, etc.--it would seem as if the etc. could be replaced by a test for an orthographic match with complex and a syntactic test for a noun, gerund, or adjective preceding complex. For a list with single word entries such as the following--bacterin, tetanobacterin, typhobacterin, etc.--it would seem possible to test for the orthographic match of bacterin without bothering to deal with syntax. Since almost all the tests involving syntax will entail recognizing an adjective, noun, or gerund followed by a given 'semantic suffix' the syntactic parsing entailed by the test can be a relatively simple recognition procedure. However, the necessity of treating a text in terms of multiple word entries as well as in single entries greatly complicates processing and, although we are very interested in problems relating to idioms and multiple-word items in general, we do not plan to deal with such multiple word strings, insofar as they occur in very long texts, in the immediate future. We might well, on the other hand, experiment with texts involving a relatively small number of paragraphs, pages, or lines.

Many of the lists which will entail a syntactic and orthographic test for possible candidates implied by the _etc_. will also fall into category II, which comprises those lists falling within some conceptual or semantic plane. For example, the following list -- examination, physical examination, digital examination, oral examination, etc.-- contains the word examination preceded by a noun, adjective, or gerund, with the additional constraint that all the examinations are related to the practice of medicine and possibly dentistry. In some of these cases, it may be possible to provide additions to the list which can either be added directly to the list in its thesaural form, or stored separately and addressed by the _etc_. in a given list. In some other cases, the _Thesaurus_ itself might be consulted for possible extensions. For example, in the following list--master craftsman, master workman, master carpenter, etc.--it would seem that a list of occupations, or practitioners of occupations, in addition to the syntactic and orthographic tests, would provide the necessary information to determine whether a given word string is indeed appropriately implied by the _etc_. The _Thesaurus_, itself, already contains a great many very extensive lists which might be used in cases analogous to that cited above.

For list extensions involving integers--such as fivefold, sixfold, etc.--it is conceivable that a computer program might be used which would add appropriate increments up to some predetermined limit.

All of the examples considered under category II thus far might be considered logical extensions of the list as it already exists in the

Thesaurus. The case of an extension of a list involving slang or certain types of idioms, such as metaphors, which might be thought to entail illogical extensions is extremely complicated. For example, what might be the extensions of the following list--split, blow, blow the gaff (naut.), sell out, rat, stool, fink, nark, put the finger on, snitch on, squeal on, etc.? In this case, the _etc_. might well function as a dead end pending further study of possible commonalities among the words on this list. Among other characteristics, metaphors contain words for which the semantic features are mismatched.* For example, the word "cool" and the word "million" in the metaphor,"cool million," would have features which would ordinarily block the co-occurrence of these two words. The question, for our purposes, is under what circumstances such mismatches are permitted. Are the circumstances so varied and extensive that the _etc_. implies almost all otherwise forbidden combinations of words in the English language? If so, for all practical purposes one would again have a dead end situation. It may be that measures of semantic distance, which we hope to be able to define on the _Thesaurus_ or some version of the _Thesaurus_, will be helpful in such cases. That is, there may be a given distance which is of such size as to qualify for an illogical rather than logical extension but not of such size as to include all the language. Much more work on the thesaurus, itself, is needed before any such speculation can be scientifically explored.

* Wayne Leman, a graduate student at the University of Kansas, has provided an extended discussion of this characteristic of metaphor in a paper written for my course in Computational Semantics.

The final category, III, exemplified by the name of a type of money, followed by the countries in which it is used, would seem to entail simple additions to the list providing the additional necessary factual information. Such additions are probably not to be found in the *Thesaurus* and would entail consultation with some other reference work or source.

For the immediate future, our procedure will be to look at each use of *etc.*, expanding those for which the extensions are obvious and brief, and making the others temporary dead ends until either a procedure for identifying proper replacements can be provided or the replacements, themselves, can be given. We would welcome thoughts and suggestions from other scholars and scientists concerning any of the issues we have raised concerning the use of *etc.* as a shorthand device in the *Thesaurus*.

D. Abstract Thesauri and Graph Theory Applications to Thesaurus Research

by Robert Bryan

I. Introduction

In a number of fields of language research, notably in the areas of machine translation, document retrieval, and stylistic and content analysis, thesauri and thesaurus-like structures play an important and growing role. While thesauri such as Roget's International Thesaurus (R.I.T.) have long been available to persons interested in language or literature, their application to automated language tasks requires that a description in more explicit terms than the traditional ones be available and that techniques and procedures be developed which permit not only access to the information contained in a thesaurus for a variety of tasks, but also the modification and comparison of existing thesauri and the construction of new ones. The basis for such a description should be an abstract system or model which correctly reflects the essential elements in the structure of thesauri and in the framework of which traditional and new concepts having to do with thesauri can be talked about in an explicit and unambiguous way, suitable to computer implementation. It is the purpose of this paper to introduce and present a partial development of this abstract system.

The formulation of such a model and its use to gain insight into the structure of thesauri and to aid in discovering concepts and procedures which bear on their application require nothing more than the willingness to look at the entries in the thesaurus as abstract entities about which we know only what an observer with no knowledge of English would know. Such an observer would know of the entry "impulse", for example, in 282.1 on page 166 of R.I.T. that it co-occurs with the entries "impulsion" and "impelling force" in the first semicolon group of 282.1, the set of entries with which it co-occurs in larger groupings, and, given any other entry in the thesaurus, that the entry "impulse" in 282.1 is or is not identical to that entry. (He would also know that "impulse" is in bold face, that it is not followed by a cross-reference number or brackets, and certain other information. The complete list of information carrying devices in R.I.T. is short and formal counterparts of all of those devices can easily be incorporated in the abstract system.) He would not approach the thesaurus with all of the information, impressions, and biases about the words he encountered that an English speaking user derives from his familiarity with the language. A model thus provides us with a computer's eye view of the thesaurus, for it is in precisely this way that the computer sees the entries in the thesaurus. Although it might be hard to find a linguist who will admit that he doesn't know what "hypersphere" or "transduction" means in reference to thesauri, the computer must be communicated with in language precisely defined in terms of that information accessible to it. It is that language which is the language of our model.

Because this is not strictly a mathematics paper, the proofs of the results stated in the following sections have been omitted to save space. Most are fairly straightforward and can be supplied by the reader with some background in mathematics.

II. Definition of an Abstract Thesaurus

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 $c \in C$.

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

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* and every molecule is, therefore, a word or a category or both. 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 is at most one entry. In the following, E, W, C and M always refer to the sets E, W, C and M of a thesaurus.

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

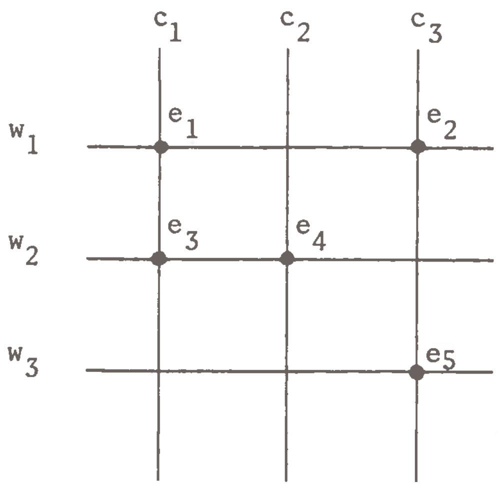
T-graph of the thesaurus $T = \langle E,W,C \rangle$ where

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

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

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. This graph of a thesaurus is given a name to differentiate it from other pictorial representations of thesauri which will be introduced later. A T-graph may be more helpful in visualizing a given thesaurus if only those line segments connecting two dots are drawn.

The following two results follow immediately from the definition of thesaurus.

<u>Theorem 2-1</u>: Distinct entries e_i and e_j cannot be both in the same word and the same category.

Theorem 2-2: If $c \in C$ and $w \in W$ and $c = w$, $|w| = |c| = 1$.

Definition 2-2: If $e \in E$, $w \in W$, and $c \in C$ and $w = c = \{e\}$, e, w and c are said to be isolated.

III. Interpretation

Roget's International Thesaurus is an instantiation of an abstract thesaurus under at least one, and probably several, interpretations 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.; "internal commas" such as those in the expression "no if's, and's, or but's" which occur within character strings obviously meant to be single entries will be made recognizable in the computer accessible version 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, being" in the first line of R.I.T.). "Comma delimited" and "bounded by semicolons" in the previous sentence should be interpreted loosely enough to handle special cases, such as the first or last entry in a grouping larger than "semicolon group", where the actual delimeter may be something other than a comma or a semicolon.

'Word' in this interpretation clearly has a special meaning distinct from its usual one as an entry in R.I.T. may contain several words in the usual sense and a 'word' in the technical sense is a set of identical entries. That (i) through (iv) in the definition of a thesaurus are true of R.I.T. under the above interpretation is immediate; (v) is true since no semicolon group in R.I.T. contains two identical entries. It is probable that there are other valid interpretations of the sets E, W and C, in particular 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.

IV. The Duality Principle

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. W and C are indistinguishable in terms of the properties they possess. It is, therefore, the case that in the development of the abstract system there is, corresponding to each true statement containing a reference to one of W or C or to elements of that set, a true statement, called its dual, formed from the first by replacing each such reference by a reference to the other of W or C. For example, since it is true of abstract thesauri that "Given any $w \in W$, w can occur at most $|w| - 1$ times in the m-chain induced by a type-6 ε-chain", the statement "Given any $c \in C$, c can occur at most $|c| - 1$ times in the m-chain induced by a type-6 ε-chain" is also a true statement in the development of the abstract system. In the development presented on the following pages, duals are sometimes given as separate statements and sometimes collapsed by means of the set $M = W \cup C$.

Although W and C are indistinguishable in the above sense in the abstract system, the "real-word counterpart" of a definition or statement in the abstract system, under any interpretation, may be quite different from that of its dual. The meanings given to "word" and "category" under the principal interpretation, for example, are very different, not least because the counterparts of elements of categories must be semantically close in some sense (under the assumption that the authors of R.I.T. grouped words into categories which bear some semantic relationship to one another that need not obtain between words in different categories--an assumption which must certainly underlie any application of R.I.T.) whereas no such claim can be made about the counterparts of elements of words.

V. Some Notation

As far as possible, standard notation is used for the mathematical concepts used in the following sections. In this section, some definitions are given which introduce the notation and terminology used in later sections for concepts concerning which there is no real agreement on notation.

Definition 5-1: Given any finite set S, $|S|$ will denote the number of elements in S. $|S| = 0$ if $|S| = \emptyset$.

Definition 5-2: "A is a subset of B" is written $A \subseteq B$ and "A is a proper subset of B" is written $A \subset B$.

Definition 5-3: Given a real number r, /r/ denotes the greatest integer less than or equal to r.

Definition 5-4: A sequence $\langle x \rangle = x_1, x_2, \ldots$ whose terms are elements of a set X is called a sequence, or a chain, over X. If $\langle x \rangle$ is finite, the initial and final terms of $\langle x \rangle$ are sometimes denoted x_I and x_F respectively. We say $\langle x \rangle$ connects x_I and x_F. Other terms of $\langle x \rangle$ are interior terms. An ordered pair (x_i, x_{i+1}), $1 \leq i \leq n-1$, is a link of $\langle x \rangle$ and a doubleton $\{x_i, x_{i+1}\}$, $1 \leq i \leq n-1$, is a block of $\langle x \rangle$. A link or block is symmetrical if $x_i = x_{i+1}$. A link or block of a chain over M is w(ord)-uniform if x_i and x_{i+1} are both elements of W, c(ategory)-uniform if x_i and x_{i+1} are both elements of C and uniform if it is either w-uniform or c-uniform. A link or a block of a chain over E is uniform if both of its components are contained in some $m \in M$. (x_1, x_2) is the initial link of $\langle x \rangle$ and (x_{n-1}, x_n) is the final link of $\langle x \rangle$. Other links of $\langle x \rangle$ are interior links of $\langle x \rangle$. Initial, final and interior blocks are defined analogously.

Definition 5-5: If $\langle x \rangle$ is a chain over X, $\{\langle x \rangle\}$ denotes the set $\{x \in X \mid x = x_i \text{ for some term } x_i \text{ of } \langle x \rangle\}$.

Definition 5-6: If $<x> = x_1,x_2,\ldots,x_n$, the *length of* $<x>$, n, is denoted $|<x>|$.

Definition 5-7: The chain $<y> = y_1,y_2,\ldots,y_m$ is a *sub-chain* of $<x> = x_1,x_2,\ldots,x_n$ if there exist positive integers $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ such that $y_j = x_{i_j}$ for all $1 \leq j \leq m$. If, in addition, $i_k - i_{k-1} = 1$ for all $2 \leq k \leq m$, $<y>$ is a *segment* of $<x>$ and an *initial segment* of $<x>$ if $i_1 = 1$. A segment of $<y> = y_1,y_2,\ldots,y_m$ is *symmetrical* if $y_i = y_j$ for all $1 \leq i,j \leq m$. A segment, $<y>$, of a chain, $<x>$, over M is *w-uniform* if all of the terms of $<y>$ are words, *c-uniform* if all of the terms of $<y>$ are categories, and *uniform* if it is either w-uniform or c-uniform. A symmetrical segment of a chain over M is clearly uniform.

Definition 5-8: If R is an equivalence relation on a set X and $x \in X$, the set $\{y \in X | \ yRx\}$ is called the *R-equivalence class determined by x* and is denoted $[x]_R$ or $[x]$ where it is clear from the context what relation is meant.

Definition 5-9: Given a set S, a collection, P, of pair-wise disjoint subsets of S whose union is S is a *partition* of S. If P_1 and P_2 are partitions of S, P_1 is a *refinement* of P_2 if given any $S_1 \in P_1$ there exists $S_2 \in P_2$ such that $S_1 \subseteq S_2$.

Definition 5-10: If f is a function from A into B, we write $f: A \rightarrow B$.

Definition 5-11: Given a set S, the set $P(S) = \{S_1 | \; S_1 \subseteq S\}$ of all subsets of S is called the power set of S.

Definition 5-12: The number $\frac{n!}{r!\,(n-r)!}$ where n and r are positive integers with $r \leq n$ will be abbreviated $\binom{n}{r}$. $\binom{n}{r} = 0$ if $n < r$ or $n = 0$.

VI. The σ- and r-operators

Definition 6-1: Let $M_1 = \{m_1, m_2, \ldots, m_n\} \subseteq M$. Then $\sigma(M_1) = m_1 \cup m_2 \cup \ldots \cup m_n$ is called the σ-set of M_1.

Theorem 6-1: Given $M_1 = \{m_1, m_2, \ldots, m_n\} \subseteq M$, $|\sigma(M_1)| \leq \sum_{i=1}^{n} |m_i|$ and $|\sigma(M_1)| = \sum_{i=1}^{n} |m_i|$ if and only if the m_i are pair-wise disjoint.

Corollary: Given $C_1 = \{c_1, c_2, \ldots, c_n\} \subseteq C$ and $W_1 = \{w_1, w_2, \ldots, w_n\} \subseteq W$,

$$|\sigma(C_1)| = \sum_{i=1}^{n} |c_i| \quad \text{and} \quad |\sigma(W_1)| = \sum_{i=1}^{n} |w_i|.$$

Definition 6-2: Let $E_1 = \{e_1, e_2, \ldots, e_n\}$ be a subset of E. Then the m(olecule)-range of E_1, denoted $r_m(E_1)$, is the set $\{m \in M | \ e_i \in m$ for some $e_i \in E_1\}$. Note that $r_m(E_1)$ is comprised of the distinct elements of the set $\{w_1, c_1, w_2, c_2, \ldots, w_n, c_n\}$ where w_i is the unique element of W and c_i the unique element of C such that $e_i \in w_i \cap c_i$. The c(ategory)-range of E_1, denoted $r_c(E_1)$, is $r_m(E_1) \cap C$ and the w(ord)-range of E_1, denoted $r_w(E_1)$, is $r_m(E_1) \cap W$.

Definition 6-3: Given $w \in W$, $c \in C$, the range of w, r(w), is $r_c(w)$ and the range of c, r(c), is $r_w(c)$.

Definition 6-4: Given $c \in C$, $w \in W$, $|r(w)|$ and $|r(c)|$ are the range indices of w and c respectively.

Theorem 6-2: Given $c \in C$, $w \in W$, $|r(c)| = |c|$ and $|r(w)| = |w|$.

The notion of range in the preceding definition can be extended to n-tuples of words and categories.

Definition 6-5: If $C_1 = \{c_1, c_2, \ldots, c_n\} \subseteq C$, $r'(C_1) = r(c_1) \cup r(c_2) \cup \ldots \cup r(c_n)$ is called the range of C_1 and if $W_1 = \{w_1, w_2, \ldots, w_n\} \subseteq W$, $r'(W_1) = r(w_1) \cup r(w_2) \cup \ldots \cup r(w_n)$ is called the range of W_1.

The σ-operator, σ, assigns to a set of molecules a set of entries and the r-operators r_m, r_c, r_w, and r assign to a set of entries a set of molecules. The range of a word, w, is the set of categories whose intersection with w is non-null and the range of a category, c, is the set of words whose intersection with c is non-null. Since

$$\sigma : P(M) \rightarrow P(E)$$

and

$$\begin{aligned} r_m &: P(E) \rightarrow P(M) \\ r_c &: P(E) \rightarrow P(C) \subset P(M) \\ r_w &: P(E) \rightarrow P(W) \subset P(M) \\ r &: C \subset P(E) \rightarrow P(W) \subset P(M) \\ r &: W \subset P(E) \rightarrow P(C) \subset p(M) \end{aligned}$$

(the same symbol is used for the last two functions since which function is meant can be gathered from the argument) the σ- and r-operators can be alternately imbedded with interesting results.

For the next three theorems, let E_1, $E_2 \subseteq E$, $C_1 \subseteq C$, $W_1 \subseteq W$, M_1, $M_2 \subseteq M$, $c \in C$, $w \in W$, and $m \in M$.

<u>Theorem 6-3</u>: i) $\sigma(\{m\}) = m$

ii) $\sigma(M_1 \cup M_2) = \sigma(M_1) \cup \sigma(M_2)$

iii) $\sigma(M_1 \cap M_2) \subseteq \sigma(M_1) \cap \sigma(M_2)$

iv) $r_m(E_1 \cup E_2) = r_m(E_1) \cup r_m(E_2)$

v) $r_m(E_1 \cap E_2) \subseteq r_m(E_1) \cap r_m(E_2)$

vi) $r_c(E_1 \cup E_2) = r_c(E_1) \cup r_c(E_2)$

vii) $r_c(E_1 \cap E_2) \subseteq r_c(E_1) \cap r_c(E_2)$

viii) $r_w(E_1 \cup E_2) = r_w(E_1) \cup r_w(E_2)$

ix) $r_w(E_1 \cap E_2) \subseteq r_w(E_1) \cap r_w(E_2)$

x) $r_m(E_1) = r_c(E_1) \cup r_w(E_1)$

Theorem 6-4: i) $r_m(\sigma(M_1)) = M_1 \cup r'(M_1 \cap C) \cup r'(M_1 \cap W)$

ii) $r_c(\sigma(M_1)) = (M_1 \cap C) \cup r'(M_1 \cap W)$

iii) $r_w(\sigma(M_1)) = (M_1 \cap W) \cup r'(M_1 \cap C)$

iv) $r_m(\sigma(C_1)) = C_1 \cup r'(C_1)$

v) $r_m(\sigma(W_1)) = W_1 \cup r'(W_1)$

vi) $r_w(\sigma(C_1)) = r'(C_1)$

vii) $r_c(\sigma(W_1)) = r'(W_1)$

viii) $r_c(\sigma(C_1)) = C_1$

ix) $r_w(\sigma(W_1)) = W_1$

x) $r_m(\sigma(w)) = \{w\} \cup r(w)$

xi) $r_m(\sigma(c)) = \{c\} \cup r(c)$

xii) $r_c(\sigma(c)) = \{c\}$

xiii) $r_w(\sigma(w)) = \{w\}$

xiv) $r_c(\sigma(w)) = r(w)$

xv) $r_w(\sigma(c)) = r(c)$

Theorem 6-5: i) $C_1 \subseteq r_m(\sigma(C_1))$

ii) $W_1 \subseteq r_m(\sigma(W_1))$

iii) $M_1 \subseteq r_m(\sigma(M_1))$

iv) $M_1 \cap C \subseteq r_c(\sigma(M_1))$

v) $M_1 \cap W \subseteq r_w(\sigma(M_1))$

VII. Measures on C and W

The following definitions present some distance measures on the sets C and W.

Definition 7-1: Given $c_1, c_2 \in C$, $o(c_1,c_2) = r(c_1) \cap r(c_2)$ is called the overlap of the pair (c_1,c_2). $|o(c_1,c_2)|$ is called the degree of overlap of c_1 and c_2 and is denoted $d(c_1,c_2)$.

There are several ways to normalize $d(c_1,c_2)$. Notation for some of these is introduced in the next definition.

Definition 7-2: i) $d^*(c_1,c_2) = \dfrac{d(c_1,c_2)}{|c_1 \cup c_2|}$

ii) $d'(c_1,c_2) = \frac{1}{2}\left[\dfrac{d(c_1,c_2)}{|c_1|} + \dfrac{d(c_1,c_2)}{|c_2|}\right]$

iii) $d''(c_1,c_2) = \dfrac{d(c_1,c_2)}{\min(|c_1|,|c_2|)}$

iv) $d'''(c_1,c_2) = \dfrac{d(c_1,c_2)}{|r(c_1) \cup r(c_2)|}$

Overlap, degree of overlap, and the measures d^*, d', d'' and d''' are defined analogously on W.

Theorem 7-1: i) Given $c_1, c_2 \in C$, $|r(c_1) \cup r(c_2)| \leq |c_1 \cup c_2|$ and $|r(c_1) \cup r(c_2)| = |c_1 \cup c_2|$ if and only if $d(c_1,c_2) = 0$.

ii) Given $w_1, w_2 \in W$, $|r(w_1) \cup r(w_2)| \leq |w_1 \cup w_2|$ and $|r(w_1) \cup r(w_2)| = |w_1 \cup w_2|$ if and only if $d(w_1,w_2) = 0$.

Theorem 7-2: Given $w_1, w_2 \in W$, $c \in C$,

$$\begin{aligned} |(w_1 \cup w_2) \cap c| &= 2 \text{ if } w_1 \cap c \neq \emptyset \text{ and } w_2 \cap c \neq \emptyset, \\ &= 1 \text{ if } w_1 \cap c \neq \emptyset \text{ and } w_2 \cap c = \emptyset \\ &\quad \text{or } w_1 \cap c = \emptyset \text{ and } w_2 \cap c \neq \emptyset, \\ \text{and } &= 0 \text{ if } w_1 \cap c = \emptyset \text{ and } w_2 \cap c = \emptyset \end{aligned}$$

and analogously for $c_1, c_2 \in C$, $w \in W$.

VIII. Chains

The term connectivity in reference to the study of thesauri refers to concepts and relations definable in terms of chains over the sets E, W, and C having certain properties. Possible types of chains and their properties are best investigated by starting with the most general chains of the smallest elements in the system, arbitrary chains of entries, and building other chains over E and over W and C by placing increasingly stringent restrictions on arbitrary chains. The ten types of chains introduced in the following defintions are so defined that a type n chain is also a type m chain if $m \leq n$.

Definition 8-1: Chains over E, M, C, and W are called ε-, m-, c-, and w-chains respectively.

Definition 8-2: Arbitrary chains over E are called type 1 ε-chains. We denote by E^n the set of all type n ε-chains over E.

Type 1 ε-chains are clearly of no interest. In Definition 8-3, we add the minimum property which characterizes chains which are of interest in the study of connectivity.

Definition 8-3: If $<\varepsilon> = e_1, e_2, \ldots$ is an ε-chain in E^1, $<\varepsilon>$ is a type 2 ε-chain if for all $i = 1,2,\ldots,$ there exists $m \in M$ such that $e_i \in m$ and $e_{i+1} \in m$, i.e., if each link of $<\varepsilon>$ is uniform.

A type 2 ε-chain, $<\varepsilon>$, can be traced on a T-graph by staying on lines representing words or categories and turning only at dots. The geometrical configuration so formed is called the trace of $<\varepsilon>$. In type 3 ε-chains adjacent terms are required to be non-identical.

Definition 8-4: $<\varepsilon> \in E^2$ is a type 3 ε-chain if no link of $<\varepsilon>$ is symmetrical.

Type 3 ε-chains are not yet sufficiently restricted that they cannot be infinite sequences. An ε-chain of the form $e_i, e_j, e_i, e_j, \ldots,$ for example, is a type 3 ε-chain if $\{e_i, e_j\}$ is uniform. The restriction imposed in Definition 8-5 will require type 4 ε-chains to be finite in length.

Definition 8-5: $\langle\varepsilon\rangle \in E^3$ is a type 4 ε-chain if no two links of $\langle\varepsilon\rangle$ are identical.

If we use "block" in reference to T-graphs to mean a line segment joining two dots, we may say that in the trace of a type 4 ε-chain a block may not be retraced in the same direction.

Definition 8-6: $\langle\varepsilon\rangle \in E^3$ is a type 5 ε-chain if no two blocks of $\langle\varepsilon\rangle$ are identical.

Clearly, a type 5 ε-chain is a type 4 ε-chain.

Definition 8-7: $\langle\varepsilon\rangle \in E^2$ is a type 6 ε-chain if its terms are pairwise distinct, i.e., if $e_i, e_j \in \{\langle\varepsilon\rangle\}$, $i \neq j \Rightarrow e_i \neq e_j$. A type 6 ε-chain is also called a non-repeating ε-chain.

It is immediate from the definitions of the ε-chains thus far considered that a type n ε-chain is also a type m ε-chain if $m \leq n$.

Since given arbitrary entries $e_i, e_j \in E$, there is at most one molecule $m \in M$ such that $e_i \in m$ and $e_j \in m$, there is a unique chain of molecules associated in a natural way with each element of E^3. Further restrictions are placed on ε-chains by restricting their induced m-chains.

Theorem 8-1: If $<\varepsilon> \in E^3$ and $<\varepsilon> = e_1, e_2, \ldots, e_n$, there exists a unique sequence $m_1, m_2, \ldots, m_{n-1}$ of elements of M having the property that for all $1 \leq i \leq n-1$, $e_i, e_{i+1} \in m_i$.

Definition 8-8: Given $<\varepsilon> \in E^3$, the unique m-chain whose existence is asserted in the previous theorem is called the m-chain induced by $<\varepsilon>$ and is denoted $<m>_\varepsilon$. An m-chain induced by some $<\varepsilon> \in E^3$ is an induced m-chain. The subchain of $<m>_\varepsilon$ formed from $<m>_\varepsilon$ by deleting all m_i which are not words is the w-chain induced by $<\varepsilon>$ and is denoted $<w>_\varepsilon$, that formed from $<m>_\varepsilon$ by deleting all m_i which are not categories is the c-chain induced by $<\varepsilon>$ and is denoted $<c>_\varepsilon$. $<m>_\varepsilon$,$<w>_\varepsilon$, and $<c>_\varepsilon$ are said to allow $<\varepsilon>$. We also say that $<m>_\varepsilon$ induces $<c>_\varepsilon$ and $<w>_\varepsilon$ and that $<c>_\varepsilon$ and $<w>_\varepsilon$ allow $<m>_\varepsilon$.

It is clear that an ε-chain in E induces a unique (possibly null) w-chain and a unique (possibly null) c-chain. However, a w-chain or c-chain may allow more than one induced m-chain and an induced m-chain may allow more than one ε-chain.

Theorem 8-2: Every uniform link of an induced m-chain is symmetrical.

This theorem says that adjacent terms of an induced m-chain cannot be distinct words or distinct categories.

Definition 8-8: Given an m-chain $<m>$ and a molecule $m_0 \in M$, the number of times m_0 occurs as a term of $<m>$ is called the frequency of m_0 in $<m>$ and is denoted $f_{<m>}(m_0)$ or $f(m_0)$ if $<m>$ is understood. $q_{<m>}(m_0)$ or $q(m_0)$ denotes the number of links, (m_i, m_{i+1}), of $<m>$ having the property that $m_i = m_{i+1} = m_0$ and $h_{<m>}(m_0)$ or $h(m_0)$ has value 2, 1, or 0 depending on whether both, exactly one, or neither of m_I and m_F are equal to m_0.

Theorem 8-3: Suppose $<m>$ is the m-chain induced by a type 6 ε-chain. Then the number of distinct type 6 ε-chains allowed by $<m>$ is

$$\prod_{m_0 \in \{<m>\}} \frac{[|m_0| - 2(f(m_0) - g(m_0)) + h(m_0)]!}{[|m_0| - 2 \cdot f(m_0) + g(m_0)]!}$$

where the product is taken over all $m_0 \in \{<m>\}$.

Theorem 8-4: Let $<\varepsilon> \in E^6$ and $m \in M$. Then $f_{<m>_\varepsilon}(m_0) \leq |m_0| - 1$ and $f_{<m>_\varepsilon}(m_0) = |m_0| - 1$ if and only if all of the terms of $<m>_\varepsilon$ equal to m_0 are adjacent.

Theorem 8-5: Let $<\varepsilon> \in E^6$ and $m_0 \in M$. Then if $g_{<m>_\varepsilon}(m_0) = 0$, $f_{<m>_\varepsilon}(m_0) \leq \Big/ \frac{\lfloor m \rfloor}{2} \Big/$.

Theorem 8-5 says that, under the hypotheses of the theorem, the frequency of a given molecule, m_0, in <m> is $\frac{|m_0|}{2}$ if $|m_0|$ is even and $\frac{|m_0| - 1}{2}$ if $|m_0|$ is odd.

Theorem 8-6: Let $\langle\varepsilon\rangle \in E^k$. Then given $m_0 \in M$, $f_{\langle m\rangle_\varepsilon}(m_0) \leq 2 \cdot \binom{|m_0|}{2}$ if $k = 4$ and $f_{\langle m\rangle_\varepsilon}(m_0) \leq \binom{|m_0|}{2}$ if $k = 5$.

Definition 8-9: An m-chain, <m>, is w(ord)-restricted if no two adjacent terms of <m> are words and c(ategory)-restricted if no two adjacent terms of <m> are categories. <m> is alternating if it is both w-restricted and c-restricted, i.e., if no link of <m> is uniform. $\langle\varepsilon\rangle \in E^3$ is alternating if $\langle m\rangle_\varepsilon$ is alternating.

Except for the initial and final terms, an alternating m-chain allows at most one ε-chain. This is stated formally in Theorem 8-7.

Theorem 8-7: Given an alternating m-chain $\langle m\rangle = m_1, m_2, \ldots, m_n$ and entries e_i, e_j such that i) $e_i \in m_1$

ii) $e_j \in m_n$

and iii) for all links (m_i, m_{i+1}) of <m>, $m_i \cap m_{i+1} \neq 0$,

there exists a unique $\langle\varepsilon\rangle \in E^3$ such that <ε> connects e_i and e_j and induces <m>.

We defined an alternating m-chain to be an m-chain none of whose links is uniform. Theorem 8-8 allows us to weaken the condition under which an m-chain is alternating.

Theorem 8-8: An m-chain, <m>, is alternating if no link of <m> is symmetrical.

The next theorem follows directly from Theorem 8-5 and the fact that no two adjacent terms of an alternating m-chain are identical.

Theorem 8-9: If $<\varepsilon> \in E^6$ and $<m>_\varepsilon$ is alternating, $f_{<m>_\varepsilon}(m) \leq /\frac{|m|}{2}/$ for all $m \in M$.

Definition 8-10: An m-chain is *non-w-repeating* if no two of its terms are equal to the same word. A *non-c-repeating* m-chain is defined analogously. An m-chain is *non-repeating* if its terms are pairwise distinct.

Note that a molecule which is both a word and a category cannot be a term of an induced m-chain. We needn't be concerned then about a molecule appearing as a term of an m-chain once as a word and once as a category.

Definition 8-11: A *type 7 ε-chain* is a type 6 ε-chain whose induced m-chain is either non-w-repeating or non-c-repeating.

Since adjacent links of an induced m-chain cannot be distinct words or categories, a non-repeating m-chain is, in fact, alternating.

Theorem 8-10: If <m> is a non-repeating m-chain, <m> is alternating.

Definition 8-12: A *type 8 ε-chain* is a type 6 ε-chain whose induced m-chain is non-repeating. Type 8 ε-chains are also called *proper ε-chains*. Clearly, $<w>_\varepsilon$ and $<c>_\varepsilon$ are also non-repeating.

The requirement that the ε-chain in Definition 8-12 be of type 6 was not redundant for, although no term of an ε-chain, <ε>, whose induced m-chain is non-repeating can be equal to another interior term of <ε>, it may be that the initial and final terms of <ε> are identical. Such an ε-chain is given a special name in the next definition.

Definition 8-13: A type 3 ε-chain whose induced m-chain is non-repeating and whose initial and final terms are identical is called a *ring*.

Theorem 8-11: $E^n \subseteq E^m$ for all $1 \leq n \leq m \leq 8$.

Theorem 8-12: Given a type 8 ε-chain, $<\varepsilon> = e_1, e_2, \ldots, e_n$, and $m \in M$,

$$|m \cap \{e_2, e_3, \ldots, e_{n-1}\}| = \begin{cases} 2 & \text{if } m \in \{<m>_\varepsilon\} \\ 0 & \text{if } m \notin \{<m>_\varepsilon\} \end{cases}$$

and if $m \cap \{e_2, e_3, \ldots, e_{n-1}\} = \{e_i, e_j\}$, $\{e_i, e_j\}$ is a block of <ε>, i.e., one of (e_i, e_j) and (e_j, e_i) is a link of <ε>.

Definition 8-14: A link (e_i, e_{i+1}) of a type 3 ε-chain is a c-link if $e_i, e_{i+1} \in c$ for some $c \in C$ and a w-link if $e_i, e_{i+1} \in w$ for some $w \in W$.

Definition 8-15: Given $<\varepsilon> \in E^3$,

i) a c-link (e_i, e_{i+1}) of $<\varepsilon>$ is strong if $d(w_1, w_2) > 1$, where (w_1, w_2) is the unique pair of distinct words such that $e_i \in w_1$ and $e_{i+1} \in w_2$, and weak otherwise.

ii) a w-link (e_i, e_{i+1}) of $<\varepsilon>$ is strong if $d(c_1, c_2) > 1$, where (c_1, c_2) is the unique pair of distinct categories such that $e_i \in c_1$ and $e_{i+1} \in c_2$, and weak otherwise.

Theorem 8-13: If $e \in E$, $m \in M$, $e \in m$ and m has range index 1, then e cannot be an interior term of an alternating ε-chain.

Definition 8-16: A type 8 ε-chain, $<\varepsilon>$, is w(ord)-strong if each of its c-links is strong and w-weak otherwise, and c(ategory)-strong if each of its w-links is strong and c-weak otherwise. $<\varepsilon>$ is strong if it is both w-strong and c-strong and weak otherwise.

The terminology of Definition 8-16 stems from the fact that $<\varepsilon> \in E^8$ is w-strong if and only if $d(w_i, w_{i+1}) > 1$ for each link of $<w>_\varepsilon$ and c-strong if $d(c_i, c_{i+1}) > 1$ for each link of $<c>_\varepsilon$. Strong links represent a first, but seemingly fairly good, approximation of a formal counterpart to links between words or categories in the thesaurus which are semantically valid in that the link represents a true semantic closeness and not the semantically void relationship between homographs. In terms of strong links, we can formulate a formal device which approximates the relationship between homographs and between entries corresponding to the literal and metaphorical use of a word, although we have no way of choosing one as the literal use. In fact, this information may very well not be contained in R.I.T.

Definition 8-17: Entries $e_i, e_j \in E$ are homographs if and only if

i) (e_i, e_j) is a w-link

and ii) there does not exist a strong type 8 ε-chain connecting e_i and e_j.

This is a stronger condition than the requirement that (e_i, e_j) be strong since a strong ε-chain in E^8 may connect e_i and e_j even if (e_i, e_j) is weak.

We define type 9 and 10 ε-chains in terms of "strength" and extend the embedding of the sets E^n to the new chains.

Definition 8-18: $<\varepsilon> \in E^8$ is a type 9 ε-chain if it is either w-strong or c-strong. We denote by E^9_c the set of c-strong elements of E^8 and by E^9_w the set of w-strong elements of E^8. Of course, $E^9 = E^9_c \cup E^9_w$.

Definition 8-19: $<\varepsilon> \in E^8$ is a type 10 ε-chain if it is strong.

Theorem 8-14: $E^n \subseteq E^m$ for all $1 \le m \le n \le 10$.

IX. Neighborhoods

We define a neighborhood of an entry $e \in E$ to be the set of entries in E which are terms of chains emanating from e of a given type and length.

Definition 9-1: $<\varepsilon> = e_1, e_2, \ldots, e_n$ is said to emanate from e_1 for any $<\varepsilon> \in E^2$.

Definition 9-2: Let $e \in E$ and let n and r be positive integers with $2 \le n \le 10$. Then

i) $S'^n_r(e) = \{<\varepsilon> \in E^n \mid <\varepsilon> \text{ emanates from } e \text{ and } |<\varepsilon>| = r\}$

ii) $S^n_r(e) = \{<\varepsilon> \in E^n \mid <\varepsilon> \text{ emanates from } e \text{ and } |<\varepsilon>| \le r\}$

iii) $S^n_\infty(e) = \{<\varepsilon> \in E^n \mid <\varepsilon> \text{ emanates from } e\}$

iv) $N'^n_r(e) = \bigcup_{<\varepsilon> \in S'^n_r} \{<\varepsilon>\}$

v) $N^n_r(e) = \bigcup_{<\varepsilon> \in S^n_r} \{<\varepsilon>\}$

vi) $N^n_\infty(e) = \bigcup_{<\varepsilon> \in S^n_\infty} \{<\varepsilon>\}$

$S'^n_r(e)$, $S^n_r(e)$, and $S^n_\infty(e)$ are called stars of e and $N'^n_r(e)$, $N^n_r(e)$, and $N^n_\infty(e)$ are neighborhoods of e. A star or neighborhood with superscript n is a *type n star or neighborhood*. A star or neighborhood with subscript r is *bounded* and *of radius r* and with subscript ∞ is *unbounded*.

Since elements of E^4 are finite in length, there is associated with any thesaurus a positive integer r_b having the property that for any $e \in E$, $4 \leq n \leq 10$, and $r \geq r_b$, $S^n_r(e) = S^n_\infty(e)$ and $N^n_r(e) = N^n_\infty(e)$, the smallest such r_b being the length of the largest member of E^4 emanating from e.

Stars are sets of chains emanating from a given entry and neighborhoods are the subsets of E "covered" by stars. More precisely, $S'^n_r(e)$ and $S^n_r(e)$ are the collections of all type n chains emanating from e having length equal to and less than or equal to r respectively and $N'^n_r(e)$ and $N^n_r(e)$ are the sets of entries which appear as terms of chains in $S'^n_r(e)$ and $S^n_r(e)$. If e is isolated, then for any n and r, $S^n_r(e) = \{\langle\varepsilon\rangle\}$, where $\langle\varepsilon\rangle$ is the chain whose single term is e, and $N^n_r(e) = \{e\}$. For any n, r, and e we, of course, have $S^n_r(e) = S'^n_r(e) \cup S'^n_{r-1}(e) \cup \ldots \cup S'^n_1(e)$ and a corresponding statement for neighborhoods. Theorem 9-1 gives some other immediate results of the definitions of stars and neighborhoods.

Theorem 9-1: For all r, s, n, m, and e

i) $S^n_r(e) \subseteq S^n_s(e)$ if $r \leq s$,

$S'^n_r(e) \subseteq S^n_r(e)$,

and $S^n_r(e) \subseteq S^m_r(e)$ if $n \geq m$.

ii) $N^{n}_{r}(e) \subseteq N^{n}_{s}(e)$ if $r \leq s$,

$N'^{n}_{r}(e) \subseteq N^{n}_{r}(e)$,

and $N^{n}_{r}(e) \subseteq N^{m}_{r}(e)$ if $n \geq m$.

That we don't have $N'^{n}_{r}(e) \subseteq N'^{n}_{s}(e)$ for $r \leq s$ can be shown by example. Let $T = \langle E,W,C \rangle$ be the thesaurus whose T-graph is shown in Figure 1, where (i,j) labels the dot denoting the entry in $w_i \cap c_j$.

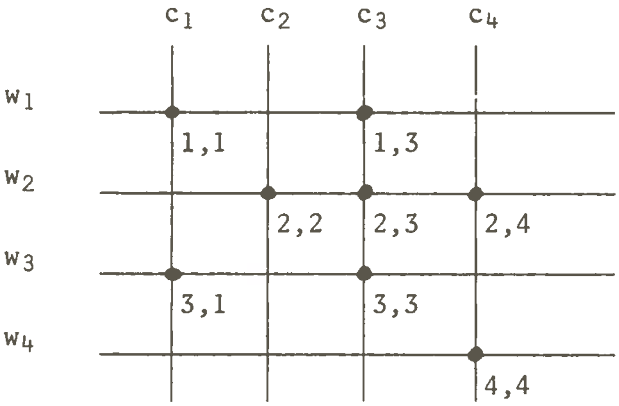


Fig. 1

The chains in $S'^{8}_{3}(e_{1,3})$ are

$\langle \varepsilon_1 \rangle = e_{1,3}, e_{1,1}, e_{3,1}$

$\langle \varepsilon_2 \rangle = e_{1,3}, e_{2,3}, e_{2,2}$

$\langle \varepsilon_3 \rangle = e_{1,3}, e_{2,3}, e_{2,4}$

and $\langle \varepsilon_4 \rangle = e_{1,3}, e_{3,3}, e_{3,1}$

so that $N'^{8}_{3}(e_{1,3}) = \{e_{1,3}, e_{1,1}, e_{3,1}, e_{2,3}, e_{2,2}, e_{4,3}, e_{3,3}\}$. The chains in $S'^{8}_{4}(e_{1,3})$ are

$$\langle\varepsilon_1\rangle = e_{1,3}, e_{1,1}, e_{3,1}, e_{3,3}$$

$$\langle\varepsilon_2\rangle = e_{1,3}, e_{3,3}, e_{3,1}, e_{1,1}$$

$$\text{and} \quad \langle\varepsilon_3\rangle = e_{1,3}, e_{2,3}, e_{2,4}, e_{4,4}$$

and therefore $N'^{8}_{4}(e_{1,3}) = \{e_{1,3}, e_{1,1}, e_{3,1}, e_{2,3}, e_{2,4}\}$. Since $e_{2,2}$ is in $N'^{8}_{3}(e_{1,3})$ and not in $N'^{8}_{4}(e_{1,3})$, $N'^{8}_{3}(e_{1,3}) \not\subseteq N'^{8}_{4}(e_{1,3})$. The same example shows that we needn't have $N^{n}_{r}(e) \subseteq N'^{n}_{r}(e)$ for given n, r and e since $e_{2,2} \in N^{8}_{4}(e_{1,3})$ but $e_{2,2} \notin N'^{8}_{4}(e_{1,3})$. Although $S^{n}_{r}(e)$ for a fixed r and $2 \leq n \leq 10$ are in general 9 distinct sets, $N^{n}_{r}(e)$ for a fixed r and $2 \leq n \leq 10$ are not. In fact, we will show that for a given positive integer, r, and a given entry, e, $N^{n}_{r}(e) = N^{m}_{r}(e)$ for all $2 \leq n$, $m \leq 8$. Since we already have $N^{n}_{r}(e) \subseteq N^{m}_{r}(e)$ for $n \geq m$, we need only show that for $2 \leq m < n \leq 8$, $N^{m}_{r}(e) \subseteq N^{n}_{r}(e)$. We first state two theorems of which the desired result is then an immediate consequence.

Theorem 9-2: Suppose $\langle\varepsilon\rangle \in E^{n}$ and $\langle\varepsilon\rangle = e_1, e_2, \ldots, e_n$. Then for each term e_i of $\langle\varepsilon\rangle$, the initial segment $e_1, e_2, \ldots, e_{i-1}, e_i$ is a type n ε-chain.

It follows from Theorem 9-2 that given e, n and r, $N^{n}_{r}(e)$ is the same as the set of entries connected to e by a type n ε-chain of length less than or equal to r. Therefore, to show that for a given e and r, $N_r^m(e) \subseteq N_r^n(e)$ if $2 \leq m < n \leq 8$, it suffices to show that given any distinct $e_i, e_j \in E$, if there exists a type m ε-chain connecting e_i and e_j, there exists a type n ε-chain connecting e_i and e_j whose length is less than or equal to r where $2 \leq m < n \leq 8$. We show in fact that for all $2 \leq m \leq 7$, if $<\varepsilon> \in E^m$ and $<\varepsilon>$ connects given entries e_i and e_j, $e_i \neq e_j$, $<\varepsilon>$ contains a subchain of type m+1 which connects e_i and e_j. Suppose, for example, that $e_i, e_j \in E$, $e_i \neq e_j$, $<\varepsilon>$ connects e_i and e_j and $<\varepsilon> \in E^2$. Let $<\varepsilon> = e_1, e_2, \ldots, e_n$. If $<\varepsilon>$ has no symmetrical links, it is also of type 3 and the desired subchain of $<\varepsilon>$ is $<\varepsilon>$. Suppose $<\varepsilon>$ has one or more symmetrical links. Let k be the least subscript such that (e_k, e_{k+1}) is symmetrical and construct the sequence $e_1, e_2, \ldots, e_k, e_{k+2}, \ldots e_n$. Clearly, through a finite number of such steps the desired subchain of $<\varepsilon>$ can be constructed from $<\varepsilon>$. Constructing in an analogous way, subchains of type m+1 for each $3 \leq m \leq 7$ completes the proof of the next theorem.

Theorem 9-3: Given any $e_i, e_j \in E$ with $e_i \neq e_j$, if $<\varepsilon> \in E^m$ where $2 \leq m \leq 7$ and $<\varepsilon>$ connects e_i and e_j, then $<\varepsilon>$ contains a subchain of type m+1 connecting e_i and e_j.

The next theorem is an immediate consequence of Theorem 9-3.

Theorem 9-4: Give $e \in E$ and a positive integer r, $N_r^n(e) = N_r^m(e)$ and $N_\infty^n(e) = N_\infty^m(e)$ for all $2 \leq n,m \leq 8$.

In view of Theorem 9-4, we can drop the superscript on neighborhoods of type 2 through 8.

Definition 9-3: Given $e \in E$ and a positive integer r, $N_r(e)$ denotes $N_r^8(e)$ and is called the disk neighborhood of e of radius r and $N_\infty(e)$ denotes $N_\infty^8(e)$ and is the unbounded neighborhood of e.

That we need not in general have $N_r^8(e) \subseteq N_r^9(e) \subseteq N_r^{10}(e)$ for a given r and e can be shown by example. In the thesaurus whose T-graph is shown in Figure 2, $e_{5,3} \in N_3^8(e_{2,2})$ but $e_{5,3} \notin N_3^9(e_{2,2})$, and in the thesaurus of Figure 3, $e_{5,3} \in N_3^9(e_{2,2})$ but $e_{5,3} \notin N_3^{10}(e_{2,2})$.

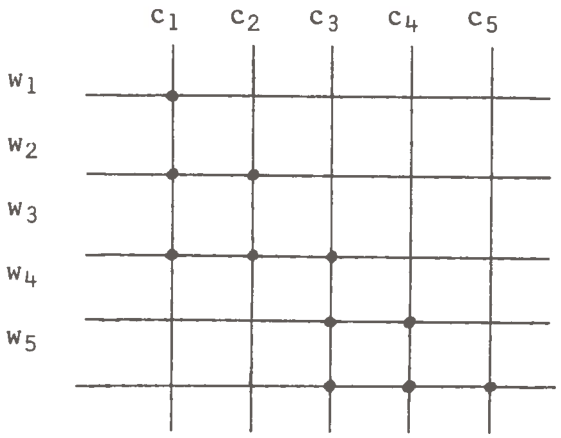


Fig. 2

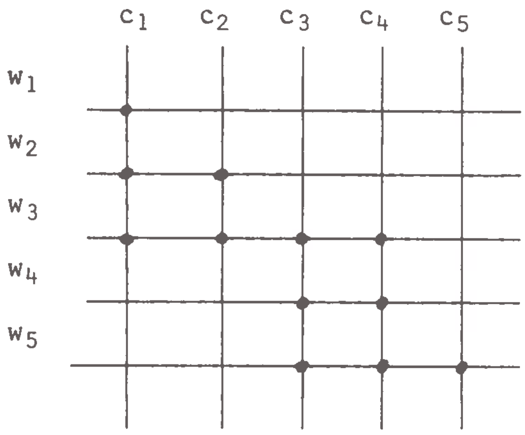


Fig. 3

An interesting property of thesauri is that the existence of a strong c-link implies the existence of a certain strong w-link and conversely.

Theorem 9-5: Suppose $c_1, c_2 \in C$ and (c_1, c_2) is strong. Then for some $w_1 \in r(c_1)$, $w_2 \in r(c_2)$, (w_1, w_2) is strong. If $w_1, w_2 \in W$ and (w_1, w_2) is strong, then (c_1, c_2) is strong for some $c_1 \in r(w_1)$, $c_2 \in r(w_2)$.

It is possible in thesauri for an ε-chain of a given type to reach a "dead end" in the sense that it is impossible to proceed to another entry without violating the restrictions imposed by the definition of ε-chains of that type. In Figure 2, for example, there is no type 8 ε-chain of length 5 whose first four terms are $e_{4,3}$, $e_{4,4}$, $e_{5,4}$, $e_{5,5}$ although there is such a chain of type 7. We call such ε-chains terminal chains and consider the set of all terminal chains emanating from a given entry.

Definition 9-4: $<\varepsilon> \in E^n$ is terminal if no ε-chain in E^n distinct from $<\varepsilon>$ contains $<\varepsilon>$ as an initial segment.

Definition 9-5: Given positive integers n and r with $2 \leq n \leq 10$ and $e \in E$,

i) $T^n_\infty(e) = \{<\varepsilon> \in E^n \mid <\varepsilon> \text{ emanates from } e \text{ and } <\varepsilon> \text{ is terminal}\}$

ii) $T'^{n}_{r}(e) = T^{n}_{\infty}(e) \cap S'^{n}_{r}(e)$

iii) $T^{n}_{r}(e) = T^{n}_{\infty}(e) \cap S^{n}_{r}(e)$

The sets defined in i), ii), and iii) are called terminal stars. The terminology applied to stars in Definition 9-2 is applied to terminal stars in the obvious way.

Although for any thesaurus we clearly have $T'^{n}_{r}(e) \subseteq S'^{n}_{r}(e)$ and $T^{n}_{r}(e) \subseteq S^{n}_{r}(e)$ for a given n, r and e, it is easy to construct examples to show that the inclusion is in general only one way.

The output from one of the existing V.I.A. programs is $T^{8}_{r}(e) \cup S'^{8}_{r}(e)$ for specified r and e. A second V.I.A. program uses as a data base a thesaurus which differs from the thesaurus of Definition 2-1 in that one entry from each category is distinguished. This thesaurus is characterized formally in the next definition.

Definition 9-6: A list structure thesaurus is a quadruple $\langle E,W,C,P \rangle$ where $\langle E,W,C \rangle$ is a thesaurus and P is a subset of E such that $|P \cap c| = 1$ for all $c \in C$. If $c \in C$ and $P \cap c = \{e_1\}$, e_1 is the primary entry of c and elements of $c \sim \{e_1\}$ are the associate entries of c. We say $\langle E,W,C,P \rangle$ is embedded in $\langle E,W,C \rangle$.

It follows immediately from the definition of a list structure thesaurus that the number of distinct list structure thesauri embedded in a given thesaurus, $\langle E,W,C\rangle$, is the product of the orders of c over all $c \in C$.

<u>Theorem 9-6</u>: Given a thesaurus $T = \langle E,W,C\rangle$, there are $\prod_{c \in C} |c|$ list structure thesauri embedded in T.

If e_p, e_q, e_r are entries in a thesaurus $\langle E,W,C\rangle$ such that e_p and e_q are connected by $<\varepsilon_1> \in E^8$ and e_q and e_r are connected by $<\varepsilon_2> \in E^8$, then there exists $<\varepsilon_3> \in E^8$ connecting e_p and e_r. For example, suppose $<\varepsilon_1> = e_1,e_2,\ldots,e_n$ and $<\varepsilon_2> = f_1,f_2,\ldots,f_m$. (Then $e_1 = e_p$, $e_n = f_1 = e_q$, and $f_m = e_r$.) Let i_0 be the least i such that $e_i = f_j$ for some $1 \leq j \leq m$. Suppose $e_{i_0} = f_{j_0}$. If $i_0 = 1$, then $<\varepsilon_3> = f_{j_0}, f_{j_{0+1}}, \ldots, f_m$ is a type 8 ε-chain connecting e_p and e_r. If $j_0 = m$, then $<\varepsilon_4> = e_1,e_2,\ldots,e_{i_0}$ is a type 8 ε-chain connecting e_p and e_r. Suppose neither of $i_0 = 1$ and $j_0 = m$ is true. Let m_1 be the unique element of M having the property that $e_{i_{0-1}} \in m_1$ and $e_{i_0} \in m_1$. Let m_2 be the unique element of M having the property that $f_{j_0} \in m_2$ and $f_{j_{0+1}} \in m_2$. Now one of $m_1 = m_2$ and $m_1 \neq m_2$ is true. If $m_1 = m_2$, $<\varepsilon_5> = e_1,e_2,\ldots,e_{i_{0-1}}, f_{j_{0+1}} \ldots f_m$ is a type 8 ε-chain connecting e_p and e_r, and if $m \neq m$ $<\varepsilon_6> = e_1,e_2,\ldots,e_{i_0}, f_{j_{0+1}},\ldots,f_m$ is of type 8 and connects e_p and e_r. Furthermore, given $e \in E$, the ε-chain of length 1 whose single term is e is in E^8 and connects e to itself and given $e_i, e_j \in E$, if $<\varepsilon> \in E^8$, $<\varepsilon> = e_1,e_2,\ldots,e_n$ and $<\varepsilon>$ connects e_i and e_j, then $e_n, e_{n-1}, \ldots e_1$ is of type 8 and connects e_j and e_i. We have shown in this paragraph that being connected by a type 8 ε-chain is reflexive, symmetric, and transitive on E.

Theorem 9-7: Define a relation R on E as follows:

Given $e_i, e_j \in E$, $e_i \; R \; e_j$ if there exists $<\varepsilon> \in E^8$ such that $<\varepsilon>$ connects e_i and e_j. Then R is an equivalence relation on E.

Theorem 9-8: If $e \in E$ and R is defined on E as in Theorem 9-7,

$[e]_R = N\infty(e)$.

The fact that unbounded neighborhoods are equivalence classes increases our knowledge of the sets $N_\infty(e)$. It follows from Theorem 9-8, for example, that for any $e_1, e_2 \in E$, $N_\infty(e_1)$ and $N_\infty(e_2)$ are either equal or disjoint. We cannot make the same statement about bounded neighborhoods, of course.

Definition 9-7: A thesaurus $\langle E,W,C \rangle$ is *totally connected* if there exists $e \in E$ such that $N_\infty(e) = E$.

Definition 9-8: Given a thesaurus $T = \langle E,W,C \rangle$, $T_1 = \langle E_1,W_1,C_1 \rangle$ is a *subthesaurus* of T if

i) $E_1 \subseteq E$, $W_1 \subseteq W$, and $C_1 \subseteq C$

and ii) $\langle E_1,W_1,C_1 \rangle$ is a thesaurus

and a *proper subthesaurus* if, in addition,

iii) one of the inclusions in i) is proper.

Definition 9-9: Suppose $T = \langle E,W,C \rangle$ is a thesaurus and $E_1 \subseteq E$. The thesaurus $T' = \langle E_1,W_1,C_1 \rangle$ where $W_1 = \{w \cap E_1 | \ w \in W\}$ and $C_1 = \{c \cap E_1 | \ c \in C\}$ is called the subthesaurus of T induced by E_1.

Definition 9-10: Given thesauri $T_1 = \langle E_1,W_1,C_1 \rangle$ and $T_2 = \langle E_2,W_2,C_2 \rangle$,

i) $T_1 = T_2$ if $E_1 = E_2$, $W_1 = W_2$, and $C_1 = C_2$.

and ii) $T_1 \cong T_2$, read 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 for all $e \in E_1$, $w \in W_1$, $c \in C_1$, $e \in w \cap c \Rightarrow \Phi(e) \in \Phi(w) \cap \Phi(c)$.

Definition 9-11: An ordered thesaurus is a quadruple $\langle E,W,C,O \rangle$ where $\langle E,W,C \rangle$ is a thesaurus and O is a linear order on W.

X. Graph Theory Applications to Thesaurus Research

In this section, the numbers of definitions and theorems from graph theory are followed by "(G)".

Definition 10-1 (G): A graph, G, is a pair $\langle V,B \rangle$ where V is a non-empty, finite set and $B \subseteq \{V_1 | \ V_1 \subseteq V \text{ and } |V_1| = 2\}$. V is called the vertex set of G and its elements are the vertices of G. B is called the branch set of G and its elements are called branches.

Theorem 10-1: Let $T = \langle E,W,C\rangle$ be a thesaurus. Then

i) $\langle E,B_1\rangle$ is a graph where $B_1 = \{\{e_i,e_j\} \subseteq E \mid e_i \neq e_j$ and $e_i,e_j \in m$ for some $m \in M = W \cup C\}$

ii) $\langle W,B_2\rangle$ is a graph where $B_2 = \{\{w_i,w_j\} \subseteq W \mid w_i \neq w_j$ and $O(w_i,w_j) \neq \emptyset\}$

iii) $\langle C,B_3\rangle$ is a graph where $B_3 = \{\{c_i,c_j\} \subseteq C \mid c_i \neq c_j$ and $O(c_i,c_j) \neq \emptyset\}$

Theorem 10-1 gives just three of a larger number of correspondences under which pairs of sets definable in the theory of thesauri are graphs. Because of the close relationship between thesauri and graphs, many results from graph theory have important implications in the study of thesauri. The discussion in this section is limited to the correspondence given in iii) of Theorem 10-1, in which vertices are categories and branches are pairs of categories whose overlap is non-null, and definitions and results from graph theory which bear on the formulation of 'grouping devices' in the development of abstract thesauri. By 'grouping devices' is meant formalisms in the abstract system which define subsets of the sets E, W, and C whose counterparts in concrete thesauri can be investigated empirically. Grouping devices we have already seen include subsets of E, W, and C defined in terms of the σ- and r- operators, bounded neighborhoods of various types, and unbounded neighborhoods, which were shown to partition the set E of entries. In the following, certain additional grouping devices which are motivated by graph theory concepts are briefly discussed.

The correspondence of part iii) of Theorem 10-1 gives rise to a second geometric representation of thesauri. The representation of a thesaurus, T, in which categories are represented by dots and elements of $O(c_i,c_j)$ by arcs connecting the dots corresponding to c_i and c_j is called the C-graph of T. The C-graph of the thesaurus, T, whose T-graph is given in Figure 4 is shown in Figure 5.

T-graph:

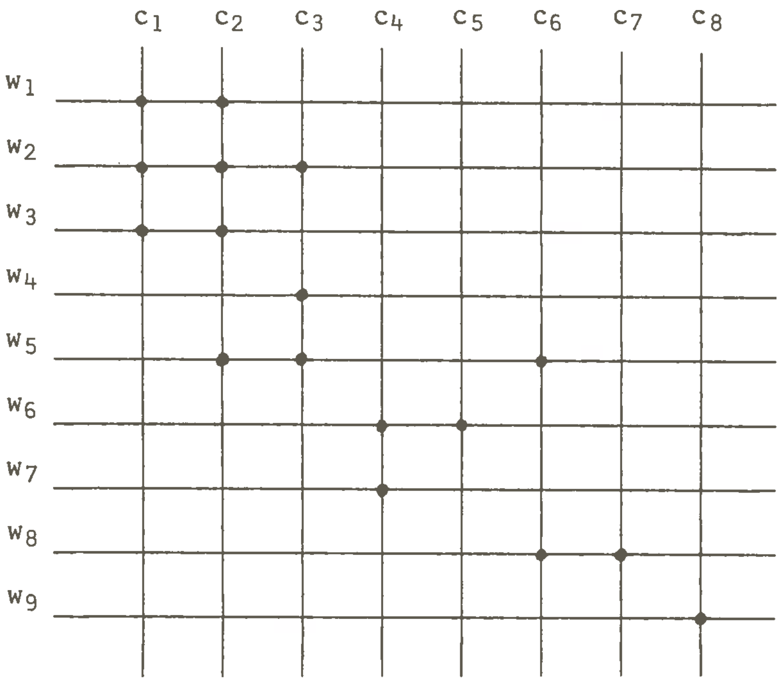


Fig. 4

C-graph:

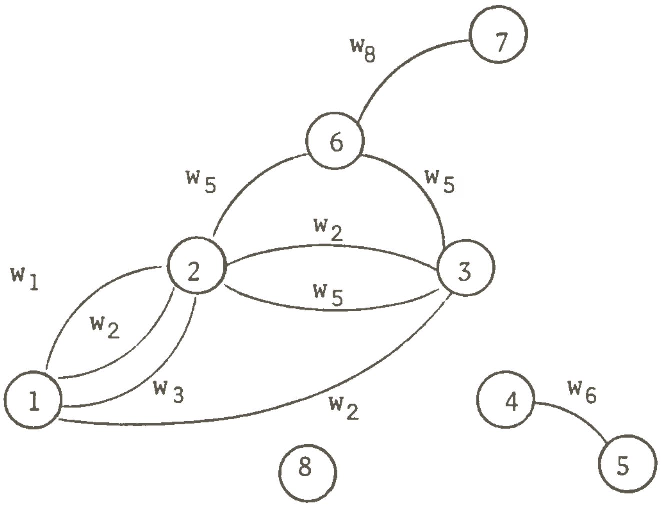


Fig. 5

The configuration in Figure 6, formed from that in Figure 5 by connecting with a single arc vertices connected by at least one arc in Figure 5, is called the reduced C-graph of T. The reduced C-graph of T is the geometrical representation of the graph $\langle C,B_3\rangle$ of part iii) of Theorem 10-1.

Reduced C-graph:

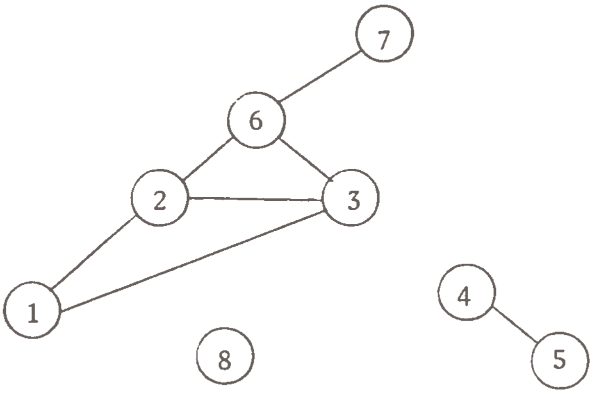


Fig. 6

The three groups into which the eight categories of T are divided in Figure 6 are called components in graph theory and are evidently closely allied with the sets $N_\infty(e)$ of T. A discussion of this and other grouping devices in graphs requires that some terminology from graph theory be agreed upon. Given a thesaurus $T = \langle E,W,C\rangle$, G_T denotes the graph $\langle C,B_3\rangle$ of part iii) of Theorem 10-1.

Definition 10-2 (G): If $G = \langle V,B \rangle$ is a graph, $G' = \langle V',B' \rangle$ is a subgraph of G if G' is a graph and $V' \subseteq V$ and $B' \subseteq B$ and a proper subgraph if one of those inclusions is proper.

If, given a thesaurus $T = \langle E,W,C \rangle$, we let $G_1 = G_T$, $G_2 = \langle C,B_4 \rangle$ where $B_4 = \{\{c_i,c_j\} \subseteq C |\ c_i \neq c_j \text{ and } d(c_i,c_j) \geq 2\}$, $G_3 = \langle C,B_5 \rangle$ where $B_5 = \{\{c_i,c_j\} \subseteq C |\ c_i \neq c_j \text{ and } d(c_i,c_j) \geq 3\}$, etc., then for each $i \geq 1$, G_{i+1} is a subgraph of G_i formed from G_i by deleting n branches where n is an integer ≥ 0.

Definition 10-3 (G): Suppose $G = \langle V,B \rangle$ is a graph and $V = \{v_1,v_2,\ldots,v_n\}$ and $B = \{b_1,b_2,\ldots,b_m\}$. If $b_k = \{v_i,v_j\}$, b_k is said to join v_i and v_j and v_i and v_j are incident with b_k. v_i and v_j are adjacent vertices in G. If $b_l = \{v_h,v_i\}$, b_k and b_l are adjacent branches.

Definition 10-4 (G): If $G = \langle V,B \rangle$ is a graph, $V_1 \subseteq V$ and $V_1 \neq \emptyset$, the graph $\langle V_1,B_1 \rangle$ where $B_1 = \{\{v_i,v_j\} \in B |\ v_i,v_j \in V_1\}$ is called the subgraph of G induced by V_1.

Definition 10-5 (G): Given a graph $G = \langle V,B \rangle$ and $v \in V$, the valency of v in G, $\text{val}_G(v)$ or $\text{val}(v)$, is the number of branches in B incident with v.

It is clear that given a thesaurus $T = \langle E,W,C\rangle$ and $c \in C$, the valency of c in G_T is $|r'(r(c))| - 1$. It therefore follows from the fact that, given a graph $G = \langle V,B\rangle$ with vertices $v_1,v_2,\ldots,v_p$ and q branches, $\sum_{i=1}^{p} \mathrm{val}(v_i) = 2q$, that if $C = \{c_1,c_2,\ldots,c_n\}$, $\sum_{i=1}^{n} |r'(r(c_i))| = |B_3| + n$. Many other results from graph theory having to do with vertex valencies have obvious analogs in the discussion of abstract thesauri.

Definition 10-6 (G): Suppose $G = \langle V,B\rangle$ is a graph and $u,v \in V$ (u and v not necessarily distinct). Then the sequence $\langle p\rangle = v_1,b_1,v_2,b_2,\ldots,v_{n-1},b_{n-1},v_n$ is a u-v path in G if

i) $v_i \in V$, $b_i \in B$ for all $1 \leq i \leq n$,

ii) $v_1 = u$ and $v_n = v$,

iii) $b_i = \{v_i,v_{i+1}\}$ for all $1 \leq i \leq n-1$,

and iv) except that v_1 may equal v_n, the v_i are pair-wise distinct.

The sequence $v_1,v_2,\ldots,v_n$ is the vertex sequence of $\langle p\rangle$. A u-u path other than the sequence whose single term is u is a cycle.

If $T = \langle E,W,C\rangle$ is a thesaurus, $e_i,e_j \in E$ and $<\varepsilon> = e_1,e_2,\ldots,e_n$ is a type 8 ε-chain connecting e_i and e_j, then there exists a c_i-c_j path, $\langle p\rangle$, in G_T, where $e_i \in c_i$ and $e_j \in c_j$, which "contains" <ε> in the sense that the vertex sequence of $\langle p\rangle$ is $\langle c\rangle_\varepsilon$. The converse, however, is not true as not every path contains a type 8 ε-chain.

Definition 10-7 (G): If $G = \langle V,B \rangle$ is a graph and $u,v \in V$, *u and v are connected* if there exists a u-v path in G. *G is connected* if every two vertices of G are connected and *disconnected* otherwise. The relation R = "is connected to" is an equivalence relation on V. Each subgraph induced by the vertices in an R-equivalence class of V is a *component* of G. We will also use "component" to refer to the vertex set of a component.

Given a thesaurus $T = \langle E,W,C \rangle$, if $G_1, G_2, \ldots$ are the graphs defined above and K_i is the set of components of G_i for all $i \geq 1$, then K_1 is a partition of C and K_i is a refinement of K_{i-1} for all $i \geq 2$. If $T = \langle E,W,C \rangle$ is a thesaurus and $K_1, \ldots, K_n$ are the components of G_T, $\{\sigma(K_1), \ldots, \sigma(K_2)\} = \{N_\infty(e) \mid e \in E\}$. Components in graphs, then, correspond to a grouping device that we already have in the formal apparatus of thesauri. Cycle groups, which further subdivide components into sets of tightly connected categories loosely connected to other such sets, introduce a new grouping device to our theory. In order to define cycle groups, we introduce the notion of cut vertices.

Definition 10-8 (G): If $G = \langle V,B \rangle$ is a graph and $v \in V$, G-v is the graph whose vertex set is $V-\{v\}$ and whose branch set is the set of branches in B not incident with v.

Definition 10-9 (G): $c(G)$ denotes the number of components of the graph G.

Definition 10-10 (G): Given a graph $G = \langle V,B \rangle$ and $v \in V$, v is a cut vertex of G if $c(G - v) > c(G)$.

That is, v is a cut vertex of G if its removal disconnects a component of G. Theorem 10-2 gives an interesting characterization of cut vertices.

Theorem 10-2 (G): If G is a graph, w is a cut vertex of G if and only if there exist vertices u and v (distinct from w) in the same component of G such that w is on every u-v path in G.

If $T = \langle E,W,C \rangle$ is a thesaurus and a category c in C is a cut vertex of G_T, then there exist entries $e_i, e_j \in E$ such that for some $e \in E$, $e_i, e_j \in N_\infty(e)$ and c is a term of $<c>_\varepsilon$ for every $<\varepsilon> \in E^8$ connecting e_i and e_j.

Definition 10-11 (G): The cut set of a graph G is the set of all cut vertices of G.

Definition 10-12 (G): If G is a graph and K is a component of G, a subgraph of K whose vertex set has order ≥ 2 and contains no cut vertices, and which is not a proper subgraph of a subgraph of K whose vertex set contains no cut vertices is called a cycle group of G.

Definition 10-13: If $T = \langle E,W,C \rangle$ is a thesaurus and $C_1 \subseteq C$ is a cycle group of G_T, C_1 is also called a cycle group of T, as is $\sigma(C_1)$.

Cycle groups and type 10 neighborhoods represent different methods of defining sets of entries in a thesaurus smaller than unbounded neighborhoods. It is likely that the sets of entries defined by both methods are semantically uniform, although the testing of this hypothesis must await the availability of the complete index for R.I.T. The formulation, purely in connectivity terms and independent of the hierarchy of categories in R.I.T., of formal devices which define semantically uniform sets of entries is an important goal in thesaurus research. Relationships between entries of a thesaurus based on the co-occurrence of those entries in the same category at some level in the hierarchy reflect directly the judgment of the authors of the thesaurus about the semantic closeness of those entries. Relationships between entries defined in terms of connectivity, on the other hand, while dependent on the conscious decision of the author of the thesaurus to group entries into categories in a certain way, are relationships of which he was not aware when making those decisions, and are, in this sense, a step away from human judgment about how close two words are in meaning and a step nearer natural relationships present in the language.

We conclude this section with a discussion of "trees", which provide a method for deleting entries of a thesaurus without disconnecting categories.

Definition 10-14 (G): A *tree* is a connected graph which contains no cycles. If $G = \langle V,B \rangle$ is a connected graph and $G_1 = \langle V_1,B_1 \rangle$ is a subgraph of G, G_1 is a *spanning tree* of G if G_1 is a tree and $V_1 = V$. If $G = \langle V,B \rangle$ is a graph with components $K_1,K_2,\ldots,K_n$ and $T_1,T_2,\ldots,T_n$ are n graphs such that T_i is a spanning tree of K_i for all $1 \leq i \leq n$, the collection $T' = \{T_1,T_2,\ldots,T_n\}$ is a *spanning forest* of G.

Theorem 10-3 (G): Given any graph G, there exists a spanning forest of G.

Suppose $T = \langle E,W,C \rangle$ is a thesaurus and T' is a spanning forest of G_T. Suppose $T' = \{T_1,T_2,\ldots,T_n\}$ and for each $1 \leq i \leq n$, $T_i = \langle V_i,B_i \rangle$. By the definition of G_T, given any $1 \leq i \leq n$ and any $b_{i_j} \in B_i$, if $b_{i_j} = \{c_{i_{j_1}},c_{i_{j_2}}\}$, $O(c_{i_{j_1}},c_{i_{j_2}}) \neq \emptyset$. Construct a set of words by choosing one element of $O(c_{i_{j_1}},c_{i_{j_2}})$, say w_{i_j}, for each choice of $1 \leq i \leq n$ and $b_{i_j} \in B_i$ and let $E_0 = \bigcup_{i=1}^{n} [\bigcup_{b_{i_j} \in B_i} [(w_{i_j} \cap c_{i_{j_1}}) \cup (w_{i_j} \cap c_{i_{j_2}})]]$. Then if $E_0 \subseteq E_1 \subseteq E$, and $T' = \langle E_1,C_1,W_1 \rangle$ is the subthesaurus of T induced by E_1, then, given $c_K,c_1 \in C_1$, if c_K and c_1 are connected in G_T, c_K and c_1 are connected in $G_{T'}$. Moreover, E_0 is minimal in that there exists no proper subset of E_0 having this property.

The brief discussion in this section certainly does not exhaust the results from graph theory whose consequences in thesaurus theory deserve to be investigated. The definition of the graph G_T associated with a thesaurus T provides more than one natural way to assign numerical values to the branches and vertices of G_T so that results about weighted graphs have a direct application. In addition, graph theory provides useful optimization algorithms and construction techniques. Further research on thesauri should further expand the theory of abstract thesauri, explore the possible contributions to that theory from graph theory and other well developed branches of mathematics and, with the availability of the computer accessible version of R.I.T. and a complete index, formulate and test hypotheses in the framework of that theory about the nature of the thesaurus and its application to automated language analysis.

III. Studies in the Semantics of English Prefixation

by Samuel L. Warfel

TABLE OF CONTENTS

Chapter I: Introduction 93

Chapter II: The Problems 99

Chapter III: Abstract Structures 105

Chapter IV: Scope Relationships 116

Chapter V: Surface Structures 128

Chapter VI: Presupposition 144

Chapter VII: Conclusions 149

Footnotes 150

Bibliography 153

Appendix A 156

CHAPTER I

INTRODUCTION

Approaches to the analysis of prefixes in English have varied over the years as the theoretical winds have shifted. Nineteenth century work on English was tied, as was all linguistic analysis, to a primary concern with the history of its words. Thus most of the work on prefixes done by traditional and neogrammarian linguists was directed toward discovering the etymology of English prefixed words being defined often in terms of Latin, French, or German grammatical rules of prefixation. After the Saussurian revolution linguists' emphasis of study shifted to synchronic analysis and centered on the most synchronic data, instances of phonetic utterance. In addition, the phonemic principle which developed from phonological analysis was soon applied to higher levels of abstraction, morphology in particular. The result was that prefixes in English were defined largely in terms of phonetic information, such as stress, and in terms of distributional patterns based on substitution frames.

Generative-transformational grammatical theory continued the emphasis on synchronic analysis although this emphasis emerged not in a preoccupation with phonetic utterances, but rather in its announced goal of describing the competence of the idealized speaker-hearer with the corollary notion that this competence is developed without recourse to historical information save that available in the present structure of the language. Because the process of analysis respective to structural practice was inverted (syntax to phonology as opposed to phonology to syntax) morphological problems have been largely ignored by transformationalists to this point. However, as syntactic structures have become more abstract and semantics has come into what is considered the proper domain of linguistics, questions about the organization of the lexicon and the relationship of semantic constructs to both words and syntactic structures have become increasingly important. It is to certain of these questions that the following paper is directed.

The term 'prefix' is a relatively new term in linguistic description. According to Marchand (1969: 129) older grammarians such as Paul, Matzner, Wilmanns do not use the term. However, by the turn of the twentieth century the term was widely accepted although not precisely defined.

Three works of the first half of this century stand out as major contributions to the study and classification of elements productive in the formation of words in English: Handbuch der englischen Wortbildungslehre by Herbert Koziol, A Modern English Grammar on Historical

Principles, Part VI by Otto Jespersen and The Categories and Types of Present-Day English Word-Formation by Hans Marchand. (It is indicative of the interest in word formation in America that all three authors are Europeans.) Each follows the same basic format of including a general discussion of word formation followed by sections on compounding, suffixation and prefixation. In each case the treatment of prefixes is reviewed largely from an historical viewpoint and a list of prefixes is given with a discussion of the source, the semantic range, and the syntactic restrictions of each prefix along with manifold examples.

Koziol's, the earliest of these major works (1937), is largely a compilation of prior analyses by European grammarians. The prefix section leaves the primary issues largely unresolved, although several are raised and discussed. The basic problem he raises concerning a definition of the term is whether an element which also occurs as an independent word should be considered a prefix when preposed in combination:

> Die Ungrenzung des Begriffes "Vorsilbe" bereitet vor allem bei historischer Darstellung einige Schweirigkeit. Wenn man einen bestimmten Sprichzustand betrachtet, so kann eine Umgrenzung leichter gegeben werden: man kann dann festsetzen, dass unter "Vorselben" nur solche Sprachelements zu vertehen sind, die nicht mehr als selbständige Wörter vorkommen, also z. B. im Ne. die heimischen Silben be-, mis- and un-. (1937: 78)

A little later, however, he expands the definition.

> In der folgenden Besprechung is der Begriff "Vorsilbe" siemlich weit gefasst und es werden daher nicht nur solche Silben angeführt, die als selbständige Wörter nicht mehr vorkommen, sondern auch solche, die auch heute noch als selbstandige Präpositionen oder Adverbien gebraucht werden. (1937: 79)

Beyond the criterion "boundness" explicitly stated, one can discern from his treatment of specific prefixes that his decision to treat 'silben' as prefixes is based largely on etymological information, as is evidenced by his division of the subject into native and foreign prefixes.

He discusses the problem of productivity only perfunctorily with a mention of living and dead prefixes. He includes the dead prefixes in his treatment often labeling them clearly as, for example, the OE prefix ed- 'again' although other prefixes are not so clearly categorized. After his discussion of whether to include as prefixes elements which occur as independent words he does include them in his analysis commenting that often semantic content is different in the two uses of the element (1937: 79). Based as they are on etymological considerations, his prefixed words are not limited to those which combine bound and free forms, but include also words such as _protocol_, _belief_, _forlorn_, _precocious_.

Jespersen discusses Koziol's work in the introduction of the sixth volume of his monumental _A Modern English Grammar_ although he says that two-thirds of his manuscript had been completed when Koziol's book appeared. He recognizes that the two works cover approximately the same ground, but points out several differences. The major differences, as Jespersen sees them, are that Jespersen is more concerned with synchronic analysis, does not recognize the difference between word-formation and flexions, is more concerned with phonetics than spelling and provides more original data.

In spite of the differences, however, Jespersen's list of prefixes is quite close to Koziol's with the exception that Jespersen does not include prefix forms which do not occur even in frozen forms[1] in modern English. Jespersen does not attempt to give a definition of prefix. He does appear to have a better understanding of productivity, and while including frozen forms of dead or dying prefixes, he is more careful than Koziol to mark them as non-productive. Jespersen does include as prefixes, forms which occur as independent words although he divides the set into elements of compounds and prefixes. 'There can be no fixed boundary between words derived by means of a prefix and compounds with a particle as their first element.' (Jespersen 1945: 490) Like Koziol, Jespersen does include words like _precept_, _prelude_, _retroject_, _pantology_ in the list of prefixed words although they do not leave a free form when segmented.

The most recent of the three major works on English word-formation is _The Categories and Types of Present-Day English Word-Formation_ by Hans Marchand. Although the title promises a strong synchronic treatment, the section on prefixes is similar to the two aforementioned books in its discussion of the historical origins of the prefixes listed. A quote from Marchand will present his aim and suggest his view of productivity.

> As my method is primarily synchronic, I have considered as English coinages such words also as are adaptations of foreign words, provided they have been actualized (reinterpreted) in English. That _multidentate_, _subprior_, _transfuvial_ by origin represent L _multidentatus_, _subprior_, _transfluvialis_ may be of historical interest; but what matters linguistically is that these words are analysed as English coinages by the present-day speaker. On the

> other hand, such words as *digonous*, *dimerous*, *dipterous* (representing Neo-Latin words in *-us*) whose second elements do not exist as words in English, have remained outside a formative pattern which is shown by the pronunciation [di] and the stress on the first syllable (as compared with regularly stressed dicoccus, dipolar). (Marchand 1969: 133)

It is important to note that he holds it to be necessary that the present-day speaker be able to segment properly the forms which are to be considered prefixed based on synchronically available information. His requirement that only elements which combine with free forms be considered prefixes forms the primary criterion of his definition.

> Prefixes are bound morphemes which are preposed to free morphemes. In a syntagma AB they fill the position A, i.e. they normally function as determinants of the word B to which they are prefixes. (1969: 129)

This definition places Marchand at odds with Koziol and Jespersen on the question of independent words being considered as prefixes and on the question of prefixes combining with bound forms. Marchand also differs from his predecessors in showing considerably more concern with the semantics of prefixed words and the possible grammatical categories with which particular prefixes can combine.

In the United States nothing approaching the completeness of the three works mentioned above has been written. This is perhaps due at least in part to the theoretical turn which Linguistics took here in the early part of this century. Both Bloomfield (1933) and Sapir (1921) use the term prefix without explicit definition in much the same way as Koziol, at least with reference to English. When using the term to refer to morphemes preposed to words in contexts other than a discussion of English the usage of the term is based on their definitions of morpheme, of which the prefix is but a positional variant.

As post-Bloomfieldians became more absorbed in working out rigorous discovery procedures based on phonetic and distributional criteria with the attendant restriction against mixing levels of analysis, the term prefix either underwent considerable redefinition or other terms were used to refer to the phenomenon previously known as prefixation. Two examples will stand for others which could be given.

Archibald Hill (1958) prefers to reserve the term prefix for elements which participate in concord relationships and coins the term 'prebase' to refer to a number of phonologically and distributionally determined elements which include the traditional prefix *un-* and such elements as [pə] in potato and [s] in svelte. (Hill, 1956: 120) The only elements in English which he is willing to label with his term prefix are the initial consonants in the personal pronouns. His discussion is limited to the examples given and are of little help in analysis of the traditional prefixes.

Zellig Harris (1951) working more from distributional patterns than phonetic information arrives at the same basic list of prefixes which traditional grammarians discuss. (Although he does not give a list, the criteria given would lead to comparable set of forms.) He does not discuss the problem of productivity and considers as prefixes forms which combine with bound forms, i.e. re-ceive, con-ceive, de-ceive. In addition to the traditional prefix forms, he is also willing to accept an analysis of glide and slide in which gl-, sl-, and -ide would all be considered separate morphemes. (Hill, 1958: 193-4) He admits, however, that such an analysis is tentative and would probably be negated by other criteria.

With the challenge to American linguistic theory led by Noam Chomsky in the late 50's came a reversal in the methodology of linguistic analysis. Post-Bloomfieldian linguists were bound by their theoretical position to begin analysis with phonetic information which could then be organized into a phonological statement which in turn provided data for analysis at the morphemic level. The syntactic level was included in theoretical statements but rarely played much of a role in the grammars of the 40's and 50's. Chomsky, on the other hand, began with syntactic structures in his analysis and worked toward the phonetic 'output.' A great deal of work has gone into syntactic analysis by the Post-Chomskyians and considerable work has resulted from an application of the basic syntactic presuppositions to phonology. However, the rules for getting from the output of the syntactic rules to the input of the phonological rules have received little attention. What attention this area has received has come relatively recently as the result of an attempt to understand better the relationship between syntax and semantics, an area of study in which the lexicon plays a key role. A short review of the development of lexical insertion in generative-transformational theory will help to establish a backdrop on which to draw the conclusions which will follow. (A more complete review forms part of Chapin's dissertation (1967).)

Chomsky's early formulation of generative-transformational grammar (Chomsky 1957) dealt only cursorily with the problem of lexical insertion. In Syntactic Structures lexical items are introduced as terminal elements in the phrase structure component of the grammar and thus form part of the tree upon which transformations operate. It soon became obvious that if the espoused goal of generating all and only the grammatical sentences of a language were to be taken seriously some modification of the ways lexical items were introduced would be necessary since the 1957 model of the grammar produced strings which no native speaker would consider part of the English language. First attempts to solve this problem resulted in a proliferation of grammatical subcategories in an attempt to block the insertion of a transitive verb into a tree with only one noun phrase or an intransitive verb into a tree with two noun phrases, to mention only two problems. At this point the lexical insertion was still done with a rewrite rule in the phrase structure component.

In Aspects of the Theory of Syntax Chomsky attempted to clear up this problem of lexically specific context sensitivity by dividing the Base into a categorial (phrase structure) component and a lexicon which inserted lexical formatives in positions marked by 'dummy symbols' through lexical transformations. This allowed the categorial component to remain context free and uncluttered with hundreds of grammatical categories, and still allowed lexical formatives to be placed only in positions compatible with their idiosyncratic syntactic restrictions. This modification (with the notion of syntactic features, which facilitates the modification) worked quite well in a grammar which did not have to interface with a semantic component. However, when work was begun on the problems of providing just such a component a number of problems arose.

Chomsky argued in Aspects that the structures emanating from the Base were sufficient to the interpretation of the sentences which would result upon the application of the transformational rules. However, by the time that the book was published, or shortly thereafter, there was evidence from a number of studies based on the Aspects model by first generation generative-transformationalists such as Lakoff (1965) Gruber (1965) and Chapin (1967) that a proper interface of semantic and syntactic structures required that some transformations occur earlier in derivations than the point at which lexical formatives were inserted. In fact these and later studies pointed to a grammar in which the semantic structures were formally identical to the categorial component, and lexical insertion occurred both before and after transformations had applied to the Base structures. Chomsky has rejected this position and maintains that all lexical formatives are inserted at one point in the derivation although he has modified his position to allow semantic interpretation rules to take into consideration intermediate and surface structure as well as the deep structure he maintained was sufficient in Aspects.

Another related problem which has separated those who see semantic structures as generative (generative semanticists) from those who see them as the results of interpretative rules (interpretivists) concerns the organization of the lexicon relative to lexical formatives which are interpreted similarly although they function as different parts of speech in surface structure, e.g., nominalization of verbs. Robert Lees has argued in an early generative-transformational monograph (Lees 1960) that considerable economy in a grammar could be effected if noun phrases containing a deverbal noun with the same semantic reading as sentences containing the same elements were related transformationally. This position was generally accepted without question until the discussion of the relationship of semantics and syntax began to heat up after 1965. The generative semanticists saw Lees proposal as corroboration for their contention that transformations precede at least some lexical insertions while Chomsky and the interpretivists maintained that only a limited number of nominalizations could profitably be treated as derived from the application of transformational rules. While the heat of these discussions has abated somewhat the smoke still obscures a number of issues.

The implications of a semantic and syntactic analysis of prefixation in English are considerable for the theory of generative-transformational grammar and a determination of the correctness of the two sub-theories presented above. In the next chapter I will set forth what I believe these implications to be and the basic arguments from prefix analysis which I see as relevant to the issues involved in the general theory.

CHAPTER II

THE PROBLEMS

It is the thesis of this paper that some words which have traditionally been considered as prefixed must be considered as the result of a process similar to the semantic-syntactic processes which result in sentences. At least this must be the case if a grammar is to be formally explicit in its description of the indefinitely large number of sentences possible in the language and if the grammar either provides a semantic reading for each sentence or at least is compatible with a semantic component which provides such a reading. It is further maintained that the grammar should only account for those prefixes which are synchronically productive in the formation of words and then only when they are attached to free forms with a semantic reading which can be described in terms of the meaning of the prefix and the free form and the structural relationship between the two. In this chapter I will discuss several issues involved in this thesis and introduce the arguments which will be made in subsequent chapters.

The term 'productive' (along side the older term 'living') has been used to describe prefixes for a long time, yet I have not been able to find a work which deals with the specific definition of the term nor with a number of ramifications related to it. As used here productivity will refer to word-formation processes which combine semantically and syntactically stable elements into new configurations to produce words for which the semantic reading is predictable. A prefix will thus be considered productive if it is possible to attach the prefix form to a member of a class of independent words with a resulting word whose meaning is clear to a hearer in the normal communication process. However, prefixes cannot be easily divided into productive and non-productive classes since there are degrees of productivity as a look at dictionary entries for words with *pre-* and *un-* will indicate. This degree of productivity is here understood in terms of the number of types with which combination is possible as opposed to the number of tokens with which a prefix occurs, although it may be that this number is significant in the acquisition of prefix processes in children, i.e., the fact that *untie* is an often used word may contribute to the acquisition of *un-*.

Based on this definition of productivity a study was done by Warfel and Harris[1] (1971) with 55 first through sixth grade children which required that the children analyze and produce prefixed words based on the prefixes *un-* 'reversal' and *re-* 'again' plus nonsense words. The results show that even third graders are able to form new words by combining the prefix *un-* with words they have never heard before and can do so with considerable consistency. The same is true for the prefix *re-* although the acquisition of this process appears to be later, at about the sixth grade. This study provides strong evidence that a grammar which accounts for all the sentences of the language must deal with prefixation as a productive process.

This view of productivity also forces a reevaluation of the relative productivity of syntactic processes and word-formation. Heretofore, in linguistic theory, it appears to have been tacitly assumed that the lexicon was the most static of the elements of a grammar and that the syntactic component was the most productive. This seems to be based on the relatively large number of sentence structures in which a given word can occur as compared to the number of words which are found in combination with a given prefix. However, if transformational rules are considered instead of possible syntactic structures, prefixation appears to be more productive in some cases. For example, while the number of words which freely combine with the prefix *un-* without any question of grammaticality may run into the thousands, the number of verbs which allow the application of the not-transportation rule[2] can be numbered on the fingers of two hands. It therefore appears that the distinction between the lexicon and the syntactic component of English is less well-differentiated than has been assumed.

Given the view of productivity discussed above prefixation must be defined in terms of several criteria.[3] These criteria can best be understood by looking at the prefix as a bundle of phonological and semantic information. Presuming that there are a limited number of semantic primes which function in some cognitive grammar, the rules of language map these semantic elements and relationships onto a linear string of phonologically encoded morphemes. These rules plus a lexicon must include both the regularities of the language and the word-specific idiosyncrasies of the language. Because of these idiosyncrasies the mapping is quite complex. Relative to the problem of prefix analysis, these non-unique correspondences look something like this: (The semantic terms in brackets are merely illustrative and are not claimed to be primes.)

SEMANTICS		[reversal]	
PHONOLOGY[4]	un-	dis-	de-
	untie	disown	deactivate

Or seen from the perspective of phonology:

PHONOLOGY		de-	
SEMANTICS	[reversal]	[remove from]	[get off of]
	deactivate	dethrone	detrain
		[derive from]	
		deverbal	

The mapping is even more complex when the phonology and semantics of both the prefix and the root are taken into consideration. The following chart shows some viable possibilities for both prefix and root of a number of words relative to the prefix pre- and the semantic element [before]. Under the PREFIX heading the pluses and minuses indicate whether the word at the left has the form pre- and/or the meaning [before]. In the ROOT column the signs indicate whether the string of phonemes remaining after the prefix form is removed constitute the form of a proper word and whether this form has a meaning consonant with the meaning of the entire word.

	PREFIX		ROOT	
	Phonology	Semantics	Phonology	Semantics
predetermine	+	+	+	+
prevent	+	-	+	-
predict	+	+	-	
preach	+	-	-	
foreknow	-	+	+	+
precede	+	+	+	-

The position taken here is that only words which would be marked with pluses in all columns should be considered prefixed words, i.e., only words which are true combinations of correlate semantic and phonological elements should be considered to be prefixed words in a grammar which is semantically oriented and formally explicit.

Therefore, for the determination of whether or not a word is to be considered prefixed in the sense of productivity discussed above, four kinds of information are necessary:

1) the meaning of the prefix, that is, the possible meanings of a particular prefix form,

2) the meaning of the word which remains when the particular prefix form is removed,

3) the rules for determining the semantic relationship of this particular prefix to the roots to which it is attachable, and

4) the meaning of the entire word.

To be considered a prefixed word by this definition it must be the case that the sum of the information obtained in 1) through 3) is identical with that in 4). In other words, to be a true prefixed word the meaning of the word must be predictable from the elements and their relationship.

To apply this criterion to a word such as _prevent_ it is necessary to know what _pre-_ means, what _vent_ means, the semantic-syntactic relationship between roots and _pre-_, and the meaning of the complete form _prevent_. In this case the word is not considered to be prefixed unless the relatively unusual reading of [vent before] is given to the word since the more usual reading of [prohibit] does not agree with the composite of information about the elements of the word and their relationship.

It should be clear from this definition that relative to the scholars whose works were discussed in the first chapter this treatment will differ along several lines.

1) As opposed to criteria based on etymology, distribution, phonetics, or spelling this study will define prefixes by purely semantic criteria with some consideration of syntax (if the distinction can be made clear).

2) Only productive elements will be considered to be prefixes and these only in combinations which fit the criteria above.

3) No particular distinction is attempted between compounding and prefixation, although the prefixes discussed here will include only bound elements.

4) Words such as *receive*, *respect*, and *deduce* will not be considered to be prefixed even though the prefix forms in question are productive and the strings of phonemes remaining when the prefix form is removed are 'recurring partials' which combine with a number of other prefix forms. Only prefix forms attached to independent words will be considered true prefixes.

The decision to define prefixes by semantic criteria alone is not without adverse consequences. The most serious of these is that in order to capture a number of generalities about the stress patterns, consonant alternations, and vowel reductions for large classes of words in English a morpheme boundary of some kind must be postulated in words which do not qualify as prefixed by the definition given here. This is particularly true of Latin or French borrowings whose meanings have changed from true prefix-root composites to unanalyzable monomorphic words. Thus while it is not semantically helpful to consider *receive*, *deceive*, *perceive*, and *conceive* as prefixed words, they do form a class ending in *-ceive* in which the phonological alternations of [i] with [E] and [v] with [p] in the 'recurring partial' occur regularly before -*tion*, i.e., *reception*, *deception*, *perception*, and *conception*.

Chomsky and Halle (1968) are aware of the problem of matching the structures required for syntax with those necessary for proper analysis of phonology, in particular stress placement.

> In preceding chapters we had had occasion to note that the surface structure required as input to the phonological component will not in all cases be identical with the surface structure that can be syntactically motivated. Thus in English *Fifth Avenue* has a different stress pattern from *Fifth Street*. The rules of the phonological component will yield this difference if *Fifth Avenue* is not dominated by the node 'noun.' Syntactically, however, there is no justification for treating *Fifth Avenue* any differently from *Fifth Street*. (Chomsky and Halle, 1968: 269)

The solution which they suggest, but do not work out in detail is that the grammar will require 'readjustment rules' which will modify the syntactic surface structure to make it compatible with the phonological component. They also discuss the problem concerning morpheme boundaries, suffixes, and stress placement and conclude the following:

> It is no doubt possible to find rules of some generality governing the deletion of # before affixes, rules which will perhaps reflect (or even sharpen) the traditional distinction of derivational and inflectional processes and which may depend on a distinction between affixes added by transformation and affixes that are assigned by

> processes internal to the lexicon. But there are many obscure questions here involving the proper dividing line between the lexical and transformational components of a generative grammar, and, since we have arbitrarily excluded problems of syntax from consideration in this study, we must leave this matter in its present unsettled state. (Chomsky and Halle, 1968: 370)

I take the position that the dividing line between lexical and transformational components as regards prefixation is probably non-existent, as was mentioned in my discussion of productivity and will be discussed in following sections. However, since I have arbitrarily excluded problems of phonology from consideration in this study, I will leave the particular problem of the interface between syntax and phonology in its present unsettled state.

This is not to say that the semantic problems related to prefixation are in a settled state. However, it is hoped that this paper will provide some directions and guides for further study in the area. An attempt will be made to relate the specific problems of prefix analysis to the broader questions of current theory where some corroborating evidence will be given for several positions held by generative semanticists and some extensions of general principles will be suggested.

In particular, the problems of prefixation impinge upon generative-transformational theory most sharply in the area of lexical insertion, an area over which generativists and interpretivists are in considerable disagreement. The general theory posits a component of grammar which generates phrase structure rules by context-free rewrite rules the terminal nodes of which are ultimately replaced by lexical formatives. Whether the categories which label the nodes of the tree thus generated are syntactic, semantic, or both and whether the terminal nodes are replaced at one point in the derivation or replaced at various points form the crux of the differences between the two sub-theories.

As to prefixed words there are several possible ways that insertion could be effected:

1) Prefixed words as single units replace single nodes in syntactic phrase structure trees. This position would in effect deny the existence of prefixes since prefixed words would be treated in the same manner as monomorphic words. In addition, this treatment would fail to capture the generalization that recurring forms preposed to free forms lead to partially identical semantic readings. Thus, this treatment would greatly increase the size of the lexicon. This position is also untenable if the semantics of entire sentences with prefixed words is to be clearly and precisely analyzed. The chapters which follow will argue this point.

2) Prefix forms and roots replace separate nodes in syntactic phrase structure trees with the proper changes in boundary markings and category reassignment being achieved by transformational rule. This position appears to be compatible with the Aspects model although Chomsky does not explicitly deal with prefixation. While this treatment recognizes prefixes as separate semantic elements and avoids the redundancy of position 1), it is not capable of dealing with the semantic reading of sentences with prefixed words as will be explained in the following chapters.

3) Prefix forms and roots replace separate nodes or entire subtrees in semantic constituent structure trees often after a number of transformational rules have operated on the trees. This is basically the generativists' position regarding the insertion of lexical formatives although no work has been done which deals specifically with the problems of lexical insertion of prefixed words. I will argue that this view of lexical insertion is necessary to account both for the surface forms which occur in English and the proper semantic readings for these forms as they occur in sentences.

The arguments which I will present for this position will be organized according to the types of structure which must be accounted for in a completely explicit, semantically based grammar. In particular, the arguments will fall into four areas: abstract structures, prefix scope relationships, surface structure, and presuppositions.

CHAPTER III

ABSTRACT STRUCTURES

Before I deal with abstract structures a few words on the criteria for arguments in generative-transformational grammar is appropriate. From the first formulation of the theory one of the most important principles of syntactic argument has been that sentences or parts of sentences which 'mean the same' should have the same underlying structures at some point in their derivation. (Where that point is and whether it is the same for all sentences is a point at issue in the generativist-interpretivist debate.) This principle has not been questioned except to discuss the content of 'mean the same' especially as linguists have taken responsibility for describing inferential relationships, presuppositions, and reference. The second basic principle which is a corollary to the first is that sentences which are ambiguous should have a number of possible underlying structures equal to the number of semantic readings possible for the sentence. Again, although people have argued about what it means to be ambiguous and about the locus for the different structures in the theoretical framework, the principle has been accepted as a basic part of the task of linguists. There are other criteria for syntactic and semantic arguments the discussion of which I leave to those with greater experience than I. These two, however, are sufficient to provide a basis for the first argument I wish to make, namely, that prefixes must be derived from underlying semantic constituent structures.

Consider the following sentence pairs.

3.1a John opened the store again.
b John reopened the store.
3.2a Jane was not capable of crying.
b Jane was incapable of crying.
3.3a The linguistic student's grade before the examination was an A.
b The linguistic student's preexamination grade was an A.
3.4a The patient's condition after the operation was satisfactory.
3.4b The patient's postoperative condition was satisfactory.
3.5a The child was too active.
b The child was hyper active.
3.6a The American colonies had no central government during the period before the revolution.
b The American colonies had no central government during the prerevolutionary period.

The sentences in each pair are sufficiently synonymous to require that any semantically based grammar show that fact by deriving both sentences from closely similar if not identical semantic structures. (The degree of similarity among these pairs varies according to a number of factors, some of which will be discussed below.)

In the simplest cases this requirement means that the semantic information given as terminal nodes in the deepest semantic structure will be replaced optionally by either a free form or by a prefix. For example, sentence 3.5a will result if the lexical formative too replaces a node [too], but 3.5b will be the result if the prefix hyper- replaces [too]. (The symbol [too] is intended here to be a semantic rather than lexical term.) This analysis accomplishes two things for the grammar: 1) the two sentences and the two different elements are formally recognized as semantically equivalent, and 2) the complexity of the semantic component of the grammar is reduced since one underlying structure is required instead of two.

In the more complex cases among the sentence pairs given, it will be seen that there are ambiguities which must be accounted for by two or more underlying structures which result in the same surface structure.

The best example of this is the pair 3.3 in which the (b) sentence is ambiguous since it can be a paraphrase of the (a) sentence or may mean something like 3.7.

3.7 The linguistic student's grade on the preliminary examination was an A.

Sentence 3.3b must then have two underlying sources, one which is the same as that underlying 3.3a and one for semantic reading roughly corresponding to 3.7.

In a grammar which derives nominals such as examination from corresponding verbals, in this case examine, the difference in semantic readings can be attributed to structures which result in different surface bracketing. With sentence 3.3b there are two bracketings which correspond to the two readings as in 3.8.

3.8a (pre(examination))
b ((preexamine)ation)

The (a) bracketing corresponds to the reading of 3.3a and the (b) bracketing to that of 3.7. In deepest structure 3.8b will require the presence of two examine predicates since the reading requires that there be two 'examinings,' one of which precedes the other. Exactly what that deeper structure may look like will depend upon current work on the nature of semantic structures and the entire process of nominalization. It suffices here to show that the prefixed forms in sentences like 3.3b are ambiguous and that their readings require underlying structures which take into account the constituent structure of prefixed words.

What is tacitly assumed in the above presentation concerning the sentence pairs of 3.1 through 3.6 is that the correspondences of prefixes with prepositions or adverbs are generalizable across the language and that most sentences with prefixed words have a corresponding paraphrastic construction with a closely related meaning. Empirical evidence supports this assumption although there is a problem with the time reference for the prefix pre- under some circumstances.[1] On the other hand, there is not always a prefixed word which corresponds to a paraphrastic construction with a preposition or adverb. For example, 3.9b, while interpretable, is not an exact paraphrase of 3.9a.

3.9a John's closing of the door before Sally got there disturbed me.
b John's preclosing of the door disturbed me.

There are two problems with 3.9b. The first is that close does not combine with the prefix pre-, at least not in normal usage, and, thus, the sentence is of questionable grammaticality. This is simply a case of the already accepted fact that each word will have to be marked to indicate which prefixes it can combine with. The other problem with the sentence is that it does not specify overtly the time reference for the prefix. While 3.9a indicates that the closing of the door happened before Sally got there, sentence 3.9b simply indicates that the closing of the door was before some appointed time which is not specified.

These problems notwithstanding, positing the same semantic elements for both prefixes and corresponding prepositions and adverbs not only captures the generalization that sentences which differ only in that one has a prefix and the other has the corresponding adverb or preposition have similar semantic readings but it also accounts for certain facts concerning permissible surface structures. For example, it is generally accepted in generative-transformational theory that attributive adjectives originate in relative clauses in underlying structure, are reduced by the deletion of the relative pronoun, and are moved to their position before the noun. This analysis accounts for a number of facts in English noun phrases and is accepted here. An illustration will be helpful. Sentence 3.10 (the example and analysis are taken from Burt, 1971)

3.10 A pink panther wobbled by.

is derived from an underlying structure of which 3.11a is a rough approximation. (The parentheses enclose the imbedded S.)

3.11a A panther (a panther was pink) wobbled by.
b A panther who was pink wobbled by.

Sentence 3.11b is the result of a relative clause formation transformation. After Relative Clause Reduction has applied the structure underlying the ungrammatical 3.12 results.

3.12 *A panther pink wobbled by.

Modifier Shift will prepose the adjective, however, to give the correct attributive adjective position and the grammatical sentence 3.10. Each of these rules is well justified in the literature as necessary to capture generalizations concerning English syntax.

I take the position that considerably economy and important generalizations can be effected in the grammar of English if prefix forms are described with the same mechanisms. Look at the following pairs of sentences which result from the process described above. First the underlying structures. (No trees are given since this process is well described in the literature. See Burt, 1971: 67-93.)

3.13a The celebration (the celebration was happy) was long and loud.
b The celebration (the celebration was after graduation) was long and loud.

After relative clause formation the following sentences would result if no other transformations were applied.

3.14a The celebration which was happy was long and loud.
b The celebration which was after graduation was long and loud.

Relative Clause Reduction produces an ungrammatical (a) sentence but a correct (b) sentence.

3.15a *The celebration happy was long and loud.
b The celebration after graduation was long and loud.

Of particular interest are the next two sentences which result from the application of Modifier Shift. Notice that the grammaticality of the two sentences is reversed from the preceding stage of derivation.

3.16a The happy celebration was long and loud.
b *The after graduation celebration was long and loud.

This is a rather involved way of saying that adjectives must be preposed relative to the nouns they modify and that prepositional phrases cannot be preposed.

These constraints on possible word order are, of course, only applicable to English since there are numerous languages where adjectives are characteristically postposed and a sizable number which allow prepositional phrases to be preposed. A good example of the latter is German where a noun phrase like 3.17 is not uncommon.

3.17 Die nach Kriegsverhaltnisse . . .
The after war conditions . . .

Of interest here is the fact that where there is a 'hole' in the pattern of preposing, where prepositional phrases are concerned, English allows prefixes with basically the same semantic readings. At least this is true for several prefix-preposition pairs. Consider 3.18 which paraphrases the ungrammatical 3.16b.

3.18 The postgraduation celebration was long and loud.

If prefixed words are derived from different underlying structures than that for prepositional phrases the grammar must be complicated by rules of some sort which indicate the relatedness of the prefix and paraphrastic constructions, and by rules in addition to the adjective preposing rules are needed only for prefixed words.

Thus the analysis I propose here which derives both prepositional phrases and prefixed words (where there is a correspondence) from the same underlying structures, reduces the redundancies of the grammar, and takes advantage of movement transformations which are necessary to the grammar in any case.

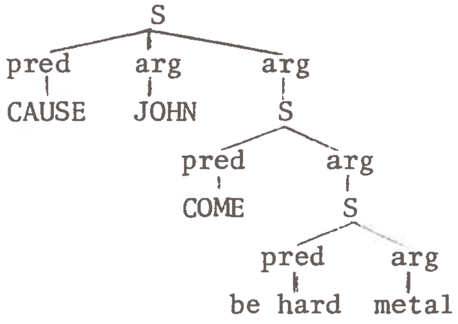
To this point the underlying structures discussed have been strongly syntactic in nature and relatively form-oriented as to level of abstraction. However, I would now like to present evidence for a more abstract underlying structure which will be shown to be necessary to deal systematically with prefixation.

One of the most appealing analyses of generative semanticists has been the causative-inchoative[2] argument for the decomposition of lexical items. This argument, the early forms of which are found in Lakoff's dissertation (1965: IV-4-IV-18), posits underlying constituent structures for single lexical formatives in the case of a number of verbs, hence the term 'decomposition.' In brief Lakoff relates the sentences of 3.18 through what he calls 'pro-verbs,' that is, verbs which are necessary in the semantic structure but may not appear in that form on the surface.

3.18a The metal is hard
b The metal hardened
c John hardened the metal

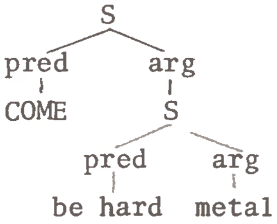
These three sentences are related in this analysis in that each sentence is imbedded in the following sentence. The Semantic structure of the (c) sentence is given as the string 3.19 where the capitalized verbs are pro-verbs.

3.19

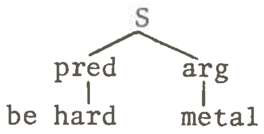


The structure of (b) and (a) are given in 3.20 and 3.21.

3.20



3.21



A considerable number of studies have been done using this framework with the result that there are several good arguments for the analysis. I will mention one argument based on work done by Binnick (1969) which will have a bearing on prefix analysis.

Action verbs can be classified as either punctiliar or durative depending upon whether the action is conceived to be non-time consuming or time consuming. The classification is corroborated by syntactic tests using time reference adverbial phrases. For example, notice that in 3.22 the verb *hit* is compatible with the puntiliar time adverb *at 4:00*

3.22 George hit the dog at 4:00.

While in 3.23 with the durative time adverbial the action must be understood as iterative.

3.23 John hit the dog for three hours.

On the other hand the verb *held* is intrinsically durative as sentences 3.24 and 3.25 illustrate.

3.24 Sally held the ball at 4:00.
3.25 Sally held the ball for 30 seconds.

Notice that 3.24 is a strange sentence which must be understood as meaning 'started to hold' if it means anything. Sentence 3.25, however, is quite natural.

Given that verbs are intrinsically punctiliar or durative, consider sentence 3.26 with verb *lock.*

3.26 John locked the lion cage at 12:00.

This seems to argue that the verb is punctiliar. However, look at 3.27 with a durative time adverbial.

3.27 John locked the lion cage for four hours.

This perfectly acceptable sentence is not understood to mean that John spent four hours turning the key but rather that the cage was in a locked condition for four hours. Notice that it is possible to include both types of time reference in the same sentence.

3.28 At 12:00 John locked the lion cage for four hours.

The facts concerning the verb *lock*, therefore, require either that the distinction between the two classes of verbs be given up, which leaves the facts regarding verbs of the types *hit* and *hold* without explanation or that another analysis be made of the sentences with verbs of the type *lock*. The causative-inchoative analysis by Lakoff referred to above provides a satisfying analysis.

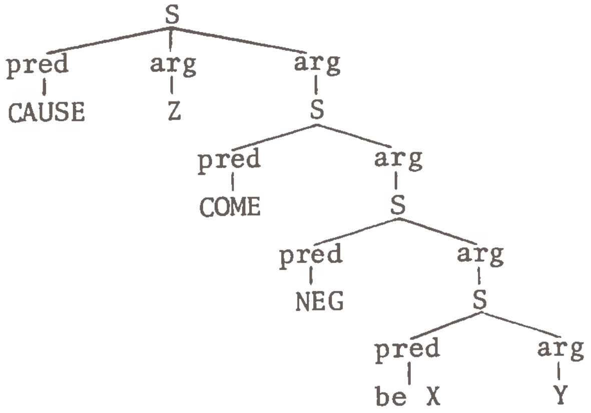
If *lock* is analyzed in the semantic structure as a structure with three predicates (pro-verbs) as in 3.29

3.29 CAUSE to COME to be locked

then the time adverbials can refer to different parts of the underlying structure and punctiliar-durative distinction preserved. In the case of 3.28 the punctiliar time reference can refer to the CAUSE verb and the durative reference to the lowest verb. This is intuitively satisfying since it is what speakers of the language understand to be the case in interpreting the sentence, i.e., at 12:00 John caused the cage to be in locked condition for the next four hours.

This analysis of verbs provides a means for simplifying the semantic component of a grammar by reducing the number of semantic primes. In particular if, as Lakoff suggested in his dissertation (1965: IX-19), the notion of reversal can be stated in the semantic structure (where X may be any of a large number of semantic predicates and Y and Z are any referents) as follows:

3.30



then the mechanism already necessary to account for the facts presented above relative to causative-inchoative sentences can be used to describe prefixes with reversal meanings.

For example, the following sentences would be related by the same underlying semantic structure in the causative-inchoative analysis.

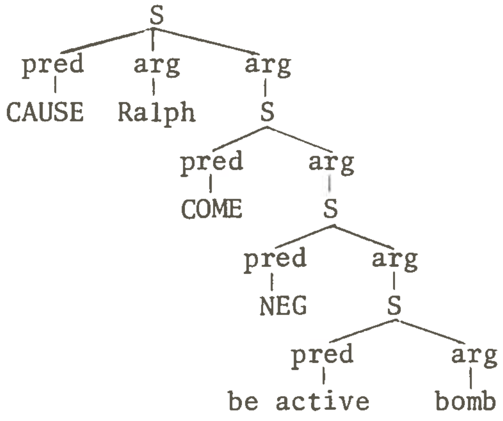
3.31 Ralph caused the bomb to come to be active.
3.32 Ralph caused the bomb to become active.
3.33 Ralph caused the bomb to activate.
3.34 Ralph activated the bomb.

I take these sentences to be paraphrases with closely similar if not identical meanings and to be roughly equivalent to the stages of collapse which the transformational rules would effect in deriving the most 'compact' of the surface structures, 3.34. What is of interest to the analysis of prefixes is that these sentences have corresponding sentences with an underlying negative which gives a consistent meaning of reversal although the surface forms differ at various stages of derivation.

3.35 Ralph caused the bomb to come to be inactive.
3.36 Ralph caused the bomb to become inactive.
3.37 Ralph caused the bomb to deactivate.
3.38 Ralph deactivated the bomb.

These sentences derived from an underlying semantic structure with the negative semantic pro-verb NEG roughly like 3.39.

3.39



This analysis claims that the semantic structure underlying these sentences has alternative surface manifestations governed primarily by the lexicon insertion rules which allow certain lexical formatives to be substituted for various subtrees in the underlying structure. In the example 3.39 the subtree corresponding to 'COME to BE' can either be realized as come to be or become. The structure underlying 'NEG active' must become a negative prefix, in this case in- (although the free form not is a marginally grammatical form) and the larger structure 'to COME to NEG BE active' surfaces either as a combination of the other elements or is signaled by the prefix de-. The prefix form de- may also signal the causative pro-verb and so ambiguously indicates either the inchoative-negative structure or the causative-inchoative-negative structure.

Applying this analysis to the sentences with *lock* discussed above works equally well although the prefix form required for *lock* is *un-*. Notice that although the form is different for the two prefixes the semantic restrictions are the same.

3.40 John caused the lion cage to come to be unlocked.
3.41 John caused the lion cage to become unlocked.
3.42 John caused the lion cage to unlock.
3.43 John unlocked the lion cage.

Notice also that while the surface form *un-* occurs in all four sentences the significance is different in the first two where *un-* signals simply the negative as does *in-* in 3.35 and 3.36. The prefix form *un-* is thus three ways ambiguous (four if the intensive of *unthaw* is considered) being the surface signal for an underlying negative, inchoative-negative, or causative-inchoative-negative structure.

Evidence that this underlying structure is necessary can be educed from sentence 3.44.

3.44 At 12:00 John unlocked the lion cage for four hours
and it made me uneasy all afternoon.

As Lakoff has argued convincingly (1970a: 154-7) the antecedents of pronouns must be constituents, i.e., a pronoun cannot refer to elements in more than one constituent unless, of course, each element is dominated by a larger constituent the whole of which is the antecedent. Given that this is true, at least one reading of 3.44 requires that the pronoun *it* refer to the state of the unlocked cage, a reference which is impossible without the lexical decomposition analysis argued for here. With this analysis, however, 'NEG be locked' is a constituent[3] to which the pronoun can refer and the reading 3.45 is formally statable.

3.45 At 12:00 John unlocked the lion cage for four hours
and its being unlocked made me uneasy all afternoon.

Another argument for lexical decomposition and abstract semantic structures underlying prefixes has to do with the ambiguities of sentence 3.46.

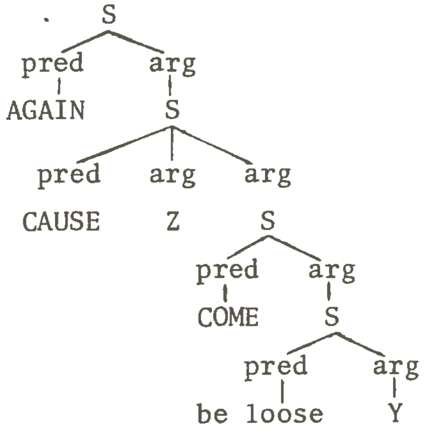
3.46 Luke reloosened the lariat.

At least two readings are possible. It may be the case that Luke is loosening the lariat which he loosened before or it may be that he is loosening a lariat which he did not loosen before, but which someone else loosened. If one accepts the necessity of deriving both *re-* and the adverb *again* from an underlying semantic prime [again] as was argued for *pre-* and *before*, the ambiguity can be specified in the semantic structure by analyzing *loosened* as the surface form of an underlying causative-inchoative construction with *loose* something like 3.47 (ignoring *again* for the moment).

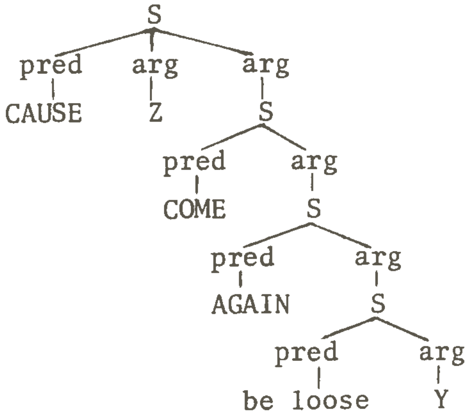
3.47 CAUSE to COME to BE loose

In this framework the ambiguity is described by specifying whether the adverbial is predicated of the lowest verb, 'BE loose' or the highest verb, 'CAUSE.' In the first case only the loosening is claimed to be a recurrence and Luke's causing it may be his first loosening. In the second case Luke is repeating an action of loosening which he has done before. These structures are given in 3.48 and 3.49.

3.48



3.49



These two semantic structures also underlie sentences 3.50 and 3.51 respectively while 3.52 is ambiguous in the same ways as 3.45.

3.50 Again Luke loosened the lariat.
3.51 Luke caused the lariat to loosen again.
3.52 Luke loosened the lariat again.

Notice that in order to indicate unambigously in the surface structure that it is the lowest verb whose action is repeated (i.e., reading 3.49) it is necessary to include the surface form of the CAUSE pro-verb. Given that fact, it should be the case that the causative verb plus the prefixed form of the lowest verb will have the same reading as 3.51. Sentence 3.53 confirms this.

3.53 Luke caused the lariat to reloosen.

Thus sentences 3.51 and 3.53 are derived from 3.49, 3.50 is derived from 3.48, and 3.46 and 3.52 are ambiguous in that they may be derived from either 3.48 or 3.49.

A complete formalization of the process which places adverbials correctly and accounts for the facts presented here will require more work on the entire problem of adverb placement, an area of study fraught with difficulties since adverbs appear to be subject to the most flexible word order constraints of any grammatical category. While lexical insertion of the re- prefix in cases such as those mentioned appears to be a relatively simple one since the rules of predicate raising which transform the deepest structure into the structure required for the insertion of a single verb will retain the semantic node [again] and the insertion rule for re- will replace that node with a prefix on lexical verb formatives which allow combinations with re-. However, correlating this process with the movement transformations is a problem which I am unable to solve at this point. (For further discussion of adverb placement see Keyser, 1968.)

The facts and arguments presented in this chapter give convincing evidence that any grammar which purports to be both semantically oriented and completely explicit will have to deal with prefixation as a process involving highly abstract underlying constituent structures.

CHAPTER IV

SCOPE RELATIONSHIPS

As linguistic theory has increasingly concerned itself with semantics it is natural that formal structures and terminology from logic have become more prevalent in linguistic literature. Of particular interest to the study of prefixation is the term 'scope' and its application to natural language phenomena. The term was used quite early in generative-transformational literature by Klima (1959 although not printed until 1964) in his ingenious treatment of negation.

> The principle grammatical notions developed in this study concern the scope of negation (i.e., the structures over which the negative element has its effect) and the structural position of the negative element in the sentence. The scope of negation varies according to the origin of the negative element in the sentence (over the whole, over subordinate complementary structures alone, or only over the word containing the negative element). (1964:316)

However, the term was not used in connection with its more technical usage in logic until natural language quantification began to cause problems with other wise well substantiated transformational rules, e.g., the semantic reading of passives with quantified noun phrases.[1] After accepting some of the framework of logicians the next 'logical' step was to experiment with using the notion of operators and their scopes in other problem areas, i.e., specificity, definiteness, question, negation, etc.

Since 1968 generative-transformationalists have begun to face the problems of interfacing the syntactic structures they have developed with semantic structures. In trying to characterize formally the semantic structures necessitated by increasingly abstract syntactic structures they have accepted large portions of the framework developed by symbolic logicians, although not without critical scrutiny and translation into forms already familiar from their work on natural language. Within the general theory, the generativists have been most in the fore of this movement. In 1968 when the differences which separate generativists from interpretivists were becoming clearly defined, Bach suggested that a system of operators be incorporated in the Base of the grammar as a way to deal with specificity, questions, reflexives, pronominalization, etc. He defines scope in terms of the familiar syntactic formalism of linguists.

> We define the _scope_ of an operator Q as the string dominated by the highest $\underline{S}$ to which Q is prefixed. If an operator Q is followed immediately by a variable _x_ then we say that every occurrence of _x_ within the scope of Q is _bound_ by Q. Among the operators will be a _generic_ operator, an _all_ operator, a _some_, a _focus_ or definiteness operator, a _question_ operator, and the like. (Bach 1968: 106)

Since that time considerable work has been done on English using the notion of operators and the number of operators has greatly increased to include the 'quantity' words and most adverbs. (See, for instance, Frazer 1971 on _even_)

This chapter will argue that any grammar which adequately accounts for the semantic readings of prefixed words will have to include the notion of scope in order to account for paraphrase relationships and ambiguities. The argument will center on analysis of sentences such as 4.1.

4.1a The girl was not attractive enough for the job.
b The girl was unattractive enough for the job.

The difference, of course, depends on whether the business man or his wife is selecting his secretary. I will argue that the semantic structure can best describe this difference by indicating the relative scope of the negative element and _enough_.

First some preliminaries are necessary. The analyses presented here will be given in the framework of generative semantics in which the phrase structure rules, semantic structures and nodes correspond to elements of symbolic logic as discussed by McCawley.

> If one accepts one of the proposals that would do away with VP as an underlying category ... then not only is the correspondence between 'deep' syntactic categories and the categories of symbolic logic exact, but the 'phrase structure rules' governing the way in which the 'deep' syntactic categories may be combined correspond exactly to 'formation rules' for symbolic logic, e.g., the 'phrase structure rule' that a Sentence consists of a 'Contentive' plus a sequence of Noun Phrases corresponds to the 'formation rule' that a proposition consists of a *n*-place predicate plus an 'argument' for each of the *n* places in the predicate.
> (McCawley 1970: 221)

Since, however, the problems of incorporating all of the facets of language of which linguists are aware within a single unitary formal structure even for a single sentence are considerable, the few structures given here should be considered fragmentary and accurate only to the degree that they reflect the point at issue for a given structure. In the particular case of prefixes and scope relationships I will treat negation and adverbs as originating in the semantic structure as predicates and will suggest necessary constraints on the derivation from these structures to the surface forms which occur with comparable semantic readings.

Consider sentences in 4.2, 4.3, 4.4, and 4.5.

4.2 Quickly Luke again loosened the lariat.
4.3a Luke loosened the lariat quickly again.
b Again Luke quickly loosened the lariat.
c Again Luke loosened the lariat quickly.
d Luke again loosened the lariat quickly.
4.4a Quickly Luke loosened the lariat again.
b Luke quickly loosened the lariat again.
c Luke loosened the lariat again quickly.
4.5a Luke quickly reloosened the lariat.
b Luke reloosened the lariat quickly.

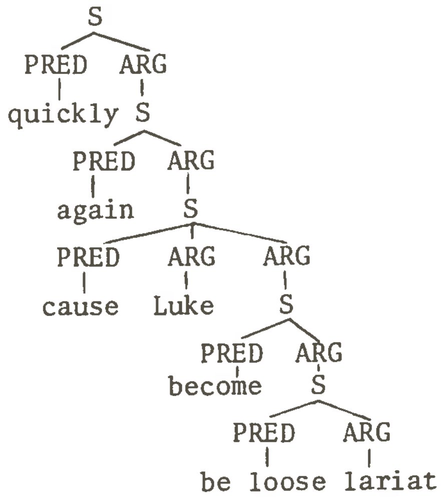
These sentences are obviously related to sentence 3.46 and others which formed the bases for the discussion in Chapter III of lexical decomposition and differ primarily in the addition of the adverb *quickly*. Although a good sample is given of possible work orders, no attempt was made to be exhaustive since the semantic variation illustrated is amply represented.

Of concern to the analysis of prefixation is the ambiguity of sentences 4.5a and 4.5b concerning whether or not the notion of *again* is relevant to the causing or the state of being loose as discussed in Chapter III. However, the addition of *quickly* raises the additional question of whether the loosening was done quickly both times or only the first or second time. My intuitions concerning the sentences are as follows:

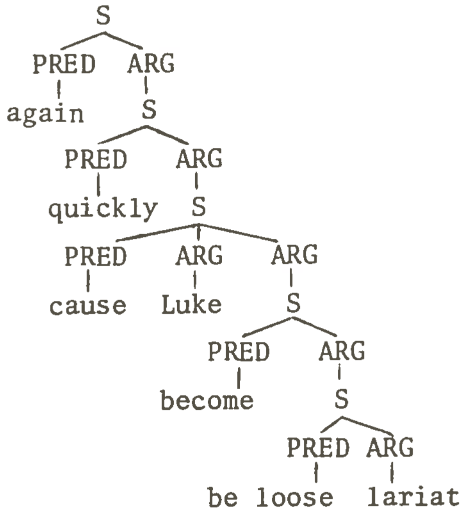
4.2'--Luke loosened the lariat twice, but only the second time was done quickly.
4.3'--Luke loosened the lariat twice and both times he did it quickly.
4.4'--Luke may or may not have loosened the lariat the first time, but in either case only the second time was done quickly.
4.5'--Both of these sentences have the same reading as 4.4.

These readings represent all the possible readings for permutations of the basic words. Notice that there is no word order which indicates that only the first loosening was done quickly, nor is there an order which specifies unambiguously that Luke was involved in but one loosening. Using formalism suggested by a number of generativists but ultimately my own, the readings paraphrased in 4.2', 4.3' and 4.4' can be diagrammed as the following trees (where the primed numbers indicate the correspondence).

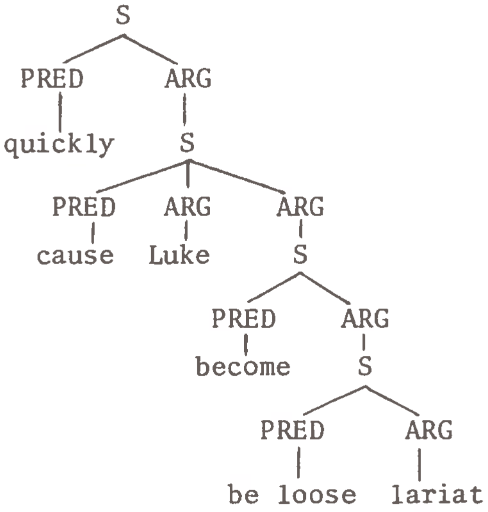
4.2"



4.3"

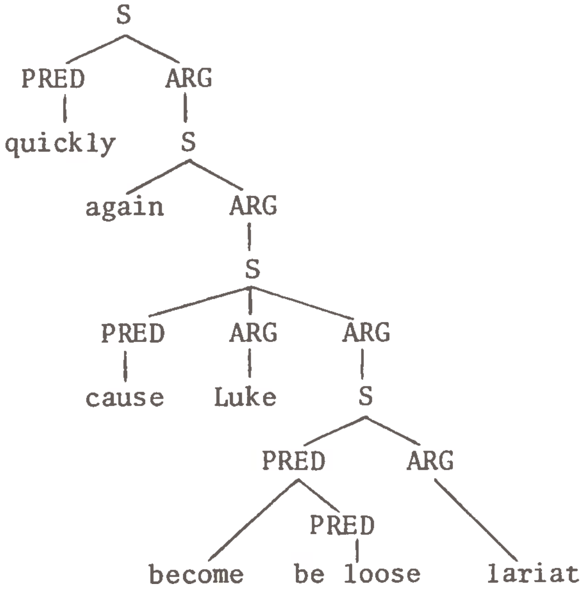


4.4" Either the following tree or tree 4.2".

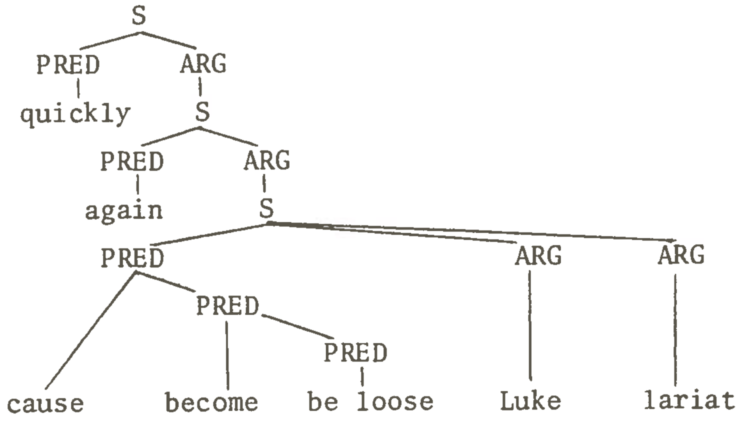


Following McCawley (1968) the predicate-raising transformation applies to the structures given to form structures to which lexical transformations can apply to insert the proper lexical items. The predicate-raising rule applies cyclically from the lowest S to the highest as is illustrated by applying the rule to 4.2" as given in 4.2"a and 4.2"b.

4.2"a

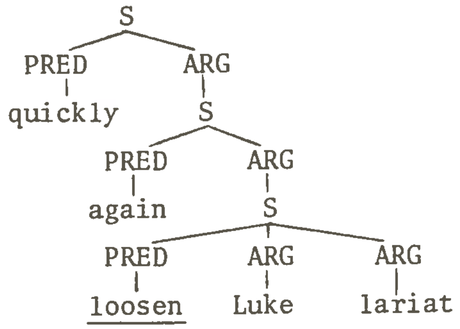


4.2"b



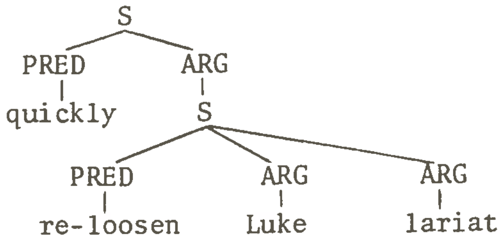
At the point in the derivation given in 4.2"b the complex predicate structural description for the lexical insertion transformation of the lexical item loosen is satisfied and the subtree for that predicate is replaced with the lexical entry to give a structure like 4.2"c.

4.2"c



(Underlined words indicate lexical formatives as opposed to the non-underlined words which represent semantic elements.) Later rules operate to provide the proper placement of the adverb predicates for sentence 4.2. However, an optional lexical transformation can apply to 4.2"c to replace the 'again' node with the prefix re- subject to the selectional restrictions of the lower predicate and other restrictions discussed below. Since loosen is marked to allow the re- prefix the rule can operate to give 4.2"d.

4.2"d



The structures given here are of course considerably truncated in that the trees must also contain information about the insertion of *Luke* and *lariat*, the proper indications of definiteness, the tense of the verb, and probably others. In addition, later rules specify the possible word order of subject, verb, object, and adverb. I shall not discuss these problems further since they have no bearing on the topic under discussion.

Turning now to the question of scope relationships and the consequent restriction on prefixation it should be noted that none of the possible permutations of the paraphrased sentences 4.2 through 4.4 allow an unambiguous reading like 4.6.

4.6 Luke loosened the lariat only the second time and that time it was done quickly.

This follows naturally if one holds that the determining factor in differentiating sentences which have the reading that both loosenings were done quickly from those which indicate that only the second loosening was done quickly is whether or not the predicate *quickly* is in the scope of (dominated by) the predicate *again* or vice versa. A look at the structures posited shows that 4.2" and 4.4", both of which indicate that the second time was the only one done quickly, reveals that *again* is dominated by *quickly* in both cases. Tree 4.3", however, has *again* dominating *quickly*. Since to indicate that Luke did not cause the first loosening it would be necessary to have the again predicate attached to the lowest S and therefore dominated by the *quickly* node which would be identical with 4.4", the underlying structures postulated thus make a correct prediction about possible interpretations.

Of particular interest is the fact that the semantic structure of 4.3" does not have a surface form with the prefix *re-*. I attribute this to a restriction on the lexical transformation for the insertion of *re-*, which for now I will specify as 4.7.

4.7 The prefix *re-* can be substituted for the semantic element 'again' just in case 'again' does not dominate a semantic element which is ultimately realized as a lexically independent adverb on the surface.

Looking at the same restriction from the surface structure it is always the case that in a sentence with the occurrence of both the prefix *re-* and an adverb (or an as yet unspecified subset of adverbs), the semantic element 'again' will always be interpreted as within the scope of the adverb. This restriction appears to hold for all verbs with *re-* although, of course, I have investigated only a sample of verbs and adverbs.

However, as stated the constraint on the re- insertion transformation appears to be both too restricted and too loose. It is too restricted in the sense that the generalization it states appears also to be true of a number of other prefixes and other classifications of semantic elements. Consider, for example, the following sentences.

4.8a John was not prepared for all his classes.
b John was not prepared for áll hìs clàsses (but he was for some).
c John was unprepared for all his classes.
d *John was unprepared for áll hìs clàsses (but he was for some).
4.9a Flossie was not welcome in all the bars.
b Flossie was not welcome in áll thè bàrs (but she was in some).
c Flossie was unwelcome in all the bars.
d *Flossie was unwelcome in áll thè bàrs (but she was in some).
4.10a Percy was not affected by all the drugs.
b Percy was not affected by áll thè drùgs (but he was by some).
c Percy was unaffected by all the drugs.
d *Percy was unaffected by áll thè drùgs (but he was by some).

Notice that the (a) and (c) sentence are synonomous at least as regards the semantic information relevant to the discussion to follow. These sentences are intended to be representations of 'normal' intonation and stress patterns. The (b) sentences, on the other hand are intended with an intonation and stress pattern which would be natural if the sentences were followed by the clauses given. The interesting thing to notice is that the (d) sentences are ungrammatical when followed by the given clauses and/or stressed as indicated.

The intonation and stress differences in the (a) and (b) sentences signal a semantic difference which I claim is a difference in the scope relationships between the underlying universal quantifier, all and the NEG element which occurs here overtly as not. In the (a) sentences the NEG element occurs within the scope of the universal quantifier with meaning of 4.11

4.11 For all x it is not true that P

where x is a variable and P is a proposition. On the other hand, the (b) sentences have an underlying structure with the quantifier occurring within the scope of the NEG element with the meaning of 4.12.

4.12 for not all x it is true that P

Since, as mentioned earlier, much of the debate on the nature of the underlying structures of sentences has centered on quantification and negation there is a considerable literature on the subject (Carden 1967, 1969, 1970, Jackendoff 1969, Hall 1970, Lakoff 1970, Heringer 1970, Labov 1972). In this literature readings which have a logical negation of the verb such as 4.11 are referred to as NEG-V and readings in which the quantifier is negated as in 4.12 are called NEG-Q (Labov 1972). I will adopt this terminology in the ensuing discussion.

In the sentences of 4.8 through 4.10 the same generalization which was made for the sentences with re- appears to hold true. Therefore, the observation of 4.7 can be generalized to 4.13.

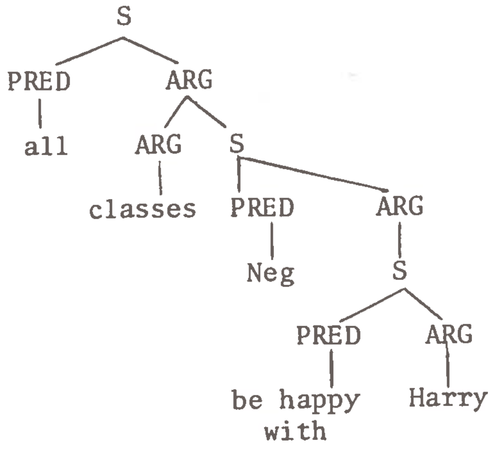
4.13 If an underlying semantic structure contains both a prefixable semantic element and another scope-involving predicate (adverbs, quantifiers, negations, etc.), prefixation can occur only if the prefixable semantic element is within the scope (dominated by) the scope-involving predicate; in all other cases the prefixable element must occur as an independent lexical item.

Consider the sentences of 4.14 which will serve more fully to illustrate this point and also will form the basis for an additional constraint later.

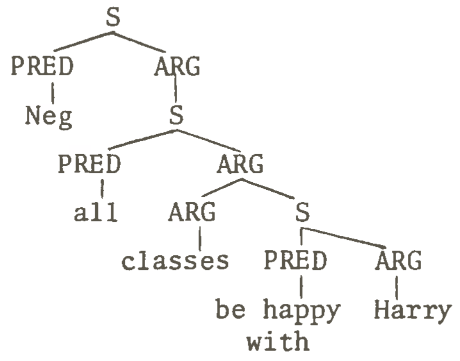
4.14a Harry was not happy with all his classes.
b Harry was not happy with áll hìs clàsses (but he was with some).
c Harry was unhappy with all his classes.
d *Harry was unhappy with áll hìs clàsses (but he was with some).

I posit the structures 4.15 and 4.16 as underlying the NEG-V and NEG-Q readings respectively. (The his has been ignored as irrelevant.)

4.15



4.16



Underlying structure 4.15 can be paraphrased roughly as 'for all classes Harry is not happy with them' and structure 4.16 as 'Harry is happy with not all classes.' Given these underlying structures the generalization of 4.13 holds true. Thus as we suggested earlier generalization 4.7 was too restricted. It remains to be shown that both generalizations 4.7 and 4.13 are too loose. Sentences 4.17 through 4.20 are relevant to the issue.

4.17a Perry wasn't able to persuade all the jurors.
b Perry wasn't able to persuade áll thè jùrors (but he was some).
c Perry was unable to persuade all the jurors.
d Perry was unable to persuade áll thè jùrors (but he was some).
4.18a The teacher wasn't qualified to test all the students.
b The teacher wasn't qualified to test áll thè stùdents (but she was qualified to test some).
c The teacher was unqualified to test all the students.
d The teacher was unqualified to test áll thè stùdents (but she was qualified to test some).
4.19a Mark wasn't willing to sell all his medals.
b Mark wasn't willing to sell áll hìs mèdals (but he was some).
c Mark was unwilling to sell all his medals.
d Mark was unwilling to sell áll hìs mèdals (but he was some).
4.20a Harry was not happy with Sue's spending all the money.
b Harry was not happy with Sue's spending áll thè mòney (but he was with her spending some).
c Harry was unhappy with Sue's spending all the money.
d Harry was unhappy with Sue's spending áll thè mòney (but he was with her spending some).

As is obvious these sentences are similar to those of 4.8 through 4.10. There are two important differences, however: 1) the (d) sentences in this group are acceptable in spite of the fact that they have NEG-Q readings and prefixed negative elements in violation of generalization

4.13, and 2) they all have quantifiers in embedded complement sentences of the type defined by Chomsky's level of deep structure, i.e., a much shallower syntactic structure than the highly abstract semantic structures proposed by generativists.

Sentences 4.21 and 4.22 should establish the facts more clearly.

4.21a The teacher wasn't qualified in all the disciplines.
b The teacher wasn't qualified in áll thè dìsciplines (but she was qualified in some).
c The teacher was unqualified in all the disciplines.
d *The teacher was unqualified in áll thè dìsciplines (but she was qualified in some).
4.22a The teacher wasn't qualified to teach all the disciplines.
b The teacher wasn't qualified to teach áll thè dìsciplines (but she was qualified to teach some).
c The teacher was unqualified to teach all the disciplines.
d The teacher was unqualified to teach áll thè dìsciplines (but she was qualified to teach some).

The crucial examples are 4.21d and 4.22d. Notice that despite the fact that the same predicate occurs in both sentences 4.21d is ungrammatical since it cannot have a NEG-Q interpretation while 4.22d is grammatical with a NEG-Q reading. I claim that these sentences differ only at a relatively shallow level in that 4.22d has an embedded sentence as complement of the verb while 4.21d has a prepositional phrase. This example indicates that the differences in the (d) sentences of 4.8 through 4.10 and the (d) sentences of 4.17 through 4.20 is not a difference in predicates since the same predicate with different complements exhibits the same regularity of difference.

The implication of these facts is considerable for the insertion of the prefix _un-_ and even more considerable for the theory of generative semantics. If the interpretation of predicates with the prefix _un-_ depends in part on whether or not the predicate is followed by an overt sentential complement then the lexical insertion transformation must have this information available at the point in the derivation where it applies. However, it must also have available the specification of scope relationships in order to block sentences such as 4.21d. The significance of these requirements becomes obvious in light of some of the basic tenets of generative semantics.

Three of the basic positions of generativists are as follows:

4.23 All information about a sentence necessary for its proper semantic reading is specified in the most underlying structure to which lexical and syntactic transformational rules apply to derive the surface structure.

4.24 A corollary to the above is that transformations do not change meaning, but simply map more abstract onto less abstract structures in such a way as to constrain the possible structures of the language to agree with the relationships of surface sentences and their semantic readings.

4.25 Although I have not found an explicit statement to this effect, the formalism indicates that generativists hold that the distinction between overt complements of the type discussed by Rosenbaum and others in the _Aspects_ model and the complements of prelexical structures is vacuous. In other words, overt complements play no role in semantic structure.

If the principles of 4.23 through 4.25 are accepted then there is no simple or 'natural' way to account for the restrictions on the lexical insertion rule for the prefix _un-_. The following approaches might be taken to avoid the difficulties and still maintain these principles:

1) Feature notation might be used to indicate negation and universal quantification on the proper elements. This would allow the insertion transformation to operate quite late in the derivation after predicate raising and pruning have left structures with overt complements thus satisfying the requirement that structural description of the insertion rule include such information. However, as I have argued here and others have argued elsewhere (Lakoff 1970b) the range of possibilities where several quantifiers are involved cannot be adequately described without the notion of scope.

2) Verbs in both the repertoire of semantic elements and the lexicon will have to be marked as to the number and kind of arguments (complements) they can or must occur with. This information might be incorporated into the insertion transformation so that only those predicates which take sentential complements allow the insertion of _un-_ under the configuration of nodes for the NEG-Q reading. However, since there is no formal differentiation between sentential complements of the type found in the deep structure of an _Aspects_ model grammar and the sentential complements which compose the core of the lexical decomposition theory the rule will not be able to operate correctly.

3) The only alternative I see which adequately accounts for the facts is to incorporate the notion of subordination into the semantic structure. This could most easily be done by adding a category to the semantic structure so that alongside S there would be a subordinate PROPosition. PROPosition could be used in lexical decomposition structures while the use of the symbol S was reserved for overt complements.

CHAPTER V

SURFACE STRUCTURES

As the underlying structures of transformational grammars have become more semantic in character, the number of categories present in the deepest structure have been reduced to a handful, the exact number varying from scholar to scholar. While this movement toward category reduction has facilitated the logical description of sentences, there remains the problem of mapping these few categories and their configurations of constituent structure onto the more numerous semantically and syntactically relevant surface categories and constructions. This problem is further complicated by the difficulties of deciding what the appropriate surface categories are. Even given relatively clear criteria for inclusion in a category there are a sufficient number of lexical items which defy easy cataloguing to make the entire enterprise untidy. This untidiness was probably first discovered by an eighth grade student of 'diagraming' who went outside the examples in his English grammar text to look at a sentence from his literature text. In this chapter I will consider some of the problems of semantics-to-surface-structure mapping as they affect the process of prefixation.

The sentences of 5.1 are not synonymous although there are semantic similarities.

5.1a The swift horse ran around the track.
   b The horse ran swiftly around the track.

The differences in surface structure are obviously the differences in the word order and the occurrence of the morpheme *-ly* in 5.1b. The semantic differences in the two sentences are traditionally described in terms of the element 'modified', i.e., *swift* modifies the noun *horse* in 5.1a and *swiftly* modifies the verb *ran* in 5.1b. Generative semantics has described the semantic differences through the use of underlying constituent structure. The adjectival construction of 5.1a is a transformation of an underlying S, an analysis which relates this sentence to 5.2.

5.2 The horse which was swift ran around the track.

The adverbial construction of 5.1b, on the other hand, has been described by analyzing the adverb as a higher 'verb' in semantic structure. Although considerable evidence has been given for this underlying structure, the transformations which place adverbs are not as well described in the literature as those involved in adjective preposing. I wish to show, however, that the relationships of underlying semantic structures and surface structures require that the mapping process be more complex than the accepted analysis summarized here.

I will first argue that a distinction must be made in the semantic structure between predications of qualities and predications of actions or processes. This distinction is the one traditionally made through the use of the terms 'adjective' and 'verb' respectively. Since these terms have been variously defined, primarily with reference to surface structure distribution, I will use the terms 'adjectival' and 'verbal' to refer to the semantic distinction which I wish to make. I will further argue that this distinction can best be made by using the pro-verbs already required by the lexical decomposition theory plus a pro-verb BE.

Consider sentence 5.3.

5.3 The window was broken.

This sentence is ambiguous in at least two ways. In the first case broken is understood as a verbal in semantic structure and is inflected as a verb in surface structure. The underlying structure has undergone both passive and agent deletion transformations to derive the surface structure. The second reading is a predication concerning the quality, state, or condition of the window, the agent of the breaking being of importance only in the sense that one's philosophical bias that there must be causes for certain states has become bound up in the English language. That these two readings are possible can be seen in the syntactic consequences of choosing one over the other. For example, the adjectival reading easily allows the use of present tense while the verbal reading does so only in special contexts. For example, look at the sentences of 5.4 and 5.5.

5.4 The window is broken.
5.5a The window was broken by John.
b The window is broken by John.
c The window is broken. (verbal)

Notice that sentence 5.3 in its adjectival sense is easily changed into present tense with only the semantic difference predicted by such a change, i.e., in 5.3 the state existed in the past while in 5.4 the state exists in the present. However, notice that while 5.5a is quite normal as a passive in the verbal reading, sentences 5.5b and 5.5c are not simple present tense equivalents but rather must be understood in special contexts. These contexts correspond to the usual contexts required for the non-progressive present tense usage of verbs in English: iterative, stage directions, on-the-spot reporting, and perhaps a few others. For example, sentences 5.6 provide contexts in which sentence 5.5c could be used.

5.6a The window is broken every time someone slams the door.
b 'Take that, and that!' (The window is broken and the light goes out.)

5.6c 'Here we are at coffee table side, folks, for the fight between John and Marsha. Marsha throws an ash tray, John counters with a lamp. The window is broken as both combatants head for the kitchen for more ammunition.'

Another syntactic consequence of the two readings is more to the point of this paper. As I mentioned in Chapter III the prefix *un-* serves four functions in English as illustrated in 5.7.

5.7a Percy unlocked the door. (causative-inchoative-negative)
b The door unlocked and the knob turned with a squeak. (inchoative-negative)
c The door was unlocked when we arrived. (negative)
d The meat had unthawed by the time we got home. (intensive)[1]

Of particular interest to the discussion of adjectivals and verbals is the fact that the negative *un-* can only occur with adjectivals. For example, look at the sentences of 5.8.

5.8a *Ralph unbroke the window.
b *The window was unbroken by Ralph.
c *The window was unbroken. (in the sense of 5.5c)

That this constraint is not a function of the lexical choice *break* can be seen in the sentences of 5.9 where the verb *tie* is used.

5.9a Sally untied the knot.
b The knot was untied by Sally.
c The knot was untied.

While these are grammatical in the sense of reversing the action of tying, they cannot be interpreted to mean that Sally did not tie the knot. The difference in the sentences of 5.8 with *break* from those of 5.9 with *tie* appears to be only that while *tie* can take both the negative *un-* and the reversal *un-*, *break* can only combine with the negative *un-*. In both cases the (c) sentences can have the negative reading only when the past participle is understood in the adjectival sense.

While the facts presented here appear quite clear there are a few cases which are not so clear-cut. These cases involve no more than a dozen words which appear to share both verbal and adjectival qualities in both the semantic and surface categories. Sentences 5.10 and 5.11 will illustrate.

5.10a The drug affected George.
b George was affected by the drug.
c George was unaffected by the drug.
d *The drug unaffected George.

5.11a The paintings impressed Bella.
b Bella was impressed by the paintings.
c Bella was unimpressed by the paintings.
d *The paintings unimpressed Bella.

Notice that the relationship between the (a) and (b) sentences appears to be that of an active to a passive construction. However, the (d) sentences with the negative un- are ungrammatical while the apparent passives of the same sentences are acceptable as the (c) sentences illustrate. If the (c) sentences are analyzed as passives several adverse consequences result.

1) The generalization that only adjectivals can combine with negative un- is no longer valid since a few verbals may also be combined.

2) The un- insertion rule must operate after the passive and then only if the passive has been applied. While I see no reason why the rules could not be thus ordered, the constraint on the insertion rules leaves unstated an important generalization, since it makes no use of the verbal-adjectival distinction which appears to be the determining factor.

3) The fact that the lexical items in question share several characteristics with the adjectivals discussed above and do not share some of the characteristics of transitive verbs calls the passive analysis of the (c) sentences into question. In particular, in addition to the fact that affect and impress take the negative un- they also allow the present tense in the usual adjectival sense as may be seen in 5.12 and 5.13.

5.12 George is unaffected by the drug.
5.13 Bella is unimpressed by the paintings.

Both these sentences are understood as referring to the state of the surface subject rather than an iterative or stage direction sense of action. That these lexical items do not share some verbal characteristics can be seen in sentences 5.14 and 5.15.

5.14a George was unaffected by the decision to sell the company.
b George was unaffected by the fact that his wife was missing.
c ?George was unaffected by Gloria.
d George was unaffected by Gloria's low-cut gown.
5.15a Bella was unimpressed by the dinner.
b Bella was unimpressed by the signing of the treaty.
c ?Bella was unimpressed by William Buckley.
d Bella was unimpressed by Buckley's record.

My intuitions concerning these sentences are not completely clear. However, it does appear to me that the (a), (b), and (d) sentences with non-animate, non-volitional objects of *by* are more acceptable than the (c) sentences with human objects. Notice that the active, positive counter-parts of all the sentences of 5.14 and 5.15 are quite acceptable. I do not pretend to understand how these sentences should be analyzed, but *affect* and *impress* do act differently from the better understood transitive verbs.

This blurring of category distinctions corroborates Ross' position that instead of dealing with discrete categories linguists may have to learn to work with a 'quasi-continuum' which he calls a 'squish.' (Ross 1972) He suggests that there is 'squish' from verb to noun which he describes as follows:

5.16

verb > present participle > perfect participle > passive participle > adjective >

preposition(?) > 'adjectival noun' (e.g., fun, snap) > noun

(Ross 1972: 316)

The facts of negative *un-* insertion substantiate this analysis since *un-* combines with categories beginning with the quasi-passive participles which I have discussed through prepositions such as *unlike*. While it is possible for negative *un-* to occur with the present participle form, I contend that these cases must be dealt with in the same way as the perfect participles. For example, 5.17 shows that the same pattern obtains with progressives as with perfects.

5.17a That Spitz stole the medals is surprising to me.
b That Spitz stole the medals is unsurprising to me.
c That Spitz stole the medals surprises me.
d *That Spitz stole the medals is unsurprising me.
e *That Spitz stole the medals unsurprises me.

Therefore, if a distinction is made between adjectivals and verbals in both the semantic structure and the surface structure the facts fit the 'squish' which Ross proposes and account for the proper insertion of negative *un-*.

Thus far I have presented what I consider convincing arguments that there must be a distinction between adjectival and verbal predicates in the semantic structure and that only those elements marked as adjectival can combine with NEG to give the surface prefix *un-* with the meaning of negation.

I am aware that as early as his dissertation in 1965 Lakoff argued that

> ... adjectives and verbs are members of a single lexical category (which we will call VERB) and that they differ only by a single syntactic feature (which we will call ADJECTIVAL). (Lakoff 1964: A-1)

His arguments for a single category are convincing and I accept them here. However, his analysis is based on a syntactic deep structure and he makes no attempt to deal with participles and the problems presented above. Since that time, of course, Lakoff's view of the deepest structure has changed and much of the analysis presented in this paper is a direct result of that change of view. Although I am not aware of his present position on the use of features to differentiate between true verbs and adjectives, I will argue below that contrary to his 1965 position the distinction between what I am calling adjectivals and verbals cannot be properly made with feature notation. Instead, I will argue that in order to capture the correct generalizations concerning the behavior of participles and prefixes, the pro-verb BE must be postulated in the semantic structure.

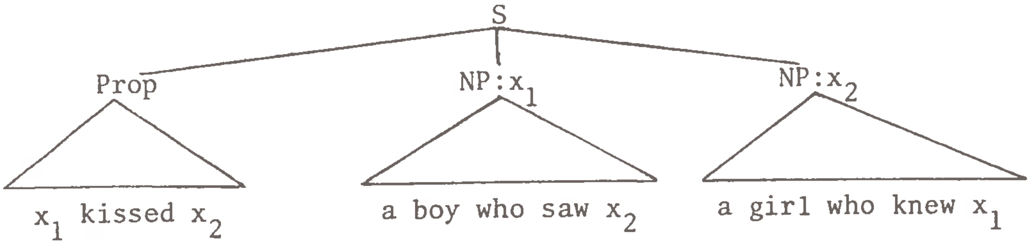
A brief excursus is necessary to provide a background for later arguments. Bach (1968), McCawley (1970), and others have presented arguments that noun phrases do not occur as arguments for propositions, but rather are inserted by transformational rule in the place of indices which do occur as arguments of propositions. This approach accounts for a number of ambiguities and permits the clarification of several reference problems. Perhaps the best example of this latter advantage is in the sentence

5.18 A boy who saw her kissed a girl who knew him.

which McCawley (1970: 176) attributes to Kuno. The argument is succinctly put by McCawley as follows:

> Under the conception of pronominalization which derives a pronoun from a copy of the antecedent noun phrase, this sentence would have to have infinitely many sentences imbedded in it: <u>her</u> would have to come from a copy of <u>a girl who knew a boy who saw her</u>, etc., and both noun phrases would thus have to be derived from infinitely deep piles of relative clauses. (1970: 177)

If the following underlying structure is posited the problem is solved by a relatively simple rule which substitutes the proper semantic information for each index.

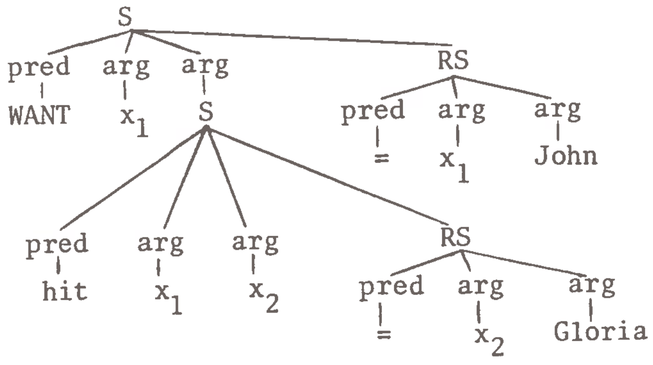


(McCawley 1970: 177)

McCawley does not discuss the structure of the semantic material which replaces the indices nor the transformations which yield the surface forms of the noun phrases. Therefore, since the discussion of participles to follow rests on these particulars I will suggest some conventions which are necessary to a clear presentation of the discussion.

I will treat the replacement structures as sentences in which the topmost verb is a pro-verb which is to be understood as meaning roughly 'is replaced by' or 'is equal to.' I will use the equal sign (=) as an abbreviation for this pro-verb. The replacement structures will be labeled 'RS' and will occur as nodes of the S which immediately dominates the highest occurrence of the index which it replaces. Although McCawley uses the point of attachment of the replacement structures to indicate different readings, I intend that no particular significance be given to where these structures are attached. Structure 5.19 is an example of the formalism just discussed.

5.19



This is the underlying structure for the sentence

5.20 John wants to hit Gloria.

with the understanding that no provision has been made in the tree for tense and that the predicates may not be as simple as indicated.

Now to the problem of characterizing the semantic relationships which must be specified in order to capture generalizations about prefixes, adjectives, and participles. The formalism should relate the following facts: 1) Verbals and adjectivals must be distinguished as separate in the semantic structure, 2) Adjectivals derived from verbals must be shown to bear the derived relationship, 3) The various semantic readings of the prefix form *un-* must be specified unambiguously for the insertion rules. The formalism presented above will provide the proper description with the addition of a pro-verb BE which I contend dominates all adjectives. Given this analysis predicates are not marked in the semantic structure as verbal or adjectival since these semantic functions are assumed by the pro-verb BE in the case of adjectivals and DO, CAUSE, COME, etc. in the case of verbals. This analysis allows the adjectivals derived from verbs (participles) to share the qualities of both and provides unambiguous structures for the proper insertion of the prefix form *un-*.

First some evidence that the BE pro-verb is necessary. Early generativists considered the copula in English to be a surface phenomenon which was inserted by rule just in case there was an adjective as predicate. The copula functioned primarily as an element to carry the inflections normally carried by the verb in sentences with true verbs. However, in the case of participles the copula appears to carry a significant load. In the case of sentence 5.3

5.3 The window is broken

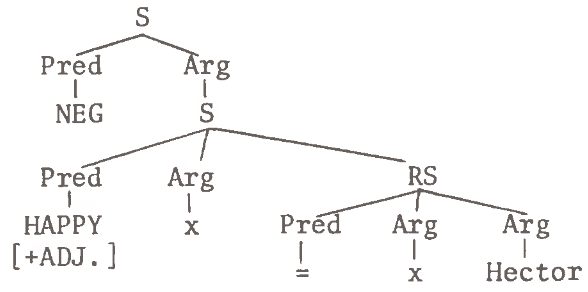
with which I began this discussion the copula is ambiguous in function as mentioned. However, consider the sentences of 5.21, 5.22, 5.23, and 5.24.

5.21a Hector wasn't happy.
b Hector was unhappy.
5.22a Jane wasn't attractive.
b Jane was unattractive.
5.23a Hester wasn't married.
b Hester was unmarried.
5.24a This trip wasn't necessary.
b This trip was unnecessary.

Notice that the sentence pairs of 5.21 and 5.22 are not synonymous while those of 5.23 and 5.24 are. The difference appears to lie in the fact that *happy* and *attractive* are considered to be points on a scale while *married* and *necessary* are conditions which either obtain or do not. Of interest here is the fact that the distinction between the overt negative *not* and the negative prefix *un-* refer to different ranges of the scalar adjectives. Sentence 5.21a allows an interpretation that Hector is anywhere from neutral (relative to happy) to completely negative, sad. On the other hand, Sentence 5.21b allows only the completely negative reading, *i.e.*, 'Hector is sad.' The sentences of 5.22 can be similarly described.[2]

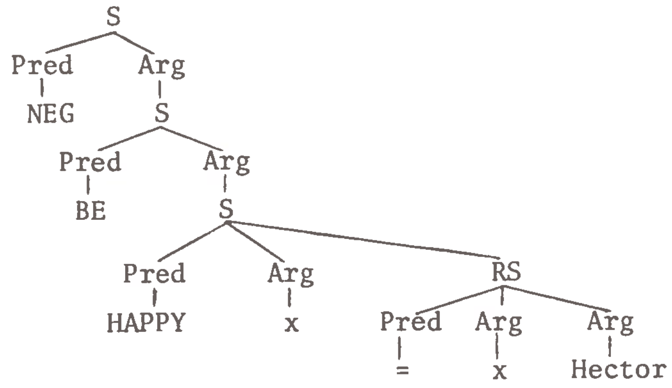
In earlier analyses the structure for 5.21a and 5.21b would have to have been something like 5.25.

5.25

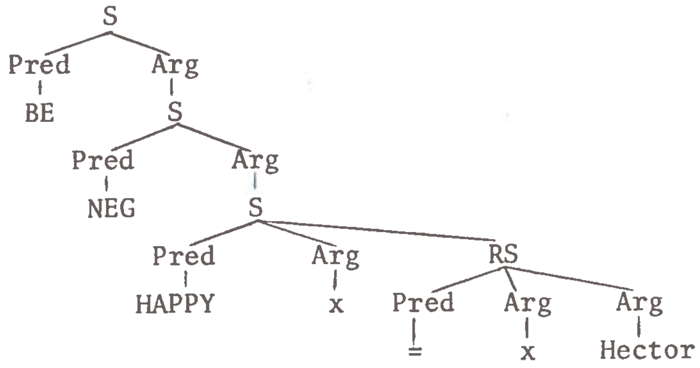


If, however, rather than attaching a feature [+ADJ.] to HAPPY the pro-verb BE is used to indicate its 'adjectivalness' the difference in 5.21a and 5.21b can be characterized as 5.26a and 5.26b respectively.

5.26a



5.26b



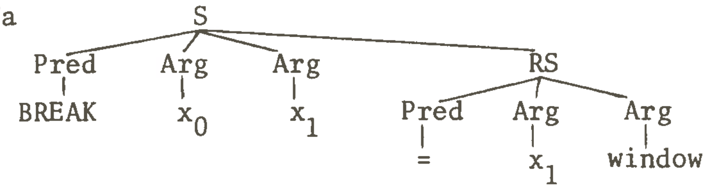
If the pro-verb BE is understood to mean something like 'to have the quality of' or 'be in the condition of' then the two sentences can be paraphrased as follows:

5.21a' Hector doesn't have the quality of being happy.
5.21b' Hector has the quality of being not happy. (sad)

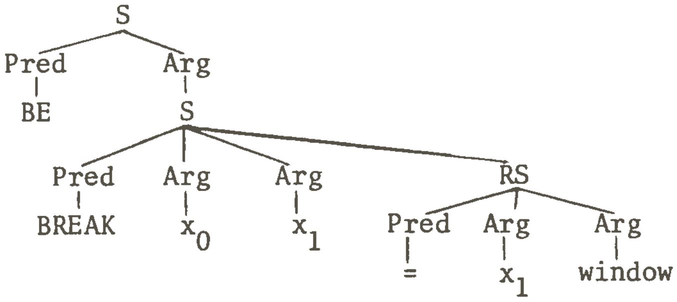
This analysis permits the specification of negation relative to degree adjectives in an intuitively satisfying manner.

More to the point of the topic of participles, this analysis permits both the verbal and adjectival nature of participles to be described. With mechanisms specified above the verbal and adjectival readings of 5.3 can be formalized as 5.27a and 5.27b respectively.[3]

5.27a



5.27b



The structure of 5.27a also underlies the active sentence 5.28 with an indefinite subject.

5.28 Someone broke the window.

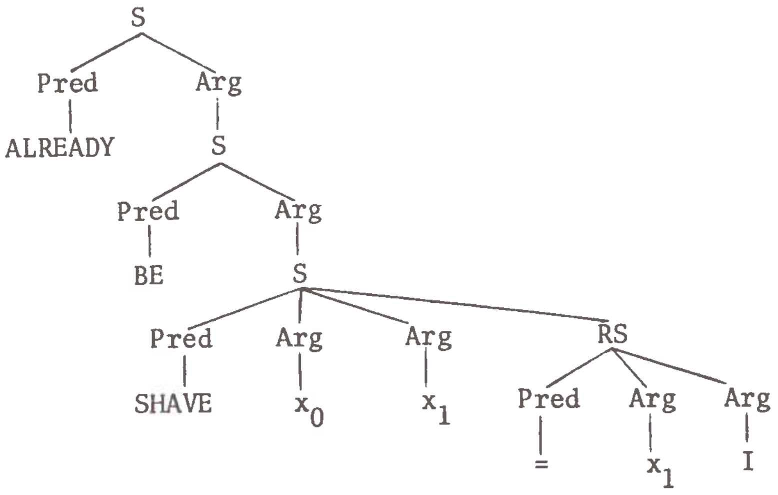
As in earlier treatments, after the passive transformation has applied the agentive argument can be deleted optionally. In the case of the adjectival structure 5.27b, however, the agentive argument must be deleted.

This analysis is further corroborated by sentences like 5.29 and 5.30.

5.29 I am already shaved.
5.30 I have already shaved.

The basic distinction between the two sentences (aside from a difference in aspect) is that 5.29 is adjectival while 5.30 is verbal. It is necessary to assume that there was an agent as the cause of the condition of 5.29. In fact given our present cultural practice of doing our own shaving most people would even identify the agent as the speaker (although I can read the sentence such that a barber did the shaving). However, the agent plays a decidedly secondary role in the semantics of the sentence. The analysis presented here reflects the fact that an agent must be assumed by including an agentive argument as part of the lowest S. On the other hand the secondary position of the agentive argument is recognized by the formalism in that no replacement structure identifies the agentive argument and the lowest S is dominated by the pro-verb BE. The semantic structure for 5.29 looks something like 5.31.

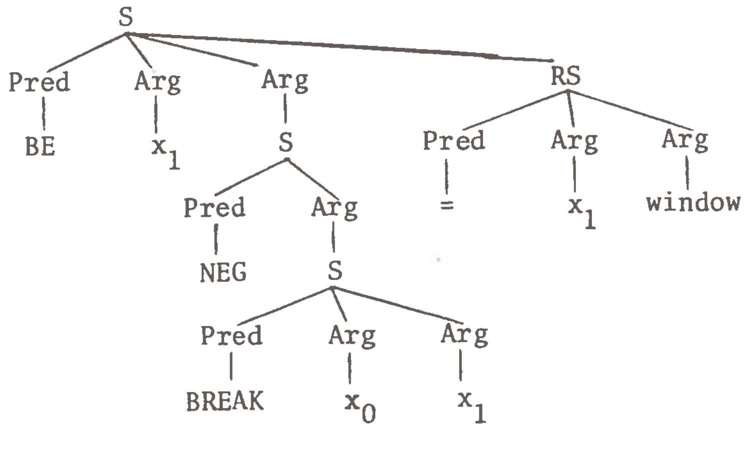
5.31



Having established some rationale for the formalism I will now turn to an analysis of several structures containing negative un-. 5.31b and 5.32b underly 5.31a and 5.32a respectively.

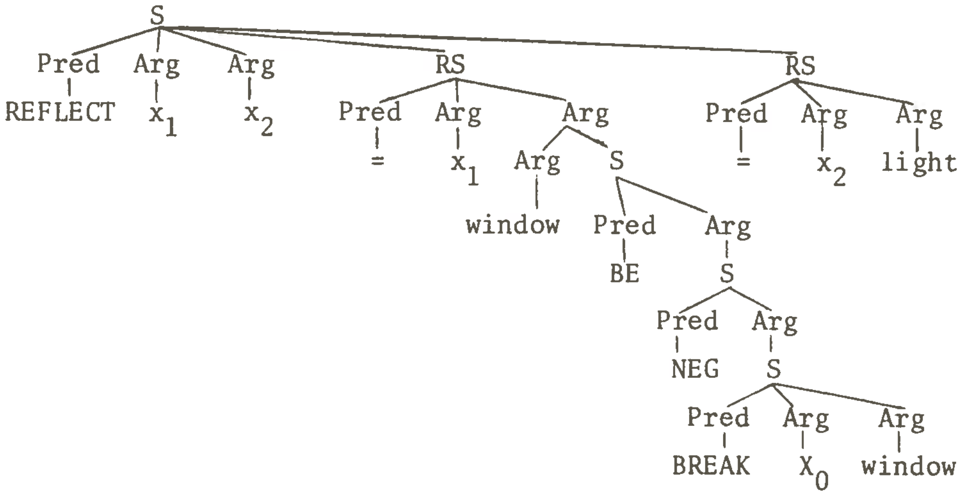
5.31a The window is unbroken.

5.31b



5.32a The unbroken window reflected the light.

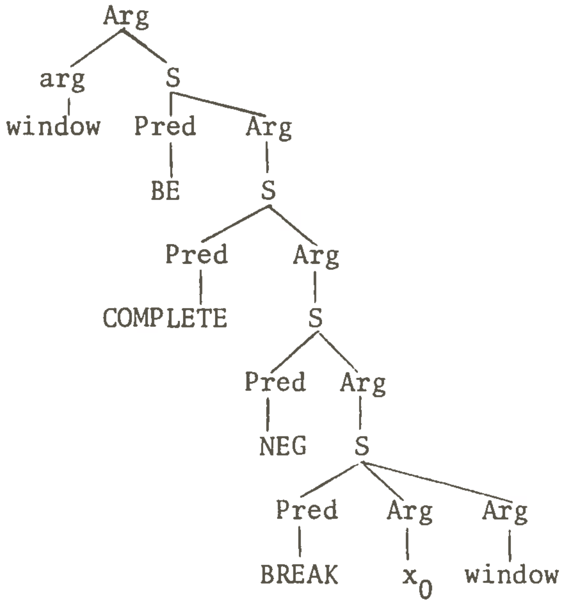
5.32b



Of interest relative to the constraints on negative un- are the noun phrases 5.33 and 5.34 which would occur as the second argument of a replacement structure in a full sentence.

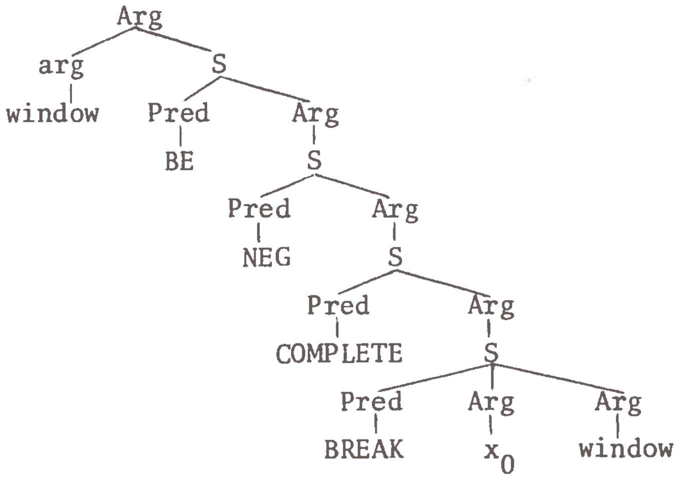
5.33a The completely unbroken window.

5.33b



5.34a The not completely broken window

5.34b



From these structures it is possible to state the constraint on the insertion of negative un- as 5.35.

5.35 Negative un- insertion is obligatory when NEG immediately dominates a predicate to which it is attachable but only when both that predicate and NEG are dominated by BE.

To claim that this constraint is valid is to claim that only adjectivals occur with negative un-. This seems to be the case with the examples I have discussed thus far. However, there are a number of cases in which negative un- occurs which do not obey constraint 5.35 when analyzed in any plausible way.

Consider sentences 5.36, 5.37, and 5.38.

5.36a The construction proceeded undelayed.
b The construction proceeded undelayed by the weather.
5.37 The window remained unbroken throughout the storm.
5.38 The dish lay unbroken on the floor.

The prefixed words in each sentence appear to have an adverbial sense of modifying the verb, e.g., in 5.36a undelayed describes how the construction proceeded rather than the construction itself. Of interest also is the fact that the prefixed participle may have an overt agentive phrase as in 5.36b. These facts argue for the inclusion of certain types of adverbs in the 'squish' mentioned above. When the details of such a continuum are clearer the relationship of these constructions to the adjectival analysis presented here should become clear.

In addition to the restrictions of surface categories on the insertion of prefixes, there are some rather knotty problems involving surface constituent structure, some of which have been discussed in the literature on immediate constituent analysis. What I propose to do here is to present the semantic structures which I consider to underly the surface structure and discuss the difficulties of mapping the former onto the latter.

Look first at noun phrase 5.39 which can have the surface structure of either 5.40 or 5.41.

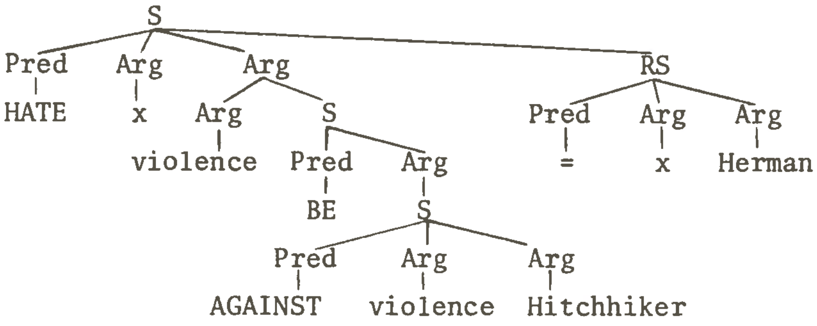
5.39 anti-hitchhiker violence
5.40 (anti-hitchhiker) (violence)
5.41 (anti-) (hitchhiker violence)

The surface structures of 5.40 and 5.41 would be appropriate in sentences 5.42 and 5.43 respectively.

5.42 Herman hates anti-hitchhiker violence.
5.43 Herman is anti-hitchhiker violence.

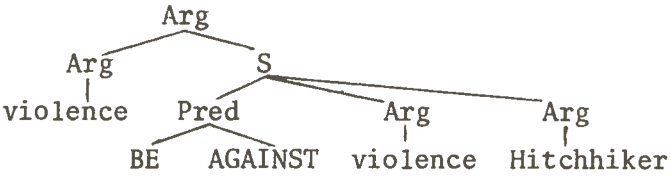
Although there are a number of problems involved in establishing the underlying structure and the transformational rules which will describe these sentences, I will propose the following derivation for sentence 5.42:

5.42a



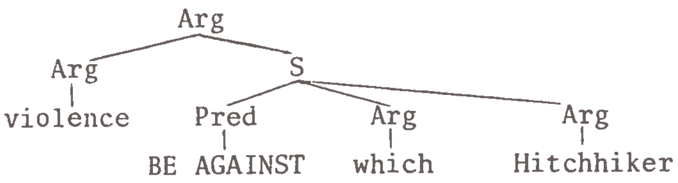
Predicate Raising applies to produce 5.42b.[4] (Only the 'Arg' node in question will be given.)

5.42b



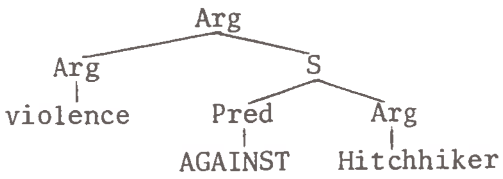
Relative Clause Formation produces 5.42c which would surface in the sentence 'Herman hates violence which is against hitchhikers' if only Subject Fronting and other obligatory rules were applied from this point.

5.42c



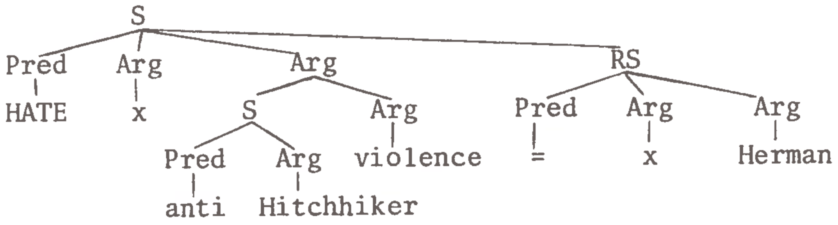
Relative Clause Reduction produces 5.42d which would result in the sentence 'Herman hates violence against hitchhikers' if only Subject Fronting and obligatory rules were applied.

5.42d



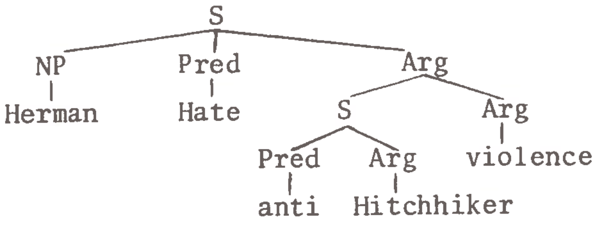
I am not sure of the order of rules at this point. Both preposing and the insertion of the prefix must occur to derive 4.2. If preposing is applied prefix insertion must occur: if it does not apply then insertion must not occur. In any case after preposing and insertion the structure looks like 5.42e.

5.42e



When Index Replacement and Subject Fronting have taken place the surface structure is something like 5.42f.

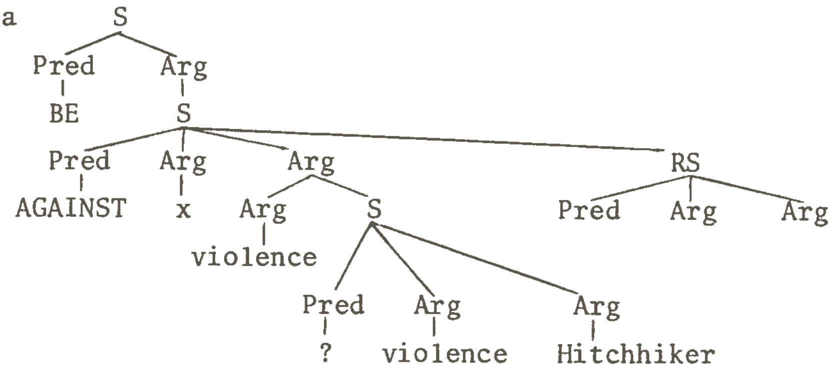
5.42f



Presumably there should be some mechanism which would insert the proper surface category nodes at some point in the derivation. I know of no discussion in the literature of this problem and could offer only an *ad hoc* solution here. I have elected not to do so. Of primary interest here is that rules already suggested by generativists derive the proper surface constituent structure for sentence 5.42.

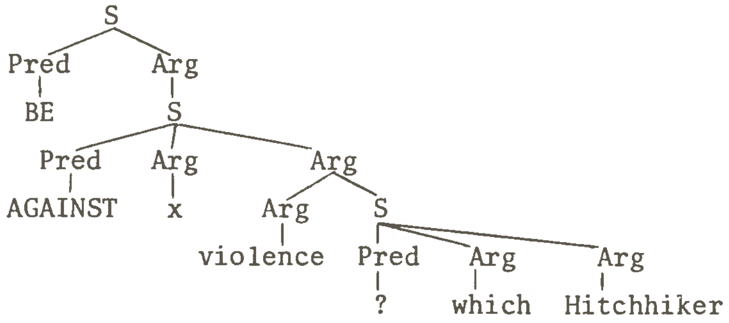
Consider now the derivation of sentence 5.43 from the underlying structure 5.43a.[5]

5.43a



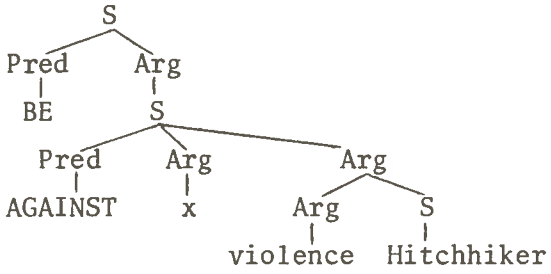
Relative Clause Formation produces structure 5.43b (disregarding the index replacement structure).

5.43b



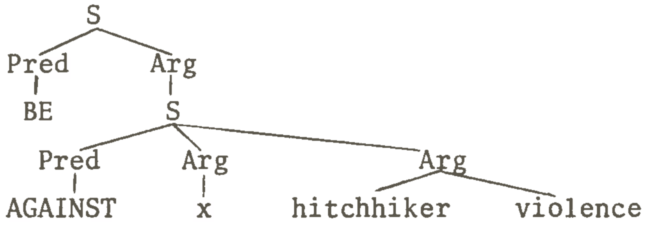
Relative Clause Reduction produces 5.43c.

5.43c



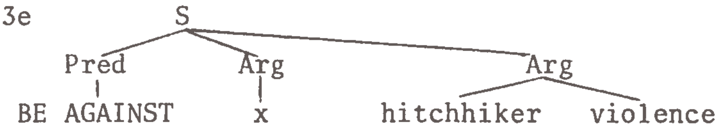
Preposing and Pruning give structure 5.43d.

5.43d



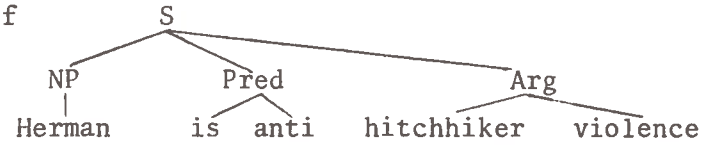
Predicate Raising operates to produce 5.43e.

5.43e



When Lexical Insertion, Index Replacement and Subject Fronting have operated the surface structure is 5.43f.

5.43f



Again as with 5.42 I am not sure how the surface categories are added, but the constituent structure appears to be correct.

The above derivations should be considered quite tentative since the art of mapping semantic structures onto surface structures is new and the processes not too well understood. However, I would argue that the derivations above indicate that such mapping is feasible in the framework given.

CHAPTER VI

PRESUPPOSITION

As linguists ventured into a serious study of natural language semantics it became obvious that they were treading on ground where philosophers had trodden for many years. Only now are philosophers and linguists becoming aware of the rich information each has for the other. Nowhere has this become more clear than in areas of logic such as implication, predication, and presupposition. At this point philosophers have never attempted to generate all the sentences of a language and linguists have not felt responsible for including such semantic nuances in their descriptions of sentences. The result is that not only are the descriptive mechanisms for dealing with presuppositions and implications in a grammar undeveloped, but the goals of linguistics itself have been called into question since one does not want to have to describe the universe in order to describe language. Without attempting to deal with the larger issues involved I would like to present some observations concerning prefixation and presupposition that are revelant to prefixing.

Although there are various ways of looking at presupposition in natural language, it appears that two views have been prevalent: 1) Sentences have presuppositions, and 2) speakers have presuppositions. The first of these views deals with the structures of sentences and the propositions which must be assumed in order to interpret the sentence, i.e., presuppositions which are bound up in the structure of the language. The second view has more to do with the knowledge of the world which must be assumed in order to interpret sentences. It may be that these two classifications are not disjunctive but rather are extremes of a continuum.

In addition to this dimension of presupposition a complete treatment should deal with what Garner calls the 'circumstances of the locutionary act' (1971: 25). In doing so one must take into account the presuppositions of every illocutionary act or object. Only when these most elementary presuppositions are specified can the more particular presuppositions of various types of illocutionary acts and objects, e.g., orders, statements, promises, etc. be explicated. Beyond these specifications the truth value of sentences and the consequences of presupposition failure can be discussed. This brief summary of the problems of presuppositions is based on Garner (1971) which contains a more detailed presentation. The problem of relating the type of information discussed above to the structures with which linguists have been working with are formidable, and the task of correlating the work of philosophers and linguists has just begun.

The presuppositions with which I wish to deal fall into the first classification, an area which has been most discussed by linguists. Keenan defines this type of presupposition as:

> A sentence S logically presupposes a sentence S' just in case S logically implies S' and the negation of S, moreover S, also logically implies S'. In other words, the truth of S' is a necessary condition on the truth or falsity of S. Thus if S' is not true then S can be neither true nor false (and must in the formal logic be assigned a third or 'nonsense' value). (1971: 45, 46)

He gives a number of examples of which 6.1 is representative.

6.1 Fred's driving annoys (doesn't annoy) Mary. (1971: 46)

Notice that with either the positive or negative sentence it is presupposed that 'Fred drives.' If it be the case that Fred doesn't drive then sentence 6.1 is neither true nor false, but nonsense.

The first type of sentences I wish to discuss are closely related to the (by now notorious) sentence, 6.2.

6.2 Have you stopped beating your wife.

The presupposition of this sentence is obviously 'you have been beating your wife.' Either answer in 6.3 makes this presupposition also.

6.3a Yes, I have stopped beating my wife.
b No, I haven't stopped beating my wife.

Consider now the sentences of 6.4 through 6.6.

6.4a Harold didn't submit his manuscript again.
b Did Harold submit his manuscript again?
c Harold, submit your manuscript again!
6.5a Sally didn't register before registration.
b Did Sally register before registration?
c Sally, register before registration!
6.6a Leo didn't edit the book with anyone.
b Did Leo edit the book with anyone?
c Leo, edit the book with someone!

Admittedly some of these sentences sound a bit strange. However, notice that in each of the negative sentences (a) it is not the main proposition which is being negated but rather the adverb or prepositional phrase. For example, in 6.4a it is not being denied that Harold submitted his manuscript, but rather that he submitted it again. The same is true for the question (b) and imperative (c) sentences. In each case the main proposition is presupposed and the syntactically ancillary elements are denied, questioned, or commanded.

Of interest to a study of prefixation is the fact that the prefixed paraphrases have exactly the same presuppositions.

6.7a Harold didn't resubmit his manuscript.
b Did Harold resubmit his manuscript?
c Harold, resubmit your manuscript!
6.8a Sally didn't preregister.
b Did Sally preregister?
c Sally, preregister!
6.9a Leo didn't coedit the book.
b Did Leo coedit the book?
c Leo, coedit the book!

I am not sure how the propositions and presuppositions of these sentences should be related to the semantic structures proposed in previous chapters and by other linguists. However, the fact that the presuppositions for the periphrastic structures of 6.4 through 6.6 and prefixed words of 6.7 and 6.9 are the same argues that the abstract structures underlying both types of sentences should be very similar if not identical.

The second type of sentence which has a bearing on presupposition and prefixation relates to the presupposition of existence. Most philosophers who have dealt with the subject have considered that definite noun phrases carry a presupposition of existence, i.e., when one talks about something he presupposes its existence (at least in the world of the discourse he is engaged in). Linguists have discussed the same facts in dealing with indefinite articles, definite articles, deictic pronouns, and proper names. Keenan gives examples of this kind of presupposition.

6.10a John called (didn't call).
b John wrecked (didn't wreck) this truck. (1971: 46)

In the (a) sentence it is presupposed that John exists and in the (b) sentence the existence of the truck is presupposed whether or not the positive or negative versions of the sentences are taken.

The existence presupposition is further complicated by the scope relationships which exist between noun phrases and other elements of sentences. Karttunen (1969) in discussing discourse referents gives sentence 6.11 as an example of ambiguity based on the scope of the existence quantifier.

6.11 Bill intends to visit a museum every day. (1969: 27)

He gives the following readings.

6.12a 'There is a certain museum that Bill intends to visit every day.'
b 'Bill intends that there be some museum that he visits every day.'
c 'Bill intends to do a museum visit every day.'

In the (a) reading the entire sentence is within the scope of the existence quantifier, in the (b) reading the verb 'intend' is outside the existence quantifier, and in the (c) reading both 'intend' and 'every day' are outside the existence quantifier. This analysis describes the facts adequately. To see that presuppositions are involved with the semantic analysis of 6.11 (at least in the sense in which I have been using the term) it is necessary only to negate the sentence to see that the same ambiguities hold relative to the number and specificity of museums involved.

6.13 Bill doesn't intend to visit a museum every day.

In this sentence the intention is negated, but the propositions involving the existence quantifier are the same as in the positive version of the sentence, thus qualifying as presuppositions under the definition given above. Heretofore the existence predicate has been considered a predicate much the same as other scope involving predicates. It appears, however, that a distinction should be made in these two types of predicates.

Prefixes relate to the scope of the existence quantifier in some interesting ways. Consider the sentences of 6.14.

6.14a Elect a great man again.
b Re-elect a great man.

I find sentence (a) ambiguous in that there may be a specific man (John Doe) whom I am to elect, or this may be an admonition to elect any man just so he is great. On the other hand, I can read the (b) sentence only in the specific sense that there is a particular man who presently holds office, who is considered by the speaker to be a great man, and whom I am to help elect for another term. There is another related ambiguity depending upon whether the prefix re- is considered to mean that I helped elect the candidate the first time or not. This type of ambiguity was discussed above in Chapter IV. Sentence 6.14a can, therefore, have three readings depending on the scope of the existence quantifier and the semantic node AGAIN.

6.15a There exists a great man whom you elected before, do so again.
b There exists a great man who was elected before, see that he is elected again.
c You elected a great man before, again find one and elect him.

In my idiolect only the (a) and (b) readings are possible with 6.14b. Some provision must be made in the grammar to account for this difference in presupposition between the prefixed and periphrastic sentences.

I am not sure how to classify this last type of presupposition which involves prefixes. In all probability it should fall in the second class. Notice the presuppositions of negative questions like the (b) sentence in 6.16.

6.16a Is Joyce coming to the party?
b Isn't Joyce coming to the party?

In the first sentence the speaker is asking a true question with no indication of whether or not he thinks Joyce is coming. In the negative question, however, there is a presupposition that the speaker believed Joyce was coming and would be surprised if she decided not to come. Consider now the sentences of 6.17 which include not only the negation of the question but other negative elements.

6.17a Isn't it unlikely that she will come at 10:00?
b Isn't it not likely that she will come at 10:00?
c Isn't it not unlikely that she will come at 10:00?

In particular notice that the (a) sentence carries a presupposition that she probably won't come at 10:00 while the (b) sentence with the overt negative not has the opposite presupposition. The (c) sentence is grammatical for me but not for others I have talked to. For me it has the same presupposition as the (b) sentence. The difference appears to be similar to some of the cases discussed above involving the scope of different operators. However, in this case there is no apparent operator which can be related to NEG which will adequately describe the difference in presupposition.

While I have presented no final solutions to the problems discussed in this chapter, I have indicated directions which I think research must go in dealing with prefixes and presuppositions. Any grammar which properly analyzes prefixes must take into account the facts of presuppositions presented here, and any grammar dealing with presuppositions must account for the behavior of prefixed words as suggested.

CHAPTER VII

CONCLUSIONS

This paper set out to investigate prefixation in English to discover the type of structures which a grammar must contain if it is to be both formally explicit in its description of the indefinitely large number of sentences possible in a language and to provide either a semantic reading for each sentence or at least to be compatible with a semantic component which provides such a reading. From the preceding chapters I draw the following conclusions concerning the general process of prefixation:

1) Productive prefixation cannot be considered a simple process of inserting elements but rather must be considered more closely akin to the syntactic process.

2) A grammar which does not decompose lexical elements into more primitive semantic elements cannot account for the ambiguities of prefixed words which result from ambiguities of scope relationships.

3) In cases where there is a scope involving element in the surface structure the prefixed semantic element is always interpreted as within the scope of that element.

In the process of capturing generalizations concerning prefixation I have suggested the following changes in the generative semantic theory:

1) A distinction must be made in the semantic structure between propositions of the type used in lexical decomposition and propositions which appear on the surface as overt complements.

2) There must be a distinction between adjectivals and verbals in semantic structure.

3) The adjectival distinction can best be made by positing a pro-verb BE in the semantic structure.

FOOTNOTES

CHAPTER I

[1] By frozen forms I mean words which were historically combinations of elements and have since lost their semantic relationship with the previous combinations, e.g., _prevent_, fr. L. _prae_ 'before' + _venire_ 'to come' and _biscuit_, fr. L. _bis_ 'twice' + _coctus_ 'to cook.'

CHAPTER II

[1] The design for this study was developed jointly by Sam L. Warfel and Herbert R. Harris. The results of the study were presented in two separate unpublished papers. Mr. Harris' dealt with suffixes and mine with prefixes. The prefix paper is included with this work as Appendix A.

[2] Not-Transportation is the rule which operates on structures involving a small number of verbs which take sentential complements and which allow _not_ to be moved from the complement sentence to which it logically belongs to the matrix sentence. For example, sentence (A) would result if Not-Trnasportation were not applied and sentence (B) results from the application of the rule.

A I think I can't come.
B I don't think I can come.

[3] Some of the material presented here and throughout the paper appears in my article "Toward a theory of prefixing.' (Warfel 1971)

[4] The term 'phonology' is used here to refer only to the rough phonemic or even graphemic form of the words.

CHAPTER III

[1] It is difficult to specify the time reference for large numbers of words with _pre-_. For example, in sentence (A)

A He pretuned the receiver in the shop.

it is obvious that the tuning was before some time, yet it is difficult to specify what that time was. The reference problem is further complicated by the fact that prepositional phrases which indicate time can occur in sentence with _pre-_ words. Even phrases with the seemingly redundant _before_ seem quite natural.

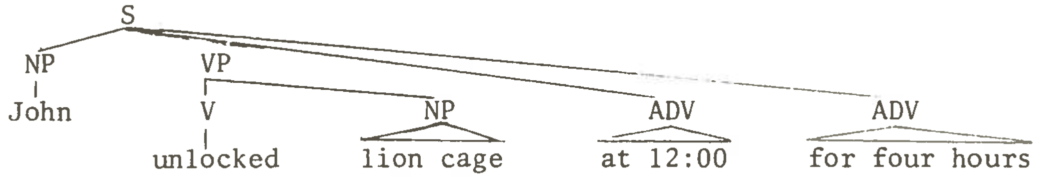
B He pretuned the receiver before it left the shop.

A complete analysis of _pre-_ which encompasses these issues will probably involve a number of presuppositions involving the relationship between the speech act and time statements.

[2] The term 'causative' is used in its obvious sense. The term 'inchoative,' on the other hand, refers to the semantic notion of 'becoming' or 'coming into being.'

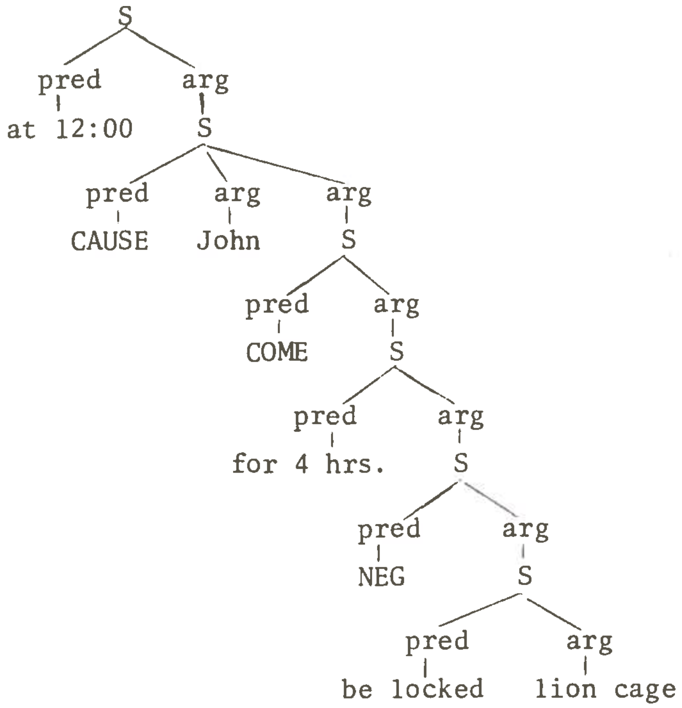
[3] In a grammar without lexical decomposition the first sentense of the compound 3.44 would look something like (A).

A



It may be that the ADV nodes should be attached to the S node. In any case there is no constituent to which the it of the second sentence of the compound can refer with the correct meaning. The pronoun must refer to the entire sentence in this analysis. However, if the structure of (B) is posited as underlying then there is a constituent to which the pronoun can refer with the meaning that the continuing openness of the cage was the source of the uneasiness.

B



[4] Although the analysis of re- presented here is original with me, Robert Binnick informs me that Jerry Morgan has presented a similar analysis in an unpublished paper the title of which I do not have.

CHAPTER IV

[1] This problem is best illustrated by the classic examples

A All the people in this room speak two languages.
B Two languages are spoken by all the people in this room.

in which the active and passive sentences are not synonymous. In the (B) sentence the same two languages are spoken by all the people. The (A) sentence, on the other hand, does not claim that all the people speak the same language.

CHAPTER V

[1] I use the term 'intensive' since it has been used traditionally. However, I do not feel that there is any intensification and the prefixed and non-prefixed forms are synonymous for me.

[2] For a discussion of prefixes and scalar adjectives see Zimmer 1964.

[3] I am using x_0 to represent an index which has no replacement structure. In cases where an unspecified noun phrase is necessary on the surface (as in 5.28) the index will be replaced by an indefinite pronoun.

[4] The rules used here are discussed in detail in McCawley 1968 and Green 1972.

[5] I am not at all sure what the lowest predicate should be so I have indicated it with a question mark. In all probability that part of the structure is considerably more complicated than indicated, involving further decomposition.

BIBLIOGRAPHY

Bach, Emmon. 1968. Nouns and noun phrases. Universals in linguistic theory, ed. by Bach, Emmon & Robert Harms, 90-122. New York: Holt, Rinehart and Winston, Inc.

Binnick, Robert. 1969. Studies in the derivation of predicate structures. PhD Dissertation. University of Chicago.

Bloomfield, Leonard. 1933. Language. New York: Holt, Rinehart and Winston, Inc.

Burt, Marina. 1971. From deep to surface structure: an introduction to transformation syntax. New York: Harper & Row.

Carden, Guy. 1967. Quantifiers as higher verbs. IBM Boston Programming Center, technical report BPC 5.

__________. 1970. A note on conflicting idiolects. Linguistic Inquiry. 1:281-90.

Chapin, Paul G. 1967. On the syntax of word-derivation in English. (Information Systems Language Studies, No. 16.) Bedford, Mass.: MITRE Corporation.

Chomsky, Noam. 1957. Syntactic structures. (Janua linguarum, 4.) The Hague: Mouton.

__________. 1965. Aspects of the theory of syntax. Cambridge: M.I.T. Press.

__________, and Morris Halle. 1968. The sound pattern of English. New York: Harper and Row.

Fraser, Bruce. 1971. An analysis of 'even' in English. Studies in linguistic semantics, ed. by Fillmore, Charles & Terence Langendoen, 150-78. New York: Holt, Rinehart and Winston, Inc.

Gleason, Jean Berko. 1958. The child's learning of English morphology. Word. 14:150-77.

Green, Georgia. 1972. Some observations on the syntax and semantics of instrumental verbs. Papers from the Eighth Regional Meeting Chicago Linguistic Society, 83-97. Chicago: Chicago Linguistic Society.

Gruber, Jeffrey C. 1965. Studies in lexical relations. Unpublished Ph.D. dissertation, M.I.T.

Harris, Zellig S. 1951. Structural linguistics. Chicago: The University of Chicago Press.

Heringer, James. 1970. Research on quantifier-negative idiolects. Papers from the Sixth Regional Meeting Chicago Linguistic Society, ed. by Campbell, Mary Ann et al, 287-96. Chicago: Chicago Linguistic Society.

Hill, Archibald S. 1958. Introduction to linguistic structures. New York: Harcourt, Brace, & World.

Jakendoff, Ray. 1969. An interpretive theory of negation. Foundations of language. 5:218-41.

Jespersen, Otto. 1942. A modern English grammar on historical principles. London: Bradford & Dickens.

Keyser, S. J. 1968. Review of Sven Jacobson, Adverbial positions in English. Language. 44:357-74.

Klima, Edward. 1964. Negation in English. The structure of language, ed. by Fodor, Jerry & Jerrold Katz, 246-323. Englewood Cliffs: Prentice-Hall, Inc.

Koziol, Herbert. 1937. Handbuch der englischen Wortbildungslehre. Heidelberg: Carl Winter's Universitatsbuchhandlung.

Labov, William. 1972. For an end to the uncontrolled use of linguistic intuitions. Paper presented to the 47th Annual Meeting of the Linguistic Society of America, December 27, 1972.

Lakoff, George. 1965. On the nature of syntactic irregularity. Computational Laboratory of Harvard University, Report No. NSF 16 to the National Science Foundation, Anthony G. Oettinger, Principal Investigator, Cambridge, Mass.

__________. 1970. Pronominalization, negation and the analysis of adverbs, in Jacobs, Roderick and Peter Rosembaum. Readings in English transformational grammar. Waltham, Mass.: Ginn and Co.

__________. 1970. Repartee, or a reply to 'Negation, conjunction and quantifiers.' Foundations of language. 6:389-422.

Lees, Robert B. 1960. The grammar of English nominalizations. (Indiana University Research Center in Anthropology, Folklore, and Linguistics, Publication 12.) Bloomington, Indiana.

Merchand, Hans. 1969. The categories and types of present-day English word-formation. 2nd ed. Munchen: C. H. Beck'sche Verlagsbuchhandlung.

McCawley, James. 1968. Lexical insertion in a transformational grammar without deep structure. Papers from the Fourth Regional Meeting Chicago Linguistic Society, ed. by Darden, Bill et al, 71-80. Chicago: Department of Linguistics, University of Chicago.

__________. 1970. Where do noun phrases come from? Readings in English transformational grammar, ed. by Jacobs, Roderick & Peter Rosenbaum, 166-83. Waltham, Mass.: Ginn & Co.

Partee, Barbara Hall. 1970. Negation, conjunction, and quantifiers: syntax vs. semantics. Foundations of language. 6:153-65.

Ross, John Robert. 1972. The category squish: endstation hauptwort. Papers from the Eighth Regional Meeting Chicago Linguistic Society, 316-28. Chicago: Chicago Linguistic Society.

Sapir, Edward. 1921. Language, New York: Harcourt, Brace, & World.

Warfel, Sam L. 1971. Toward a theory of prefixing. Automated analysis of language style and structure in technical and other documents: Technical Report #1, Sally Y. Sedelow, Principal Investigator, 76-97. Lawrence, Kansas: University of Kansas.

__________, & Herbert Harris. Unpublished study of the acquisition of prefix and suffix rules in English.

Zimmer, K. 1964. Affixal negation in English and other languages. Monograph No. 5, Supplement to Word 20.

APPENDIX A

PROLOGUE

The following is part of a report on an experiment which Herb Harris and I conducted dealing with prefixes and suffixes relative to the acquisition of English. Mr. Harris is reporting on the suffixes and I will report on the prefixes. I will deal with the general design of the experiment, but will report only on the problems which relate to prefixation.

INTRODUCTION

This study is an attempt to answer several questions about the English language as it is structured in the mind of speakers of the language and about the acquisition of this structure. In particular the study focuses on two prefixes and two suffixes: re- (meaning "again"), un- (meaning "reversal"), -y (derives adjectives from nouns), and -able (derives adjectives from verbs). The basic hypothesis is that these affixes are rule governed, that is that words with these affixes are not stored in the lexicon of the speaker with the affix but combined with them by rule. The present study tests this hypothesis with children and attempts to determine at what point in the acquisition of the language these rules are acquired. The study is thus an extension of the general approach to morphological problems of acquisition well illustrated by Jean Berko Gleason (1958). We have incorporated her technique of using a questionnaire with pictures and nonsense words as well as her view that derivational as well as inflectional morphology is rule governed. However, we have defined more precisely what it means to acquire an affix and have devised types of questions which seek to make finer distinctions in this area.

THE QUESTIONNAIRE

The questionnaire devised includes questions which require the child to process words in three ways: 1) segment a word into its morphological constituents, 2) produce a word with the addition of an affix consistent with the semantic context, and 3) explain an affixed word semantically. Berko's test required only the second process, which in effect defines for her what it means to acquire an affix. We feel that this definition is too gross and can be refined by considering what a child is able to do with the processes listed above. We will speak then of acquisition with regard to the three processes above as segmentation, production, and comprehension respectively.

The questionnaire consists of 24 questions of which 12 use real words and 12 nonsense words. Each of the four affixes is used in one sentence of each of the three process types with both real and nonsense words. (The Questionnaire is included as Appendix aA.) The real words chosen are words which we were relatively certain were familiar to even the youngest of the subjects to be tested. The semantic circumstances were established through one or two sentences and an appropriate frame was given into which the child was to insert the processed word. No pictures were used on this section of the questionnaire. One major mistake was made in writing the real word questions. Question 9, which was devised to elicit semantic information about the *un-* prefix, is ill formed in that it means 'negation' and not 'reversal.' This flaw seriously damages the data for this prefix, but does provide some interesting evidence (albeit indirectly) as will be mentioned later.

The remaining questions were constructed with one or two nonsense words which follow the morpheme structure and the syntactic-morphological rules of the language. There are two levels of nonsense for the sentences depending upon the number of nonsense words used. For example, Question 13 has the real word *cat* and only the word *wug* is unknown. However, in Question 15 both the verb *glot* and the object noun *whizzle* are unknown. The semantic content of the nonsense words was filled partially by the use of pictures of identifiable objects. (See Appendix bA.) In most cases the nonsense words were translatable into real words which would describe the object and the action. For example, in Question 15 *newspaper* can be substituted for *whizzle* and *fold* for *glot* and the question still allows the insertion of the proper prefix. On the other hand, in Question 16 while *puzzle* can be substituted for *zigger*, there is no verb equivalent of *tring* as something you do to puzzles, at least no verb from which an adjective is derivable with *-able*. However, as a subjective judgment from administering the test there appeared to be no translating by the subjects who for the most part either understood the 'game' of manipulating the nonsense words or ignored the nonsense words and answered with real words.

ADMINISTRATION

The test was administered on Friday and Monday, December 5 and 8, 1971 at the Blackburn Elementary School in Independence, Missouri. A total of 55 students were tested: 5 from the first grade and 10 each from grades 2 through 6. The selection was not entirely random. We requested children from the average sections in each grade and we have no way of knowing what criteria the teachers might have used to select the specific children which they sent to us. The test required from 5 to 12 minutes depending on the age and ability of the child. Three-fourths of the test was conducted in a separate room which was quiet and pleasant. However, one afternoon the room was in use by the school and the only available space was a storage room under a stairway next to the physical education area. These examinations were conducted among the boxes of the storage room with considerable noise and distractive movement.

Each subject was seated across a small table from the questioner with a microphone on a stand in front of him. The observer was seated behind and to one side of the questioner. The questions and responses were tape recorded and the responses were also transcribed by the observer along with any additional comments he deemed important. The questioner usually introduced himself and the observer who then got the child's name, age, grade, and sex. The questioner then began reading the introduction and then the questions. If the child supplied the wrong answer or failed to answer the questioner repeated the question. If a second response was also wrong the prompts listed on the Questionnaire in Appendix aA were given. I was the questioner for grades 1, 2, 4, and 6 and Mr. Harris questioned the subjects from grades 3 and 5. After all of the tests were completed Mr. Harris and I listened to the tape recording of the tests, compared the responses with those transcribed by the observer, and made decisions as to the correctness of the responses.

RESULTS

The results of the tests for the two prefixes are presented in Table 1, Appendix cA. The figures are given in percentages of correct responses by grade. Thus, ten percentage points in a given column represent one response for grades 2 through 6 and twenty percentage points equals one response for grade 1. It is, therefore, obvious that the interpretation of one response or various non-linguistic factors affecting two or three children can make a considerable difference in the results. The significance of the data must be judged accordingly.

Questions 1, 5, and 9 deal with un- and real words; Questions 15, 18, and 22 deal with un- and nonsense words. The results of Question 9, however, must be ignored (although I have placed the results in the table and on the charts) because of the semantic difference mentioned above. The table shows that all but one of the children tested were able both to segment and to produce the prefix with real words. With the nonsense words the results are quite different. A problem of interpretation arises with Question 15. The answer we were looking for was unglot. However, in 12 cases the un- was attached to the wrong nonsense word and the answer was unwhizzle. These answers are distributed quite evenly over grades 1 through 5 as is shown in the difference between the blue and black lines on Chart I of Appendix cA. This problem was not encountered with other nonsense words which seems to rule out the explanation that it is too difficult to hold the two words in mind for the time required. The only explanation I can think of is that the word *whizzle* has high salience value as a 'funny' word. In any case, it seems that the answer should be considered correct since the prefix was correctly added even if it is on the wrong word.

Both Charts I and II indicate that there is a rather even increase in the ability to manipulate the nonsense words with *un-*. Chart III shows the randomness of the responses to Question 22. This randomness is partially the result of the vagueness of the question and the resulting difficulty with interpretation of the answers. It was very difficult to set criteria for judging an answer as correct. We had hoped for answers such as 'He didn't want it claggified' or 'He didn't want it that way.' We did get such answers, but most were not that simple. Often we got something like 'He did it wrong' or 'It needed to be unclaggified' both of which we counted as incorrect. Interestingly, often the same answer was given for Questions 21 and 22 which may say something about the semantic relatedness of *un-* and *re-*. I plan to work on this problem at a later time.

The questions relating to *re-* are Questions 2, 6, and 10 with real words and Questions 14, 17 and 21 with nonsense words. Table 1 and Charts IV and V show the disparity between the correctness of responses to Questions 2 and 6. In particular, it is not until the 5th grade that anyone uses *re-* in *reheat*. This is probably due in part to the difficulty of writing a question which will force the use of the prefix to the exclusion of a paraphrastic construction with the adverb *again*. However, even with the prompt 'Can you think of another way to say "heat it again"' there was no result. Chart VI shows the responses to the semantic question relative to *recount*. We were quite surprised to find that few subjects could correctly answer this question until the fourth grade level. Our criteria on this question were that the answer include the words *again* or *over* or that it demonstrate in some other way that *recount* refers to a second counting.

The results with nonsense words are displayed in Charts IV, V, and VI. Again, as with *un-*, there was considerable trouble deciding on correct answers for Question 21, the semantic comprehension question. We expected answers such as 'Because she did it wrong' or 'Because she wanted to do it over.' However, we also got answers such as 'To fix it better,' 'It got tore up probably,' and 'Cause it broke' all of which we accepted. Most of the incorrect answers were simply no response, since the question has so little semantic content that the children seemed hard put for any answer at all. As with *un-* there is considerable randomness of correct answers and little can be learned from the results. This information might be of use in looking at the patterns which may emerge when individuals are considered. It might be the case that some children can answer this question when they cannot segment re- in Question 17. However, a cursory examination of several cases where this is the case indicates that the answers for the semantic question are doubtful or that the child did not learn the task on the previous nonsense questions.

CONCLUSIONS

We determined that there are several design faults in the experiment. The most serious flaw seems to be the failure to catch the different semantic meaning of un- in Question 9. We also failed to control several variables which may have prejudiced the data. We alternated as questioners and each person had his own style of presentation as to stress of syllables, when to press for another answer, how long to wait before prompting, and even the exact nature of the prompt. We both found it very difficult to do each examination exactly like the others and tended to interact with the subject. Other more subtle factors such as time of day and even day of the week may have had some effect on the alertness of the subjects. In all, however, I have confidence that the factors were sufficiently controlled as to give reliable data within the limits of the size of the sample.

I think that we have shown that our original hypothesis is correct. Children do not simply store the words which grammarians consider to have prefixes; they learn what is best described as rules for the combination of morphemes. If prefixed words were simply stored in the lexicon there would be no ability to manipulate the prefixes with nonsense words which we have demonstrated that the older children have developed the ability to do.

There is a noticeable difference between the ability to deal with real words and nonsense words. It may be the case that the cognitive abilities of children have not sufficiently developed to deal with the task of manipulating nonsense words at the point at which they can manipulate the real words. This raises a question as to what is required both cognitively and linguistically in performing the task we have established with the nonsense words. It was clear that some children did not grasp what was expected of them or did not have the ability to perform the task. Several children answered none of the nonsense questions correctly, but seemed determined to give a correct semantic answer based on the pictures. This leads us to conjecture that a training period with some other affix might have been of help to those who did not understand the task. However, I contend that while it may be that general learning ability is necessary to learn what is expected in the task, a child cannot do the task without having the linguistic competence to perform it. We may have missed some children who have the competence, but who could not grasp the task, but we did not have any children who learned the task without the competence. Part of learning the task is to know that *un-* and *re-* are separate morphemes which can be added to or removed *from* stems. Without this knowledge the task cannot be completed even if the child understands that he is to change the nonsense words in some fashion. This seems to be substantiated by Berko's study where children younger than any we tested learned essentially the same task with little difficulty. Since our youngest children had difficulties with the nonsense words, this implies that the affixes we were working with are acquired at a later period of acquisition than those tested by Berko.

It remains to posit some explanation for the considerable difference between the ability to produce and segment real words with prefixes and the ability to do the same with nonsense words. This can best be explained by looking at morphology in general linguistic theory. Morphemes exist at two levels in a grammar. On one level they are composed of phonological information and thus are recognizable by their form. On the other level they contain semantic information. Most grammarians analyze a word like *untie* as consisting of two morphemes: *un-* meaning 'reversal' and *tie* meaning roughly 'to fasten together by entwing.' The morphological analysis is possible because of all the other words which appear in the language with *un-* and in that form also mean 'reversal' plus the meaning of the word without the *un-*. The problem is that other words seem to be composites of semantic elements also, but the elements are not divisible as phonological units. For example, the relationship between *build* and *destroy* seems to be analogous to that of *tie* and *untie*. However, there is no morphological relationship between the first two since there is not even one common phonological segment.

Relating this to child acquisition of prefix rules it appears quite possible that the word *untie* is learned first as a unit, just as *destroy* is learned; both have the semantic prime element 'reversal' plus the notion of 'tie' on one hand and 'build' on the other. It is only later when he has learned enough words with the phonological-semantic bundle *un-* that he is able to formulate a rule which allows him to comprehend and produce new forms with the prefix. It thus seems plausible that the children in our experiment were not processing the real words in the early grades but recalling semantically appropriate lexical items.

A careful look at Question 2 on the questionnaire will show that the context is sufficient to allow the correct answer to be guessed without even the morphologically related word present. What you do to stories is read them. Again Question 5 can be answered without the first sentence. What you do to your shoe to keep it from coming off is to tie it. Question 1 also can be answered in similar fashion. In each case the answer is strongly suggested by the semantic context and does not require any morphological analysis if the lexical items are known. It appears that reheat is not a familiar word and thus does not suggest itself to the child in Question 6, especially when other words and constructions are available to fit the context.

The semantic context is so strong with the younger children that it appears that they may not even hear the prefixes. It is interesting to note that 5 of the children answered Question 9 to the effect that there might be germs or dirt in the bottle which indicates that the prefix un- (although the wrong un- for our purposes) was simply not comprehended and the word was interpreted as opened.

I am suggesting therefore, that the disparity between the ability to manipulate real words and nonsense words may be more apparent than real. The child may be simply recalling lexical items until the 4th or 5th grade when he begins to formulate prefixing rules. Thus, he cannot process the nonsense words because they are not in his lexicon to be recalled and he has not as yet formulated the necessary rules to make the phonological and semantic connections.

The data also suggest that un- as a productive prefix is acquired before re-. This can be seen best by looking at the charts for segmentation and production relevant to each prefix. With nonsense words un- is correctly used by 60% to 70% in the 3rd grade while re- is used correctly by only 0% to 20% in the 3rd grade. If my assessment of how prefix rules are acquired is true, this is probably attributable to a fewer number of re- words in the vocabulary of younger children. This seems to be substantiated by Chart VI where the real word recount does not seem to be understood well until the 5th grade.

It is also possible to argue from the data (recognizing how tenuous such a reliance is on a sample of this size) that the acquisition of the re- prefix rule takes place in stages in the following order: comprehension, segmentation, production. Looking at Chart VII which puts together the information on nonsense words from the segmentation and production charts (IV and V) with the information on real words from the comprehension chart (VI) it appears that this order is proper. Notice that there is a jump from 40% to 90% at the 4th grade level with the comprehension of the real word recount. The segmentation line jumps from 20% to 50% in the 4th grade and from 50% to 90% in the 5th grade with nonsense words. The production line moves little until the 6th grade when it jumps from 10% to 70%. This at least hints that the acquisition of the prefix is not something which makes the rule available to both the encoding and decoding components of the child's grammar at the same time.

A look at Chart VIII will show that data on the acquisition of un- does not show as clear a pattern. In fact if the unwhizzle answer is considered correct the order of acquisition appears to be production followed by segmentation (remember that the use of the real word comprehension question is invalid for *un-*). If the *unwhizzle* answer is considered incorrect, the order follows that given above for *re-*.

From this experiment three conclusions can be drawn with varying degrees of certainty: 1) children do develop rules for the addition, segmentation, and comprehension of prefixes, 2) The prefix *un-* precedes *re-* in order of acquisition, and 3) The acquisition of a prefixation rule is not a single process but involves stages of comprehension, segmentation, and production.

APPENDIX aA

Introduction

This test has two parts and will take about five minutes. In this part we will ask you to give us a word or to explain something. Some of the questions will seem very easy. This is because these questions are also being given to very young children. There is nothing tricky about the questions. Just give the obvious answer.

1. When you go outside, you button up your coat. When you come inside again, what do you have to do to your coat?

   PMT If the answer was 'Hang it up' or something similar the prompt was 'What do you do to the buttons?'

2. George reread a story. He did it because he couldn't remember it the first time that he __________ it.

3. My backyard is bushy. What are there a lot of in my backyard?

4. A baby rabbit is lovable. Jane wants one so that she can __________ it.

5. A boy's shoe is untied. What will he have to do to his shoe to keep it from coming off?

6. Mother heated the soup, but now it is cold. In order to get it ready to eat what will she have to do to it?

7. Jane has come in from outside with dirt all over her. Her mother asked her how she got so ___________?

8. Joan's mother bought some vitamins that can be chewed. These are called what kind of vitamins?

9. Why can't you drink an unopened bottle of Coke?

10. What would you do if you recounted a sack of nickols?

    PMT Why would you recount a sack of nickels?

11. If a bed is sandy, what does it have in it?

12. If something is drinkable, what can one do with it?

APPENDIX aA 2

In the following questions there are some words that you probably have never heard before. We are going to ask you to change them around some. Don't worry if you don't know what they mean. It isn't necessary to know their meanings in order to answer the question. In some cases there is more than one answer possible. If you think of more than one, give all the answers you think of. The answers, however, always have some part of the new words in them.

13. This cat has wugs [wə́gz] on his head, wugs on his body, and wugs on his feet. He has wugs all over him. He is a very __________ cat.

14. PMT How would you describe him?

14. This man has kented his boat. He didn't do a very good job, so he is going to have to kent it again. He is going to have to __________ the boat.

PMT How else can you say he is going to have to kent it again?

15. When this man got home he found that his kids had glotted [glátɨd] his whizzle [wɪ́zəl] all up. He doesn't want his whizzle glotted so he is going to have to __________ it.

PMT He is going to make the whizzle back like it was.

16. This zigger [zɪ́gər] is easy to tring [trɪŋ]. Even little children can tring this zigger. It is a very __________ zigger.

PMT This zigger is very __________.

17. Sally is reknopping [rinápɪŋ] her kline [klayn] because the teacher said she didn't do it right yesterday. John wasn't here yesterday so he is just starting to __________ his kline.

PMT If Sally is reknopping her kline what is John going to do?

18. George unbicked John's simlach, but John didn't want his simlach [sɪ́mlæk] or [sɪ́mlæč] unbicked [ənbɪ́kt]. He told George that he wanted it __________ back again.

PMT If George unbicked it what would you have to do to get it back like it was before?

19. If this dun [dʊn] or [dən] is very tamable [tǽməbəl], then it is easy to __________ this dun.

PMT What could you do to a dun that is tamable?

APPENDIX aA 3

20. This yap [yæp] is droonie [drúni]. He has ____________ all over him.

PMT What does he have all over him?

21. Mary has refigged [rífɪgd] her niz [nɪz]. Why do you suppose she had to refig her niz?

PMT What could refig mean?

22. Clarence unclaggifies [ənklǽgɪfayz] his radle [rédəl]. Why do you suppose he unclaggifies his radle?

PMT What could unclaggify mean?

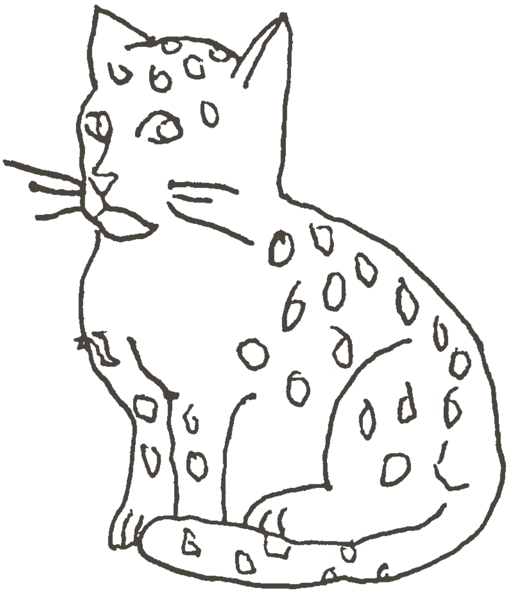
23. If a darf [darf] is drickable [dríkəbəl] what can be done to it?

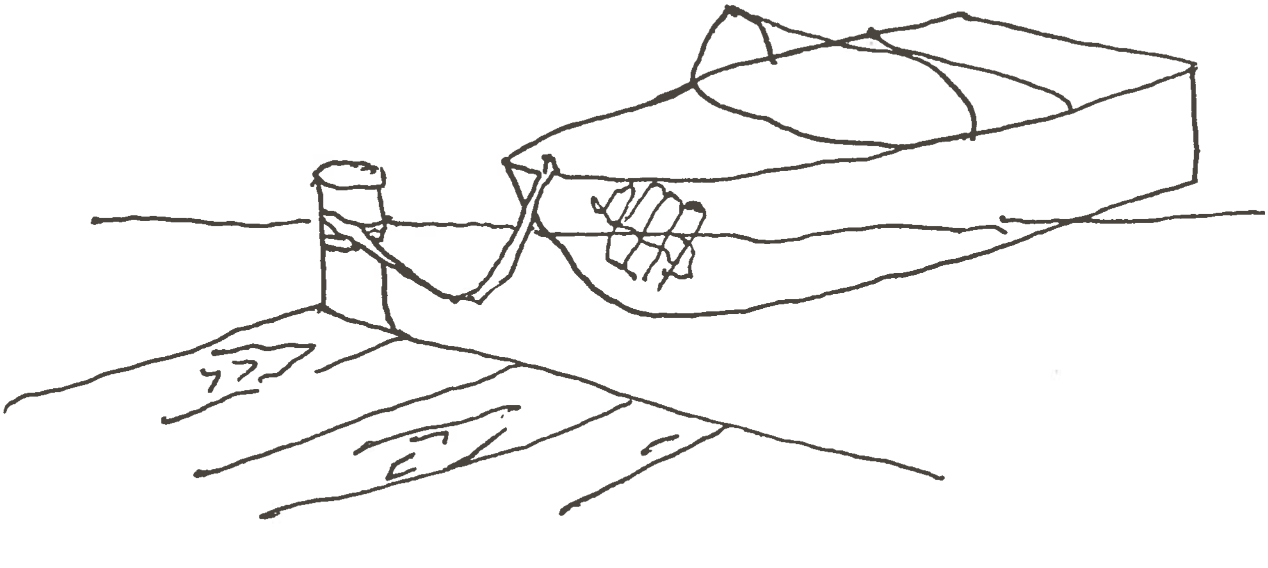
PMT What could drickable mean?

24. What do you think a toopie [túpi] dog would have all over him?

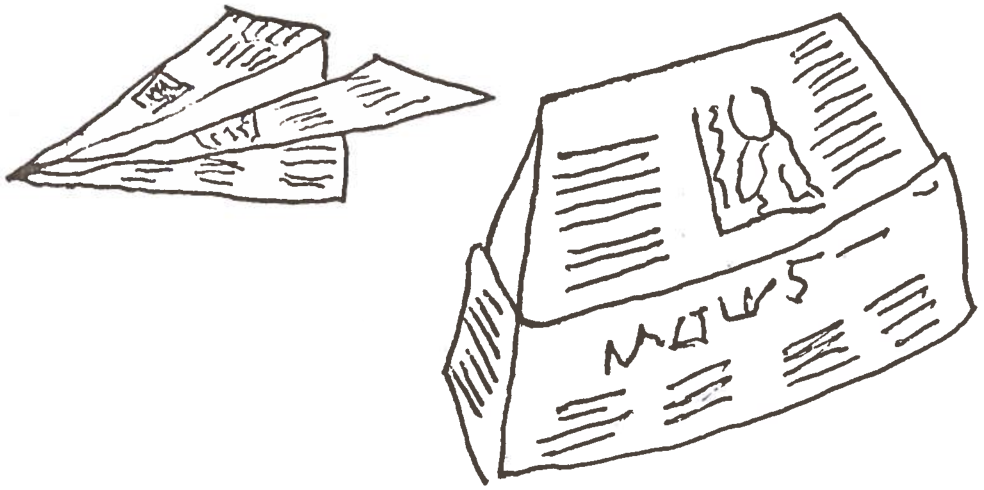
PMT What could toopie mean?

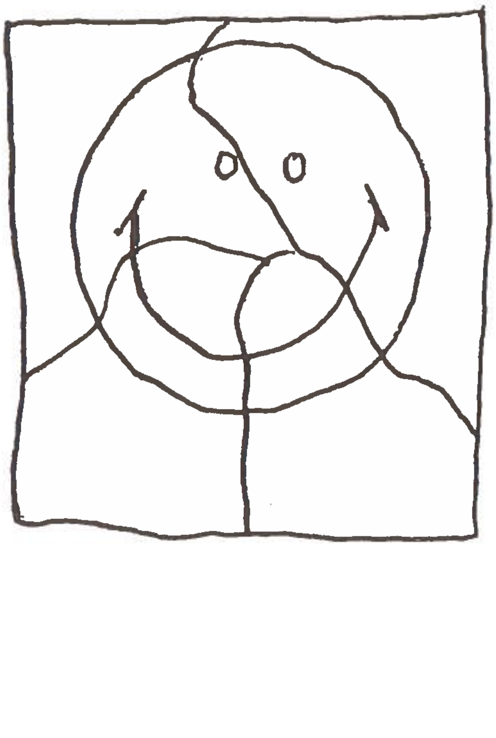
APPENDIX bA



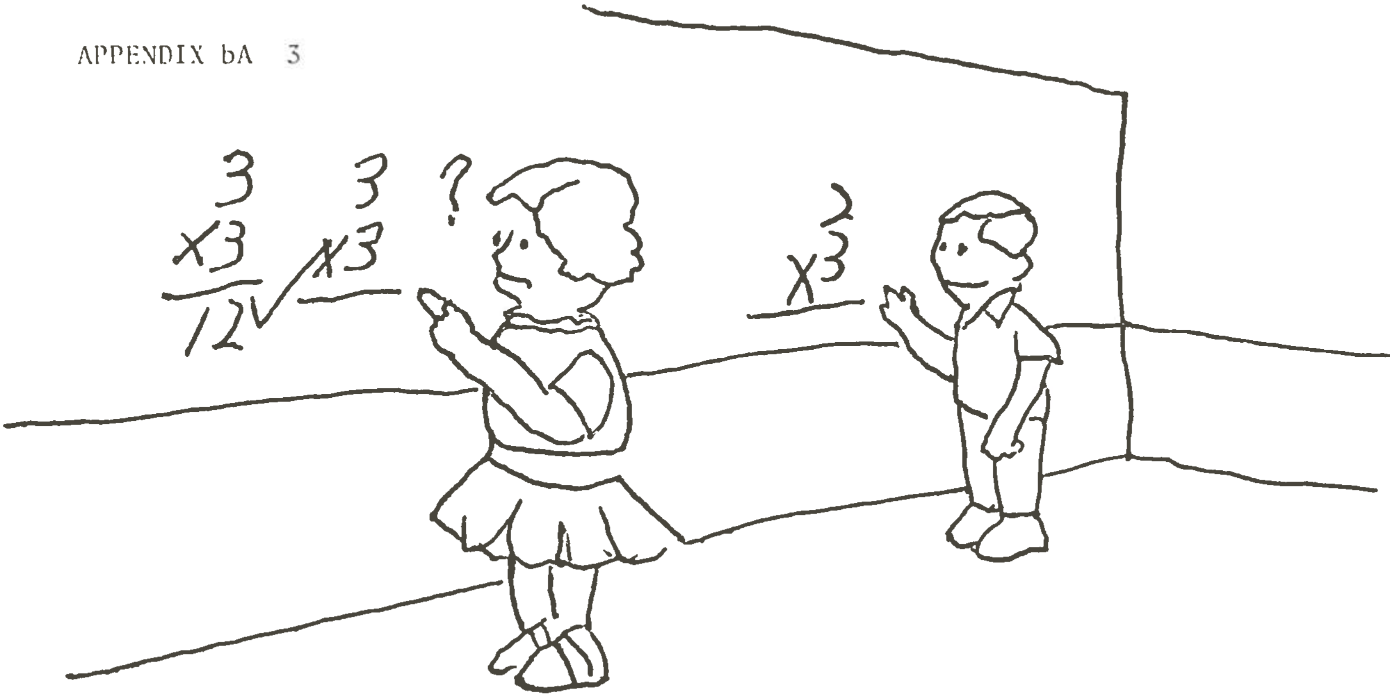


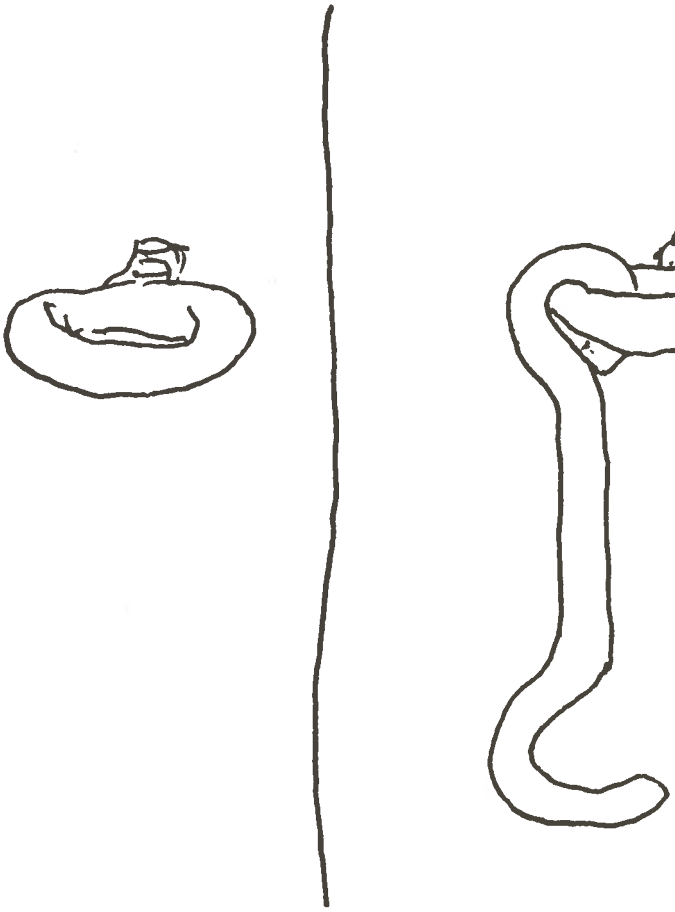
APPENDIX bA 2



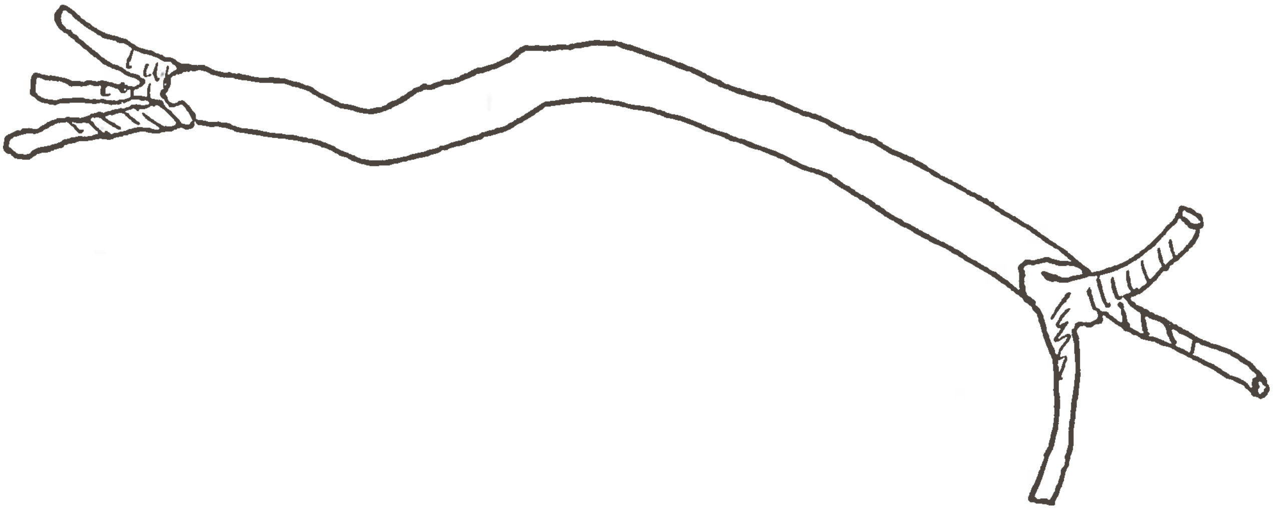


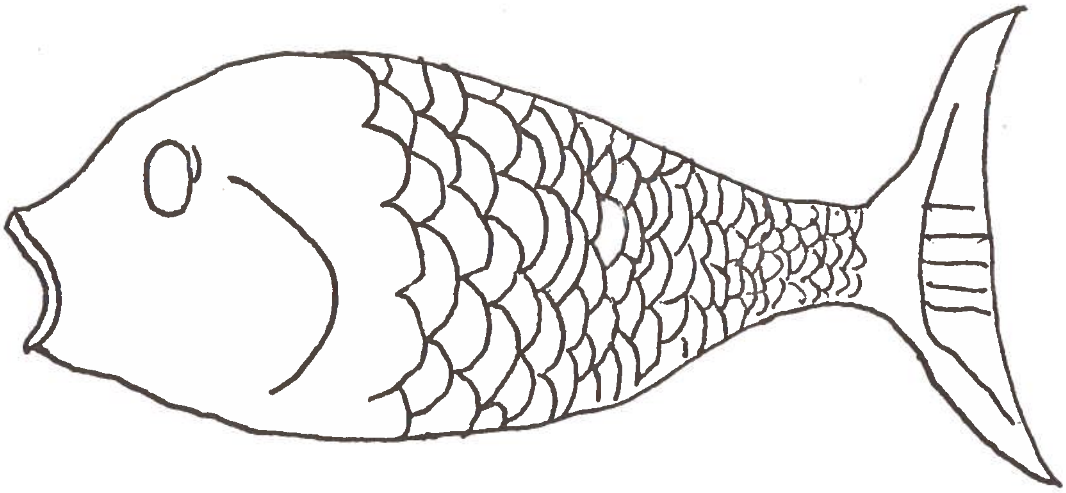
APPENDIX 6A 3





APPENDIX bA 4

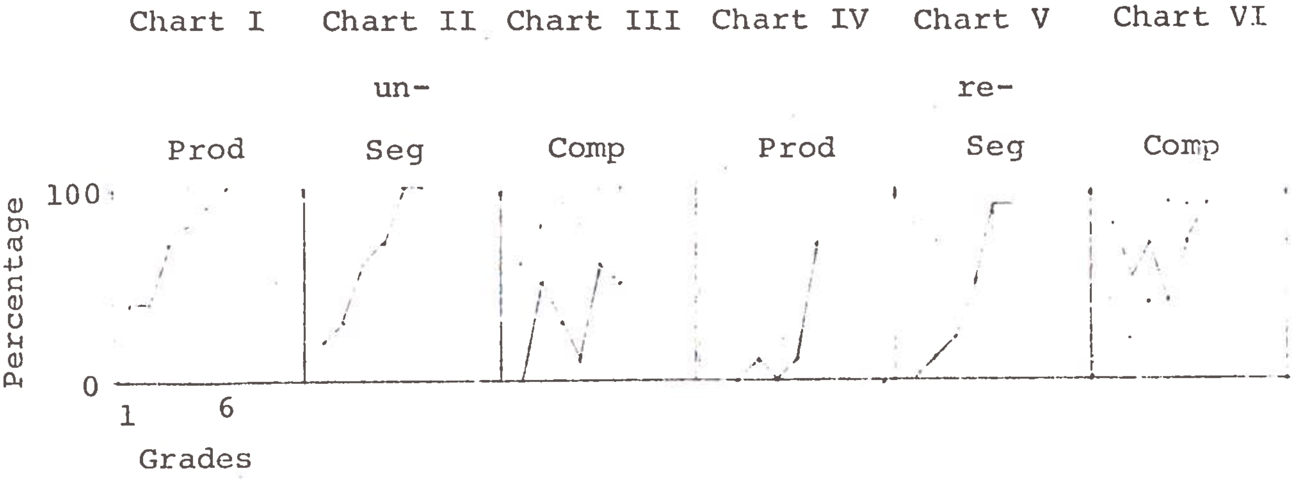




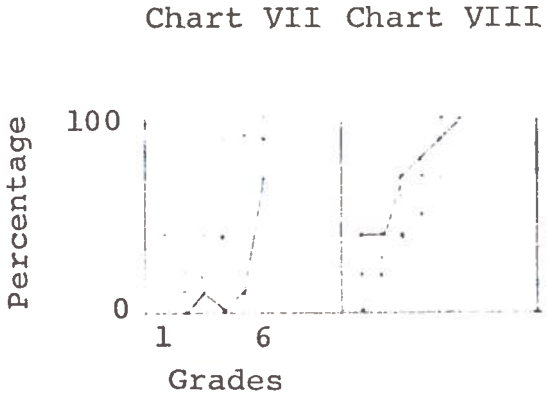
APPENDIX cA

Table 1.

	un-						re-					
Grade	real			nonsense			real			nonsense		
	Production	Segmentation	Comprehension	Production	Segmentation	Comprehension	Production	Segmentation	Comprehension	Production	Segmentation	Comprehension
First	100	100	60	40	20	0	0	80	40	0	0	80
Second	90	100	80	40	30	50	0	70	20	0	10	50
Third	100	100	90	70	60	30	0	100	40	10	20	70
Fourth	100	100	80	80	70	10	0	90	90	0	50	40
Fifth	100	100	100	90	100	60	20	100	90	10	90	70
Sixth	100	100	100	100	100	50	70	100	100	70	90	90



real
nonsense
without
unwizzle



production, nonsense
segmentation, nonsense
comprehension, real
production, nonsense, without
wizzle

IV. VIA User's Manual

by Robert Bryan and Peggy Lewis

TABLE OF CONTENTS

I. Introduction 177

II. Necessary Resources 178

III. General Overview 178

A. What the Programs Do 178

B. Options in Choice of Programs 186

IV. INDEX 188

A. Input Data; the Text 189

1. PROS 192
2. PLAY 195
3. SPOKE 195
4. POET 196
5. VPLAY 196
6. MILT 196

B. Parameter and Title Cards 198

1. Parameter Card 198
2. Title Card 199

C. Job Deck and Running Suggestions for INDEX 200

1. Control Cards 201
2. Long and X-Long Jobs 203

D. Description of Printed Output 204

V. PREFIX 206

A. Function 206

B. Job Decks and Running Suggestions for PREFIX 208

VI. SUFFIX 210

A. Function 210

B. Parameter and Title Cards 213

1. Parameter Cards 213

VI. B. 1. a. Parameter card 1 213
b. Parameter card 2 213

2. Title Card 214

C. Job Decks and Running Suggestions for SUFFIX 215

D. Description of Output 218

1. Audit and Error Message 218
2. Word Listing 219

VII. SELECT 220

A. Function and Parameter Cards 220

B. Job Deck and Running Suggestions for SELECT 222

VIII. THESR 223

A. Function and Input 223

B. Parameter Card 224

C. Job Deck and Running Suggestions for THESR 225

D. Description of Output 227

IX. SAMPLER 229

A. Capabilities 229

B. Parameter and Title Cards 229

C. 1. Parameter Cards 229
2. Title Card 232

C. Printed Output 232

D. Job Deck and Running Suggestions for SAMPLER 232

X. EXAMPLE 234

A. INDEX 234

1. Job Deck 234
2. Sample Input Text 234
3. Sample Printed Output 235
4. Job Deck for SAMPLER 239
5. Sample Output from SAMPLER 240

X. B. PREFIX 242

1. Job Deck for the SORT 242
2. Job Deck for PREFIX 242
3. Job Deck for SAMPLER 243
4. Sample Output from SAMPLER 244

C. SUFFIX 246

1. Job Deck for the Sort 246
2. Job Deck for SUFFIX 246
3. Sample Printed Output 247
4. Job Deck for SAMPLER 254
5. Sample Output from SAMPLER 255

D. THESR 257

1. Job Deck for the Sort 257
2. Job Deck for THESR 257
3. Sample Input Thesaurus 258
4. Sample Printed Output 259

Appendix A - Prefix Data File 261

Appendix B - Suffix Data File 279

Appendix C - Function Words File 285

Appendix D - Punctuation Words File 289

Appendix E - Iterative Use of THESR 291

I. INTRODUCTION

This manual describes the capabilities of and provides operating instructions for the VIA (Verbally Indexed Associations) automated language analysis programs which are currently operational on the Honeywell 635 computer installation at the University of Kansas. It is a user's, and not a programer's, manual in that it provides all and only that information necessary for the user with no previous acquaintance with the programs and no computer science background to fully understand what the programs do and what procedures he must follow in order to successfully use the programs for his particular language analysis task. As such, the VIA package of programs is treated as a black box and information about the internal structure of the programs not essential to their successful implementation by the user is not included. The user who is interested in how the programs work, who is using the programs at an installation other than the University of Kansas, or who would like information about VIA related programs currently available at other installations or about plans to expand the capabilities of the University of Kansas package should use this manual in connection with the research reports (Sedelow, 68-69), (Sedelow, 69-70), (Sedelow, 71) and (Sedelow, 71-72) which provide extensive documentation and program listings for VIA programs at this and other installations as well as descriptions of research currently being undertaken at the University of Kansas.

II. NECESSARY RESOURCES

The resources necessary for running VIA programs at the University of Kansas include a "project number" assigned by the Computation Center. This may be acquired through a department of the University, or directly from the Computation Center. When possible, current estimates appear to be used as funding guidelines for each step (VIA program or sorting activity) of an analysis.

Magnetic tape is the medium used by VIA programs for data storage, primarily because very large amounts of data may be stored and accessed at a reasonable cost, and tapes are relatively easy to acquire. From two to eight tapes may be used by VIA, depending on the number of intermediate steps for which the user wishes to save data and the number of VIA programs run. An optimum number of tapes would fall somewhere between these extremes, probably at three or four. This would allow preservation of the initially indexed data in some form along with a "backlog" of several steps so that loss of intermediate data due to unforeseen problems would be kept at a minimum. Tapes may be acquired through a department of the University, or purchased through the Computation Center.

III. GENERAL OVERVIEW

III. A. What the Programs Do

The VIA package described in this manual consists of six self-contained FORTRAN IV programs: INDEX, PREFIX, SUFFIX, SELECT, THESR, and SAMPLER. The first five programs must be run in the order in which they are listed as each takes as input the magnetic tape data file created by its predecessor (with certain exceptions as explained below). The programs will be referred to by these names, written in all capitals, throughout this manual. The activity and operation of each of these programs will be discussed in detail below. In this section an overall picture of the functions of the programs is given to introduce the reader to the type of analysis carried out by the VIA package. Throughout the following discussion, where it is important that a distinction be made, 'word token' is used to denote a single occurrence of a word in a text and 'word type' to denote a particular character string. Thus, in the preceding sentence there are 37 word tokens but only 28 word types, the word types 'word', 'is', 'a', 'to', and 'denote' being represented by more than one word token.

INDEX processes textual material (strings of English sentences) provided by the user (the text is entered into the system on punched computer cards) and produces a list of the word tokens appearing in the text together with indexing information for each giving its location in the processed text in terms of text divisions (e.g., 1st volume, 3rd chapter, 5th paragraph, 3rd sentence, 10th word token in that sentence). Punctuation symbols are treated as words by INDEX and 'word token' in the previous sentence means 'word and punctuation symbol token'. In preparing the text data to be processed by INDEX, text divisions must be marked with special symbols as explained below. For each word and punctuation symbol token in the text, INDEX produces a line of output containing the token itself and its associated indexing information. Tokens in the list will be in the order in which they appear in the text (cf. the sample output from INDEX in X.A.3.). INDEX can be run in one of six 'modes', suitable for processing prose, poetry, stage plays, verse plays, epic poems, and transcribed speech respectively.

PREFIX takes the magnetic tape output of INDEX as input and identifies word tokens with legal English prefixes. When a prefixed word token is found, it is so marked for use by SUFFIX, which will compare the remainder of the word token (minus the prefix) with other word tokens in searching for words with common stems.

SUFFIX accepts the magnetic tape output of PREFIX (or of INDEX if PREFIX was omitted) as input and identifies words in the list of text words which have the same stem, _i.e._, words that differ only in that they have different suffixes (represent-ing, represent-ation) or in that one has a suffix and the other does not (represent, represent-ing). If PREFIX was run, and only in that case, 'mis-represent-ing' would be found to have the same stem as 'represent', 'represent-ing', etc. A group of words found by SUFFIX to have the same stem is called a _root group_. SUFFIX assigns to each word processed a number called a 'matchcount' and marks words as belonging to the same root group by assigning to each word in a root group the same matchcount. In the following, "...be in the same root group as...", "...have the same stem as...", "...have the same matchcount as...", and "...match..." are used interchangeably in reference to word tokens or word types. (A word token or word type always matches itself.)

SUFFIX also counts the number of word tokens in the text for each word type and computes the number of word tokens for each root group. The number of word tokens in the text of a given word type is called the 'type frequency' of that word type and the number of word tokens in the text in a given root group is called the 'matchcount frequency' of that root group as well as the matchcount frequency of each word token and each word type in that root group. (The matchcount frequency for a given root is thus the sum of the type frequencies of word types in the root-group.) Printed output from SUFFIX lists the word types which appear in the text, grouped into root-groups, and the locations (volume, chapter, etc.) of each word token of each type together with the matchcount and matchcount frequency for each root-group and the matchcount and type frequency for each word type. This listing is ordered alphabetically by root-group, and by word type within root-groups. (cf. the sample output from SUFFIX in X.C.3.)

SELECT is an optional program in the VIA series which copies from the SUFFIX output tape only a specified portion of the processed text words as described below. This new tape may then be introduced as input to THESR in place of the original SUFFIX output tape.

THESR accepts as input the magnetic tape output of SUFFIX or SELECT plus a 'thesaurus' provided by the user. The 'thesaurus' is composed by the user using whatever standard references he wishes (dictionaries, standard thesauri, etc.)* and entered in the system on typed computer cards as explained in the following section. It consists of a set of 'primary words' and, associated with each primary word, a set of semantically related 'associate words' (cf. the sample thesaurus in X.D.3.). THESR uses the magnetic tape output of SUFFIX and this user supplied thesaurus to construct a 'text specific multilevel thesaurus'. By a multilevel thesaurus is meant an extension of the two level thesaurus supplied by the user; the extension is formed by listing for each associate word which matches a primary word its associates, for each word in those lists which matches a primary word its associates, etc. This linking process is carried out to the number of levels (up to five) specified by the user. For example, the four level extension of the two level thesaurus

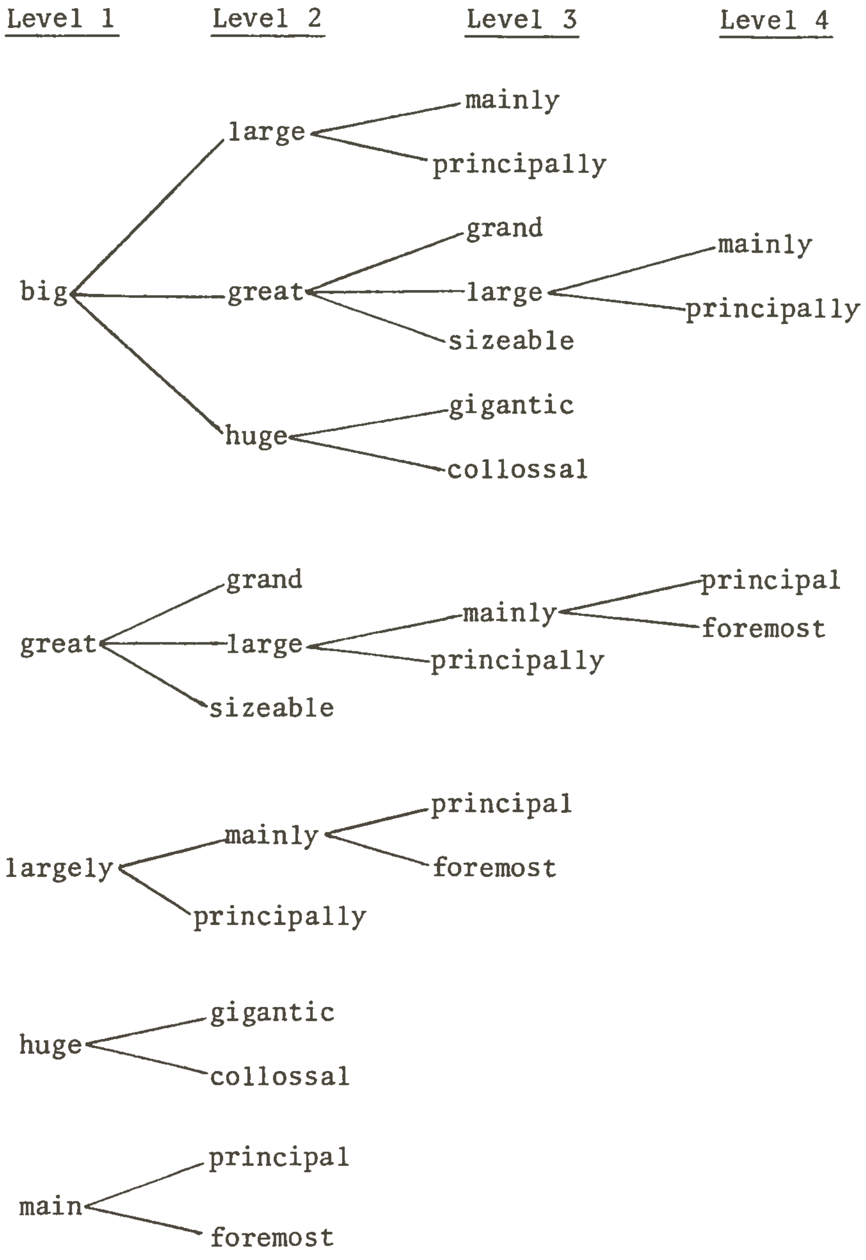
(1)

Primary Words (Level 1)	Associate Words (Level 2)
big	large great huge
great	grand large sizeable
largely	mainly principally
huge	gigantic collossal
main	principal foremost

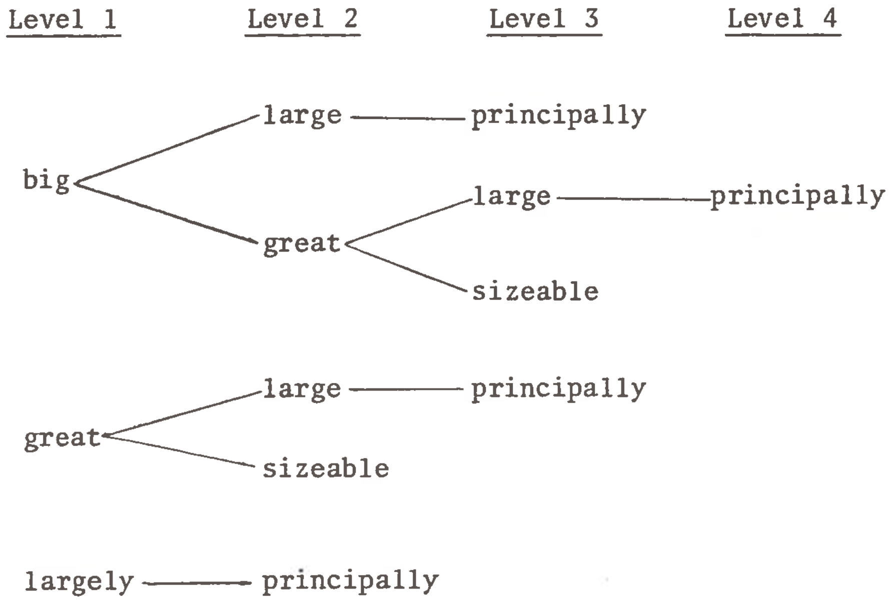
* Work is being completed on a computer accessible version of Roget's International Thesaurus which will eliminate the need for the user to compose a thesaurus.

is

(2)



'Large' keys further branching because it is in the same root group as the primary word 'largely'. By 'text specific thesaurus' is meant a (two to five level) thesaurus containing only words which match a word in the text being processed. For example, if all the words in the sample thesaurus (1) match a word in the text except, say, 'huge', 'grand', 'collosal', 'main', and 'mainly', then the text specific version of (2) would be



Redundancies in the tree are pruned as explained in VIII.D. (cf. the sample output from THESR in X.D.4.).

Each primary word which matches a text word, i.e. which is in a root group represented in the text by at least one token, appears as a 'root node' in the text specific thesaurus ('big', 'great', 'largely', 'huge', and 'main' are root nodes in (2)). The user may require, if he wishes, that only those primary words appear as root nodes which are in a root group represented in the text by ≥N tokens (i.e. which have a matchcount frequency ≥N), where N is any positive integer, by specifying the threshold value on a parameter card. For later reference, we will call a thesaurus whose root nodes are required to have matchcount frequencies greater than or equal to some positive integer N a 'threshold specified thesaurus with threshold N'. The printed output from THESR consists of a k-level extension of the 2-level user supplied thesaurus (where k is the depth of extension specified on a parameter card) which is text specific in that <u>all</u> nodes in the tree have matchcount greater than or equal to 1 and threshold specified in that all <u>root</u> nodes have matchcount greater than or equal to N (where N is the threshold specified on a parameter card). Also given in the printed output are the matchcount and matchcount frequency of each node in the tree.

Branching in a three or more level thesaurus occurs only when an associate word matches a primary word. The multilevel thesaurus constructed by THESR is thus rich in branching in proportion to the degree of overlap between the set of primary words and the set of associate words. If no associate word matches a primary word, the output of THESR will be a text specific version of the two level thesaurus supplied as input. (If N=1.)

SAMPLER is a utility program in the VIA package which may be used to print selected portions of the magnetic tape output from INDEX, PREFIX, SUFFIX or SELECT or a sorted version of those tapes as described in section IX.

III. B. Options in Choice of Programs

The overall structure of the VIA package is represented schematically in Figure 1.

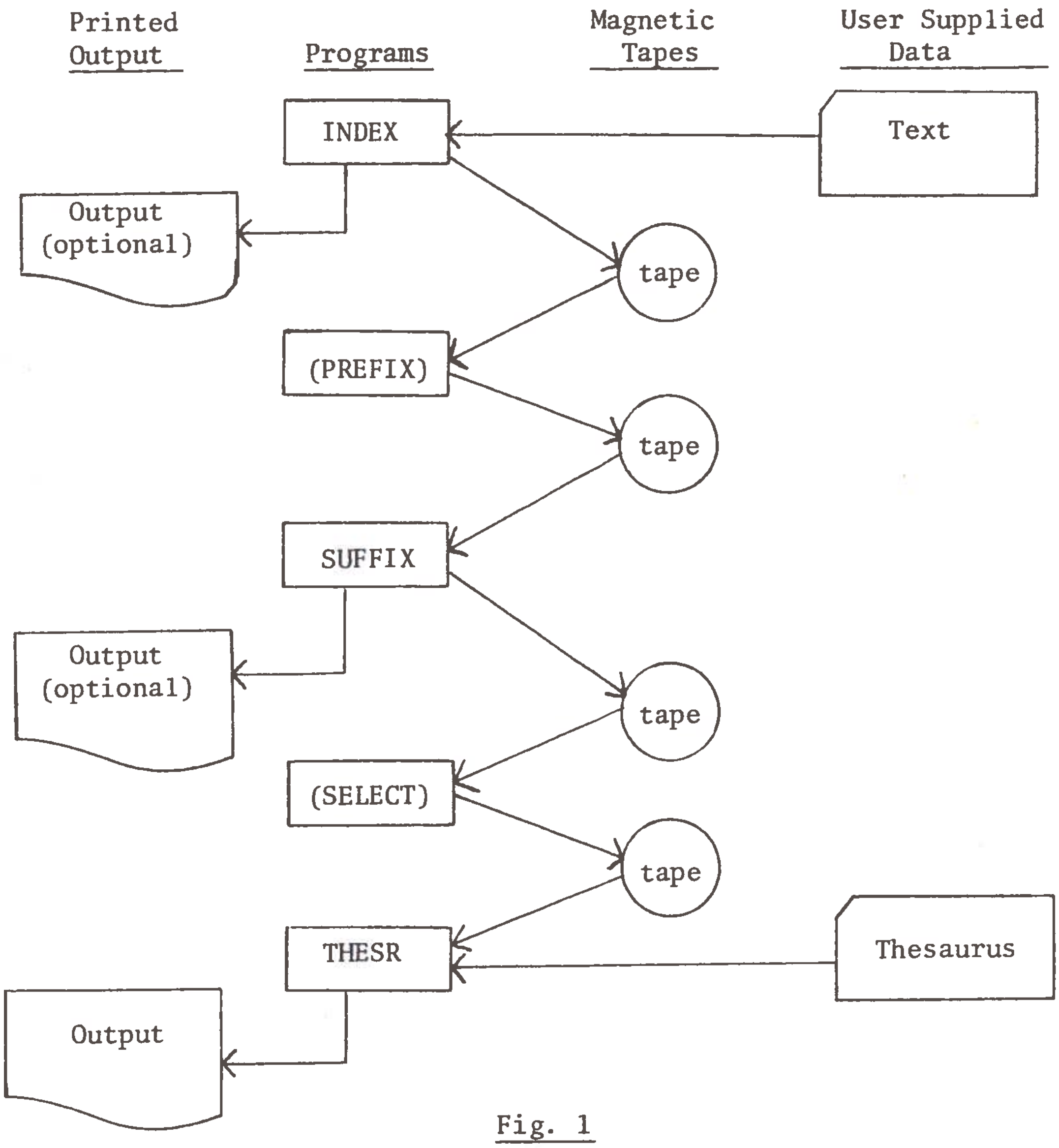


Fig. 1

Depending on the kind of output the user wants, he may use from one to all five of the programs. However, INDEX or INDEX and PREFIX must be run before SUFFIX can be run and SUFFIX or SUFFIX and SELECT must be run before THESR can be run. That PREFIX and SELECT are optional in this program sequence is indicated in the diagram by enclosing the program name in parenthesis. (If PREFIX is omitted, then of course the output tape from INDEX becomes (after being sorted) the input tape for SUFFIX, etc.) An arrow from a program box to the 'Printed Output' column indicates the option of asking for printed output. THESR always produces printed output. The user must supply the text to be processed by INDEX and a thesaurus if THESR is run. Not shown in Fig. 1 are sorts, which must be run before PREFIX, SUFFIX, and THESR as explained in the 'Job Deck' portion of the sections on those programs, and SAMPLER, which may be run against the (sorted or unsorted) magnetic tape output of any program in the series except THESR.

In terms of choice of programs, therefore, the user has the following options:

1.	2.	3.	4.	5.	6.	7.
INDEX	INDEX SUFFIX	INDEX PREFIX SUFFIX	INDEX SUFFIX THESR	INDEX PREFIX SUFFIX THESR	INDEX SUFFIX SELECT THESR	INDEX PREFIX SUFFIX SELECT THESR

The user would presumably elect to receive printed output from the last program in each case and has the option of doing so from INDEX and SUFFIX in any series in which they appear. SAMPLER may follow any program except THESR in each of the 7 series.

In the following sections each of the programs in the VIA package is discussed in detail. The discussion of each program includes an explanation of the total range of options available for that program and the operating procedures for exercising these options, descriptions of input and output data sets where appropriate, and the 'job deck' and running suggestions for that program. 'Job deck' will be used in this manual to refer to the total set of computer cards which must be keypunched, ordered, and submitted to the dispatch desk to run a program. It may include

1. data cards (in the case of INDEX and THESR)
2. parameter and title cards

and 3. control cards.

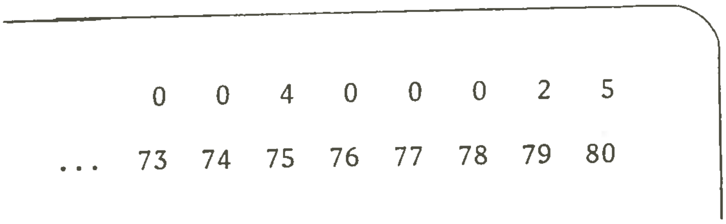
By convention, data cards and parameter and title cards are white and control cards, all of which have a '$' in column 1, are orange. Section X contains an example, based on a brief sample text, which gives examples of input and output, where appropriate, for all of the VIA programs. It will be helpful to refer to that example in reading sections IV -- IX.

IV. INDEX

INDEX can be run in one of six processing modes -- PROS, POET, MILT, PLAY, VPLAY, and SPOKE -- suitable for processing prose, poetry, long poems analogous to Paradise Lost, stage plays, verse plays, and transcribed speech respectively. The form of the input data and the printed output depends on the choice of mode, which is made on a parameter card as described below. PROS, PLAY, and SPOKE are 'sentence modes' in which indexing information is compiled for sentence and word in sentence in addition to larger textual divisions such as volume, chapter, and paragraph, whereas POET, VPLAY, and MILT are 'line modes' where indexing information is compiled for line and word in line in addition to larger textual divisions. The sentence is ignored in line modes.

IV.A. Input Data; the Text (see sample input in X.A.2.)

The text to be processed by INDEX is entered into the system on punched computer cards. In any mode the text should appear in columns 1-71 of the cards. Columns 76-80 may be used for card sequencing numbers in case the deck is dropped. Columns 73-75 are used for various purposes depending on the mode chosen as discussed below. All numbers in both the columns 73-75 and columns 76-80 fields should be 'right justified' as shown in Fig. 2, in which 4 and 25 appear in the first and second number fields respectively.



	0	0	4	0	0	0	2	5
...	73	74	75	76	77	78	79	80

Fig. 2

In addition to all words, all punctuation should be blank delimited (preceded and followed by a blank). For example, the sentence "As for me, I hate opera." should be keypunched: As for me , I hate opera . Punctuation symbols are treated as words in indexing, so that 'I' would be indexed as the fifth word in the sentence "As for me, I hate opera." Adjacent blanks are treated as a single blank. For a list of symbols recognized as punctuation by the system, see Appendix D.

In sentence modes, a hyphen should be placed in column 72 if a word is continued from one card to the next. It is important that no blanks appear between the initial word fragment and the hyphen in column 72. Words can, of course, be divided at any point. If no hyphen appears in column 72, the word counter is automatically incremented when a new card is read. Therefore, if a word ends in column 71 of one card, the next word or symbol of the text can be punched in column 1 of the next card with no preceding space. In line modes, the line counter is incremented when a new card is read unless an 'at' sign (@) appears in column 72, which signals the continuation of the line to the next card. As in sentence modes, if a word ends in column 71, the next word of the same (or a new) line can appear in column 1 of the next card and if a word is continued from one card to the next, a hyphen should be typed in column 72. In line modes, a continued word forces continuation of the line in which it appears. Since adjacent blanks are treated as a single blank, end of card situations can be avoided by leaving several blanks at the end of the text portion (columns 1-71) of the card. Legal solutions to various end of card situations are given in Figure 3.

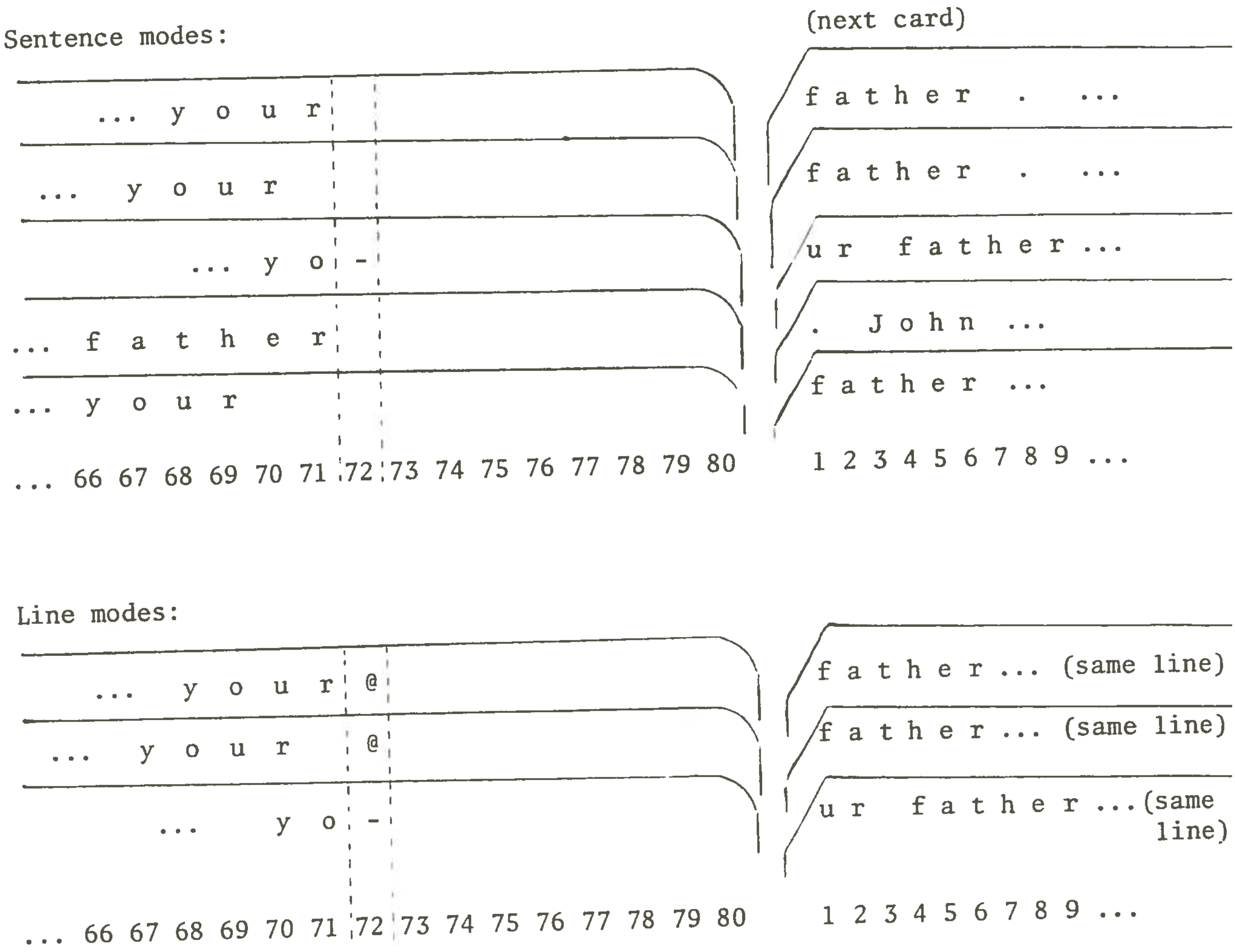


Fig. 3

Text divisions other than the sentence, which the program recognizes by finding sentence terminal punctuation (., !, ?), are indicated variously, depending on the processing mode chosen on the parameter card. The text division conventions described in the following paragraphs are summarized in Figure 5.

IV.A.1. PROS

In PROS, three levels of text division above sentence may be marked. These will be called paragraph, chapter, and volume (a chapter contains several paragraphs and a volume several chapters) in the output unless otherwise specified by the user as explained below. End of paragraph should be marked with a double period (..), which replaces the period of the last sentence in the paragraph or follows the ? or ! of the last sentence in the paragraph, e.g.

```
                    (new paragraph)

He closed the door .. John ...
Did he close the door ? .. John ...
Close the door ! .. John ...
```

End of chapter is marked with $$$ and end of volume with $$$$. End of chapter and volume markers should be blank delimited and are treated like words in end of card situations. If no end marker for a given division appears in the text submitted for processing, the counter for that division will be one for all words in the output. Only one of '$$$$', '$$$', or '..' will appear at a given point in the text since an end of division marker automatically marks the end of all lower level divisions. However, sentence terminal punctuation (., !, ?) should be retained, e.g.

```
... ever after . $$$$ Once upon ...
```

and _not_

```
... ever after .. $$$ $$$$ Once upon ...
```

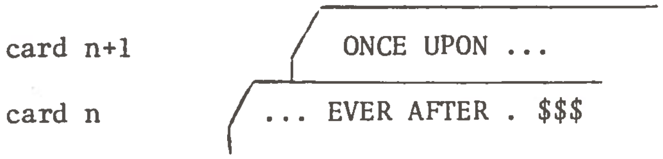
'Updating' capabilities exist for some of the text divisions in PROS and certain other modes which allow the counter for a text division to be arbitrarily set by inserting an 'update' card in the set of data cards. Volume, Chapter, Paragraph and Sentence can be updated in PROS. The update card should contain only the name of the level whose counter is to be set (volume, etc.) immediately preceded by a '$', and the number to which the counter is to be set, which follows the level name and is separated from it by a single space. For example,

```
$CHAPTER 18
```

will reset the chapter count to 18. The instruction may appear anywhere on the card. An update card will normally be inserted at the beginning of a text division of the level being updated and replaces the end of division marker (unless it precedes all data cards) of that level which would otherwise appear at that point in the data. It will reset to 1 the counters for lower level text divisions and leave unaffected the counters for higher level divisions. The use of the update card is illustrated in Figure 4.

Without update card:

Data deck:

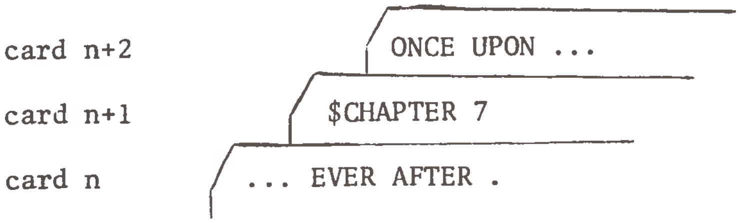


Printed Output:

	Volume	Chapter	Paragraph	Sentence	Word
:					
EVER	2	3	5	4	7
AFTER	2	3	5	4	8
.	2	3	5	4	9
ONCE	2	4	1	1	1
UPON	2	4	1	1	2
:					

With update card:

Data deck:



Printed Output:

	Volume	Chapter	Paragraph	Sentence	Word
:					
EVER	2	3	5	4	7
AFTER	2	3	5	4	8
.	2	3	5	4	9
ONCE	2	7	1	1	1
UPON	2	7	1	1	2
:					

Fig. 4

If the user wishes to use a character other than $ for the end of volume and end of chapter marker he may do so by so indicating on a parameter as explained below. The specified character should replace the $ on update cards as well. The user may put page numbers in columns 73-75 and card sequence numbers (in case the deck is dropped) in columns 76-80 for his convenience, but neither will appear in the output nor be used by the program (unless SEQ=YES appears on the parameter card. Cf. IV.B.1).

IV.A.2. PLAY

PLAY is the same as PROS in all respects except that $$$$ and $$$ are used to mark divisions which are called Act and Scene respectively in the printed output. Only Act and Scene counts can be updated, with $ACT X and $SCENE X cards respectively. (X denotes a positive integer.) Stage directions should be marked by preceding each word and punctuation symbol with an asterisk, e.g. *lights *dim *slowly *. Stage directions can then be included in or excluded from the processed data by an appropriate indication on a parameter card as explained below (cf. IV.B.1.)

IV.A.3. SPOKE

SPOKE is the same as PROS except that $$$$, $$$, and .. are used to denote what are called series, session, and speaker in the output.

The counters for the text divisions marked by $$$$, $$$, and .. are set with $SERIES X, $SESSION X, and $SPEAKER X cards respectively.

IV.A.4. POET

Since POET is a line mode, the text should be typed one line per card or continued from one card to another by putting '@' (or '-' if a word continues to the new card) in column 72 of the first card. Only one level of text division higher than line can be marked in POET. It may be used for page, stanza, or some other division as the user sees fit. It is marked by numbers in columns 73-75 and is called stanza in the output. For example, if pages are to be marked, all those cards containing lines on page 1 should have 001 in columns 73-75, those cards containing lines on page 2, 002, etc. The counters in POET cannot be reset with update cards. The markers $$$$, $$$, and .. are not used in this mode.

IV.A.5. VPLAY

VPLAY is a line mode and lines are treated as in POET. Use of $$$$, $$$, and .. and updating capabilities are as in PLAY, as is the use of columns 73-80. Stage directions are treated as in PLAY.

IV.A.6. MILT

MILT is a line mode and lines are indicated as in POET and VPLAY.

Three levels of text division above line can be marked and are called Volume, Chapter, and Paragraph in the output. Chapters and paragraphs are marked as in PROS. Volume is marked by numbers in columns 73-75. $$$$ is not used in MILT. No updating facilities exist for this mode.

The conventions pertinent to preparing the input data text for the various modes are summarized in Figure 5.

Mode Marking Device	PROS	PLAY	SPOKE	POET	VPLAY	MILT
$$$$	Volume	Act	Series		Act	
$$$	Chapter	Scene	Session		Scene	Chapter
..	Paragraph	Paragraph	Speaker		Paragraph	Paragraph
Col. 73-75	Page (Opt)	Page (Opt)	Page (Opt)	Stanza	Page (Opt)	Volume
Col. 76-80	SEQUENCING INFORMATION (OPTIONAL)					
Special Instructions				One line per card image; @ in Col. 72 for line continuation		
		* For Stage Instructions			* For Stage Instructions	

Fig. 5

IV.B. Parameter and Title Cards

Besides the data cards discussed in III.A., the job deck for INDEX must contain a parameter card and a title card. In the description of parameter cards, 'instruction' will refer to a character string containing a single equals sign and separated from the rest of the card by commas. 'PROC=PROS' is an instruction on the parameter card for INDEX, for example.

IV.B.1. Parameter Card

The parameter card has the following format:

$$\text{PROC=}\left\{\begin{array}{l}\text{PROS,}\\ \text{MILT,}\\ \text{SPOKE,}\\ \text{POET,}\\ \text{PLAY,}\\ \text{VPLAY,}\end{array}\right\}\text{PRINT=}\left\{\begin{array}{l}\text{YES,}\\ \text{NO,}\end{array}\right\}\text{SEQ=}\left\{\begin{array}{l}\text{YES,}\\ \text{NO,}\end{array}\right\}\text{STAGE=}\left\{\begin{array}{l}\text{YES,}\\ \text{NO,}\end{array}\right\}\text{DELIM=}\left\{\begin{array}{l}\$\\ \text{some other}\\ \text{character}\end{array}\right.$$

1234...

The five instructions must appear in the indicated order on the parameter card but the PRINT, SEQ, and STAGE instructions may be left off entirely, in which case NO is assumed as a default value. Braces indicate that exactly one of the choices within the braces is to be selected. The entire character string on the parameter card must begin in column 1 and must contain no internal blanks (even if one or more of the instructions is omitted) and instructions must be separated by commas. (cf. the sample job deck for INDEX in X.A.1.) The choice of one of the options in the braces after 'PROC= ' will determine the mode in which the INDEX run is made. PRINT=YES will cause printed output to be produced. PRINT=NO should be typed if no printed output is desired from the INDEX run. SEQ=YES will cause the program to check the data card sequencing numbers in columns 76-80 and return an error message without processing the data if the cards are out of order. If SEQ=NO, the data cards will be assumed to be in the correct order by the program and columns 76-80 will be ignored. SEQ=YES should not appear on the parameter card if columns 76-80 are blank. If the processing mode is PLAY or VPLAY, STAGE=YES will cause stage directions to be included in the data processed by INDEX. If STAGE=NO, stage directions will be ignored. In a mode other than PLAY or VPLAY that instruction should be left off of the parameter card. If a character other than $ is typed after DELIM= , in typing the input text that character should be used in place of a $ for end of division markers and update cards.

IV.B.2. Title Card

The title card may contain any character string anywhere on the card (e.g. the title of or other information about the text being processed). The character string typed on the title card will be printed at the top of each page of output if the PRINT=YES option is chosen.

IV.C. Job Deck and Running Suggestions for INDEX

```
Card Column:    1       8          16
_____________________________________________________________________

                $       IDENT      project number,name
                $       SELECT     2632-SEDELØW/INDEX
                $       INCØDE     IBMF
                Parameter Card
                Title Card
 (optional)     $       LIMITS     time
 (optional)     $       TAPE       06,X6SD,,tape number,,tape label,ØUT
 ØUTPUT TAPE    $       TAPE       21,X2DD,,tape number,,tape label,ØUT
                $       DATA       20,IBMF,CØPYD
                Input Text

                $       ENDCØPY
 (optional)     $       SELECT     2632-SEDELØW/LIST
                $       ENDJØB
                ***EØF
_____________________________________________________________________
```

Job Deck 1 (INDEX)

Job Deck 1 shows the input deck which must be used to run INDEX. The parameter card, title card, and input text have already been described. The remaining cards (Honeywell 635 control cards) are discussed below, along with other miscellaneous information related to the INDEX job.

IV.C.1. Control Cards

Each card in Job Deck 1 which has a dollar sign ($) in column one is a Honeywell 635 control card. This dollar sign is not to be treated in the same manner as the delimiter described for INDEX input text; it is simply a marker to identify control cards.

All control cards have a few common characteristics besides the dollar sign in column 1. The control card type is identified beginning in column 8, and the user-supplied information begins in column 16. Spacing is important after column 16, as a blank signals that there will be no further information appearing on the card. (Any that does appear after a blank will be ignored.) An additional "control" card is the ***EØF card which appears at the end of each job deck as a signal to the card reader.

The IDENT control card identifies the user and his project number to the computer so that charges for the job are made to the correct account and the printed output is readily identifiable. The four-digit project number followed by a comma and the user's name should be punched beginning in column 16.

A SELECT control card calls a particular program from storage to be run. This should not be confused with the VIA SELECT program data cards. Each SELECT card should appear exactly as shown.

An INCODE card specifies what type of keypunch was used to produce the cards which immediately follow it. 'IBMF' indicates that the cards were punched on an IBM/026 keypunch. (Those found in the DISPATCH area of the Computation Center are type 026.) 'IBMEL' indicates a keypunch of type IBM/029. (These are frequently found in other buildings on campus.) The 'IBMF' or 'IBMEL' appearing in the second field of a DATA control card has exactly the same meaning as described above.

Tapes are identified by a five-digit tape number and an alphanumeric tape label. This information on a TAPE control card must match exactly the number and label which appear on the tape itself, or the job will not run. The same tape may not be assigned twice within the same job.

The first optional control card is a LIMITS card which specifies the maximum time allowed for the job to run (in hundredths of an hour). The time it will take for INDEX to process any given input text with no printing may be estimated at .02 hours per 10000 words (output word or punctuation symbol tokens). An additional .02 hours per 10000 words may be estimated if INDEX is producing printed output. Thus an input text containing about 30000 word or punctuation symbol tokens can be expected to run in .06 hours without printing, or .12 hours if printed output is produced.

The LIMITS control card is not necessary at all if INDEX will run in .10 hours or less. The first example (30000 words with no printed output) falls in this category. More than .10 hours but less than .20 hours expected run time will cause the job to be classified as a LONG job and will require the optional LIMITS card. The second example (30000 words with printed output) would fall in this category, and the LIMITS card would appear:

```
$          LIMITS          12,12000
1 ...      8 ...          16 ...
```

If more than .20 hours of processing time is expected, the job will be classified X-LONG and the LIMITS card must appear.

The second and third optional control cards in Job Deck 1 are used only in case the INDEX print option chosen was "YES" and the expected output will exceed 5000 lines. Remember, each word token or punctuation symbol will generate at least one line of printed output. If these conditions are met, then both the second and third optional control cards must be present and the job must be submitted as LONG (to 10000 lines) or X-LONG (more than 10000 lines) at the DISPATCH window. If the input text exceeds 500 cards, the job must be submitted as LONG anyway.

The procedure for submitting a job classified as LONG or X-LONG is described below.

IV.C.2. LONG and X-LONG Jobs

Classification of an INDEX job as LONG or X-LONG has been discussed in connection with control cards in the preceding section. Briefly, the conditions may be summarized as follows:

1. More than 500 cards in input text (LONG)
2. More than .10 hours of processing (LONG),
   or more than .20 hours of processing (X-LONG)

3. More than 5000 lines of printed output (LONG), or more than 10000 lines of printed output (X-LONG)

In order to submit a LONG or X-LONG job at the DISPATCH window, it must be accompanied by a 'Job Resource Form'. These are available in the DISPATCH area and are relatively simple to fill out. In the case of INDEX with no print, there will be one activity and one output tape. The total processing time required and number of cards in the input deck should be noted. If printed output is produced, there will be two activities and two output tapes; processing time and deck size should still be noted.

An additional condition forcing the classification of a job as LONG (at least) is the use of three or more magnetic tapes at one time. This is the case with PREFIX, and with SUFFIX if printed output is to be obtained.

IV.D. Description of Printed Output (see sample output in X.A.3.)

If the PRINT=YES option was chosen on the parameter card, the INDEX run will produce printed output consisting of one line of output for each word and punctuation symbol token in the text which contains that token followed by indexing information giving its location in the text. Tokens are listed in the order in which they appear in the text. By indexing information is meant numbers indicating, in the case of PROS, the volume, chapter, paragraph, sentence, and position in sentence of the token, and analogous information for the other modes. For example, if 'dog' were the fifth word in the third sentence of the 12th paragraph of the seventh chapter of the first volume of a text processed by INDEX in PROS, the line of output for that occurrence of 'dog' would be

	Volume	Chapter	Paragraph	Sentence	Word
DOG	1	7	12	3	5

The location of a token in the text is thus given by a 5-tuple of positive integers. This is true regardless of the mode in which INDEX is run. The names given to the 5 components of the 5-tuple (Volume, Chapter, Paragraph, Sentence, and Word for PROS) in printed output from INDEX vary depending on the choice of mode and are shown in the table in Figure 6 which, it should be noted, differs slightly from the table in Figure 5 which shows how different text divisions are marked in the various modes in preparing the input text for INDEX. For MILT, for example, Volume is marked in the input data by numbers in columns 73-75 of the input data cards as shown in Figure 5, but in printed output the component of the location 5-tuple which indicates the Volume in which the token appears is the first, as shown in Figure 6. In POET, which alone among the six modes uses three rather than five parameters for indexing, the first two components of the location 5-tuple in the output will be 1 for all tokens and Stanza, Line, and

Word in line are given by the third, fourth, and fifth components, respectively.

Mode Component	PROS	PLAY	SPOKE	POET	VPLAY	MILT
1st	Volume	Act	Series	(1)	Act	Volume
2nd	Chapter	Scene	Session	(1)	Scene	Chapter
3rd	Paragraph	Paragraph	Speaker	Stanza	Paragraph	Paragraph
4th	Sentence	Sentence	Sentence	Line	Line	Line
5th	Word	Word	Word	Word	Word	Word

Fig. 6

V. PREFIX

V.A. Function

PREFIX identifies words in the list provided by INDEX which have legal English prefixes. If a word is found to be prefixed by PREFIX, then the remainder of the word, without the prefix, is used by SUFFIX in grouping together words having the same stem. Thus, if PREFIX finds 'misrepresent' to have the prefix mis-, SUFFIX will group misrepresent (and misrepresented, misrepresenting, etc.) with represent (represented, etc.). If PREFIX is not run, SUFFIX will not find the prefixed words to have the same stem as the non-prefixed words. PREFIX takes as input the magnetic tape output of INDEX (which is similar but not identical to the printed output). No further user supplied data is required.

PREFIX finds prefixes by reference to the 'Prefix Data File' which appears in Appendix A. The file lists the character strings recognized as prefixes by PREFIX in alphabetical order and marks each as an INCLUD or an EXCLUD prefix. Those marked INCLUD are followed by an alphabetical list, called an INCLUD list, of words (delimited by apostrophes) which begin with the letters of the prefix and are judged by PREFIX to be prefixed words. All words beginning with the letters of the prefix not on the INCLUD list will not be found to be prefixed. Those prefixes marked EXCLUD are followed by a list, called an EXCLUD list, of words which begin with the letters of the prefix but are judged by PREFIX not to be prefixed words. Any word beginning with the letters of the prefix which does not appear in the EXCLUD list following the prefix will be found by PREFIX to be a prefixed word. A word beginning with a character string not included in the list of prefixes in the Prefix Data File is not found to be prefixed by PREFIX. Thus, since "RE" is marked EXCLUD and "READ" appears in the following list, "READ" is not found to be prefixed. Since "REDO" does not appear in the list, it is judged to be prefixed by PREFIX. Since "EX" is marked INCLUD and "EXPORT" is in the following list and "EXTRA" isn't, "EX" is judged a prefix on "EXPORT" but not on "EXTRA". If the following " ' " immediately follows the last letter of an n-character word in an INCLUD or EXCLUD list, the test word is only required to begin with those n-characters. Thus, "ATYPICAL" and "ATYPICALLY" match "ATYPI" in the INCLUD list following the prefix "A" and are found to be prefixed. If a space precedes the following apostrophe, the test word must be identical to the word in the INCLUD or EXCLUD list. Thus, "ANTICLIMAX" does not match "'ANTIC'" in the EXCLUD list following the prefix "ANTI" and is found to be a prefixed word by the program. It is possible in this way to determine manually whether any given word will be found to be prefixed by using the list in Appendix A. It is also possible to amend the prefix file, but it must be pre-processed before use by the PREFIX program. Readers interested in exercising this option are referred to (Joyce, 71-72), which includes a description of the prefix data pre-processing programs, PREF1 and PREF2. Theoretical problems involved in programs dealing with prefixes are discussed in the articles by Warfel in (Sedelow, 71-72) and in this report.

V.B. Job Decks and Running Suggestions for PREFIX

Data contained on a magnetic tape file produced by INDEX must be ordered alphabetically before it is introduced to PREFIX as input.

This procedure is illustrated as Job Deck 2. Because the VIA system is designed to handle large amounts of information, the Honeywell 635 SORT/MERGE subsystem is used; it is classified as a NORMAL job, and should be run separately.

```
Card Column:    1       8         16
________________________________________________________________
                $       IDENT     project number,name
                $       SELECT    2632-SEDELØW/SØRT
INPUT TAPE      $       TAPE      SA,X1DD,,tape number,,tape label,IN
OUTPUT TAPE     $       TAPE      SZ,X2DD,,tape number,,tape label,ØUT
                $       167PK     S1,X3R,R,R0001,SCRATCH,PRIVATE,0/number of links
                $       ENDJØB
                ***EØF
________________________________________________________________
```

Job Deck 2 (SØRT)

The IDENT, SELECT, and TAPE control cards have already been described in Section IV.C.1. (for INDEX). The 167PK control card designates a scratch file for the sort. The "number of links" specifies the size of this file and may be estimated as 20 or 25 links per 10000 "words" (word or punctuation symbol tokens) to be ordered. Thus, a text of 30000 "words" would use a scratch file of approximately 70 links.

The output tape from the sort job now becomes the input tape for the PREFIX job, therefore it must be identified with the same number and label in each job deck. Job Deck 3 shows the input deck for running PREFIX. Control cards are the same as previously described. The optional LIMITS card must be included if the estimated processing time exceeds .10 hours. Notice, however, that the PREFIX program is automatically classified as a LONG job due to the fact that it uses three magnetic tapes at one time. (In addition to the user-specified input and output tape files, the PREFIX 'INCLUD' and 'EXCLUD' lists are stored on tape.) A reasonable estimate for processing time is .02 hours per 10000 "words" of text.

```
Card Column:    1        8        16
------------------------------------------------------------------
                $        IDENT    project number, name
                $        SELECT   2632-SEDELØW/PREFIX
 (optional)     $        LIMITS   time
 INPUT TAPE     $        TAPE     15,X15DD,,tape number,,tape label,IN
 ØUTPUT TAPE    $        TAPE     20,X20DD,,tape number,,tape label,ØUT
                $        ENDJØB
                ***EØF
------------------------------------------------------------------
```

Job Deck 3 (PREFIX)

Procedures for submitting PREFIX as a LONG or X-LONG job are analogous to those described for INDEX in Section IV.C.2., except that PREFIX will always have one activity and three tapes to be listed on the 'Job Resource Form'.

VI. SUFFIX

VI.A. Function

SUFFIX accepts the magnetic tape output (after having been sorted) from PREFIX or INDEX as input and groups words having the same stem as explained in III.A. No additional user supplied data is required. Printed output from SUFFIX will consist of (1) a list of the word types in the text grouped by root group (this output, called the 'word listing' below, is printed only if the PRINT=YES option is indicated on the parameter card as described in V.B.) and (2) an 'Audit and Error Message' which lists the parameters that were read in from parameter cards and the function words and punctuation symbols that occurred in the text. The Audit and Error Message is printed on every SUFFIX run.

SUFFIX groups words into root groups by reference to the 'Suffix Data File' printed in Appendix B. To determine from the list if two words (word here means 'word or word minus prefix' if PREFIX was run) will be assigned the same matchcount by SUFFIX, find the divergent endings of the words (e.g., -ing and -ation for 'representing' and 'representation', -e and -ing for 'come' and 'coming', ⟨blank⟩ and -ation for 'represent' and 'representation') in the alphabetical list of suffix pairs and then determine whether one of the two words appears in the exception list following the suffix pairs. The two word tokens (other than function word and punctuation symbol tokens; cf. V.B.) will be assigned the same matchcount only in case

(1) they are identical,

OR (2) a) both tokens are greater than or equal to three characters in length,

AND b) the tokens are identical in the first three characters,

AND c) the pair of divergent endings is in the list of possible suffix pairs, AND d) neither token is identical to a word token in the exception list following the suffix pair.

The list of possible suffix pairs is ordered alphabetically, with pairs one of whose members is ⟨blank⟩, the empty suffix, ordered first. For example, SUFFIX assigns "RECOMMEND" and "RECOMMENDATION" the same matchcount since "⟨blank⟩, ATION" is on the list of possible suffix pairs and neither "RECOMMEND" nor "RECOMMENDATION" is in the following exception list. "CATHODE" and "CATHOLIC" are not assigned the same matchcount since "DE, LIC" is not in the list of suffix pairs. "COME" and "COMIC" are not assigned the same matchcount for, although "E, IC" is in the list of suffix pairs, "COMIC" is in the following exception list. The LETTER rule merits a brief explanation because it somewhat misleadingly appears in the exception list. A typical example is the following: EXCEPT LETTER E. This means that if the final matching letter in the two words being examined is an E, the suffix pair is legal, unless an exception is found. If the final matching letter in the two words being examined is not an E, the program assumes it has not found a legal suffix and does not look at the rest of the exceptions.

The user may alter the contents of the 'Suffix Data File' for a particular SUFFIX run as well as the 'Function Word Data File' and 'Punctuation Data File' (Appendices C and D) which are used by SUFFIX to recognize function words and punctuation symbols. If this is done the date files must be preprocessed before use by SUFFIX. Those interested in exercising these options are referred to the documentation in (Joyce, 71) which describes the preprocessing programs SUFUNA1, SUFN1S, and SUFNS2.

VI.B. Parameter and Title Cards

VI.B.1. Parameter Cards

The job deck for SUFFIX will contain the following two parameter cards.

VI.B.1.a. Parameter Card 1

The first parameter card has the form

$$\begin{array}{l}
\text{FUNCT=}\begin{Bmatrix}\text{YES,}\\ \text{NO,}\end{Bmatrix}\text{PUNCT=}\begin{Bmatrix}\text{YES}\\ \text{NO}\end{Bmatrix}\\
1234\ \ldots
\end{array}$$

The entire character string on parameter card 1 must begin in column 1 and contain no blanks (i.e., ...NO,PUNCT...). A YES answer will cause function words or punctuation symbols, as indicated, to be included in the data (cf. VI.D., one parameter may have value YES and the other NO). A NO answer will cause the indicated set to be excluded from the data processed by SUFFIX.

VI.B.1.b. Parameter Card 2

The second parameter card has the form

$$\begin{array}{l cccccccccc}
\text{PRINT=}\begin{Bmatrix}\text{YES}\\ \text{NO}\end{Bmatrix} & X & X & X & X & X & Y & Y & Y & Y & Y\\
123\ldots & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19
\end{array}$$

where the braces indicate a choice as before and the X's and Y's indicate that two positive integer numbers are to be right justified in columns 10-14 and 15-19 respectively (e.g. 0004500050). PRINT=YES will produce printed output as described in Section V.D. If PRINT=NO, no output is printed and the number fields should be left blank. The 'PRINT= ' instruction must begin in column 1 and contain no spaces. The numbers in the two number fields 10-14 and 15-19 specify the number of lines per page in the printed output for the "Word Listing" (VI.D.1.) and "Audit and Error Message (VI.D.2.) respectively and may not exceed 58. If either number field is left blank (with PRINT=YES), 58 lines per page will be assumed as a default value.

If the PRINT=YES option was chosen, the user may elect to use the standard format for the "Word Listing" output as in the sample output in X.C.3. (this is the format described in VI.D.), or he may provide his own format as described in (Joyce, 71).

VI.B.2. Title Card

In addition to the parameter cards, the job deck for SUFFIX must contain a title card, if PRINT=YES appears on parameter card 2, which provides a title for the word listing portion of the printed output from SUFFIX. The title card may contain any character string and its contents will appear at the top of every page of the word listing output. If PRINT=NO appears on Parameter Card 2, no title card is included in the job deck.

VI.C. Job Decks and Running Suggestions for SUFFIX

Data produced by INDEX or PREFIX must be ordered alphabetically before it is introduced as input to SUFFIX. This is the case even though it may have been previously ordered before introduction to PREFIX; words which are found to have prefixes are stripped of those prefix characters, and may then appear quite out of order.

The input deck for this sort is illustrated as Job Deck 4. It is obvious that this is exactly the same as Job Deck 2 for sorting data before PREFIX. There is no difference in procedure for running the job: it is a NORMAL job, the scratch file should specify 20 to 25 links per 10000 text words of data (word or punctuation symbol tokens), and it should be run as a separate job in the VIA sequence.

```
Card Column:  1      8         16

              $      IDENT     project number,name
              $      SELECT    2632-SEDELØW/SØRT
INPUT TAPE    $      TAPE      SA,X1DD,,tape number,,tape label,IN
OUTPUT TAPE   $      TAPE      SZ,X2DD,,tape number  tape label,ØUT
              $      167PK     S1,X3R,R,R0001,SCRATCH,PRIVATE,0/number of links
              $      ENDJØB
              ***EØF
```

Job Deck 4 (SØRT)

The input deck for SUFFIX is shown as Job Deck 5. The parameter cards and title card have been previously discussed. The SELECT control card which specifies 'PARAMS' selects the default formatting information for printed output with SUFFIX. The user may choose to substitute his own format (where the SELECT card appears) according to the description in (Joyce, 71). Note that if the title card does not appear, the INCODE card immediately preceding it is unnecessary.

```
Card Column:      1     8        16

                  $     IDENT    project number,name
                  $     SELECT   2632-SEDELØW/SUFFIX
(optional)        $     LIMITS   time
                  $     INCØDE   IBMF
                  Parameter Card 1
                  Parameter Card 2
                  $     SELECT   2632-SEDELØW/PARAMS
                  $     INCØDE   IBMF
(optional)        Title Card (if PRINT=YES)
INPUT TAPE        $     TAPE     15,X15DD,,tape number,,tape label,IN
OUTPUT TAPE       $     TAPE     13,X13DD,,tape number,,tape label,ØUT
(optional)        $     TAPE     06,X6SD, ,tape number,,tape label,ØUT
                  $     SELECT   2632-SEDELØW/LIST
                  $     ENDJØB
                  ***EØF
```

Job Deck 5 (SUFFIX)

The optional LIMITS control card specifies the maximum time allowed for the program to run (in hundredths of an hour), and the amount of storage required by SUFFIX, which is always 25000 units. It is not required at all if SUFFIX can be classified as a NORMAL job. However, this is only possible for very small amounts of data and less than 5000 lines of printed output.

If more than 5000 lines of printed output are expected, then the last two optional control cards must appear in the job deck. In this case, SUFFIX will be automatically classified as at least LONG because of the three tapes.

Processing time may be estimated at .05 hours, plus an additional .03 hours per 10000 text "words," or slightly less if no printed output is to be produced. Guidelines for classifying and submitting SUFFIX are the same as those described for INDEX in Section IV.C.2., except that there is an additional tape (input file) to be listed on the 'Job Resource Form'. For example, a text of 30000 "words" to be processed by SUFFIX with printed output would be classified as X-LONG (more than 10000 lines of printed output) and the 'Job Resource Form' would note that there are two activities (SUFFIX processing and printing) using three magnetic tapes. As usual, total estimated processing time should be noted on the form.

For SUFFIX, estimates of processing time should be especially liberal, as the program is rather unpredictable for different types of text data.

SUFFIX uses three scratch files which may cause problems if they are not adjusted somewhat for very large texts. Input files ranging approximately from 10000 text words to 100000 text words should run without problems. Smaller amounts of data will run adequately, but with some wastage of resources. For adjustments, the user should refer to (Joyce, 71) and/or seek help from someone familiar with the VIA system.

VI.D. Description of Output (cf. sample output in X.C.3.)

Printed output from SUFFIX consists of two parts, an "Audit and Error Message" and, if PRINT=YES appears on Parameter Card 2, the "Word Listing".

VI.D.1. "Audit and Error Message"

The Audit and Error Message also consists of two parts. The first part is a list of the parameters that were read from the parameter cards and preceeds the word listing portion of the output if it appears. The second part follows the word listing and provides 'Record Counts' giving the total number of content word tokens, and the number of function word and punctuation symbol tokens (0 if FUNCT=NO and PUNCT=NO on Parameter Card 2) which were processed and, if FUNCT=YES and PUNCT=YES was typed on Parameter Card 2, a 'Function Words and Punctuation Summary' which lists the function word types and punctuation symbol types which occur in the text and provides counts giving the total number of function word tokens in the text, the total number of punctuation symbol tokens, and for each function word and punctuation symbol type the number of text tokens of that type. If function words and punctuation symbols are processed by SUFFIX, all function words are assigned the same matchcount (99998), and all punctuation symbols the same matchcount (99999).

VI.D.2. "Word Listing"

The word listing portion of the printed SUFFIX output lists the (content) word types which appeared in the text grouped into root groups and, for each word type, its type frequency and indexing information giving the location in the text of each token of that type. The location of a token is given by the same 5-tuple used to give its location in printed output from INDEX, but in output from SUFFIX the components of that 5-tuple are called Volume, Chapter, Paragraph, Sentence, and Word in sentence regardless of the mode in which INDEX was run. Figure 6 in IV.D. may be used to translate to appropriate division names for modes other than PROS. Also given is the matchcount and matchcount frequency for each root group. The listing is ordered alphabetically by root group and by word type within root groups. In the sample output from SUFFIX in X.C.3., 'abundant' is the single word type in the first root group and is represented in the text by a single token whose location is volume 1, chapter 1, paragraph 2, sentence 1, and the 16th word in the sentence. The matchcount frequency of the root group and type frequency of the word type are thus both 1, as shown in the 'Frequency of Occurrence' column. The fifth root group (Matchcount 5) has two word types, 'blue' and 'bluish', each represented by a single token in the text. The locations of those tokens are given in the parenthesis to the right. The root group has matchcount frequency 2. The root group with matchcount 42 has two word types, 'water' and 'waters', with type frequencies 3 and 1 respectively as shown in the 'Frequency of Occurrence' column. The locations of the three text tokens representing the word type 'water' are given in the three sets of parentheses to the right of 'water' (if 'water' had been represented by more than three tokens, the list of locations would have continued on a second line). The location of the single text token representing the word type 'waters' is given in the parentheses to the right of 'waters'. The matchcount frequency of root group 42 (the sum of the type frequency of word types in that root group) is four as shown in the 'Frequency of Occurrence' column.

VII. SELECT

VII.A. Function and Parameter Cards

If it is desired that THESR process the entire output from SUFFIX, SELECT should not be run and THESR should directly follow SUFFIX in the sequence of programs. SELECT may be used to select a portion of the original text for processing by THESR by indicating the desired portion of text on the following two parameter cards. (In the discussion of THESR in Section VIII,'text' means that portion of the original text passed to it by SUFFIX or SELECT.) The first parameter card contains a single instruction.

Parameter Card 1

$$\text{SELECT=}\left\{\begin{matrix}1\\2\\3\\4\end{matrix}\right.$$

One of 1,2,3, or 4 must follow "SELECT= ". The spacing on the card is immaterial as blanks are ignored. The number chosen to follow "SELECT= " indicates the level at which the text portion is selected. 1,2,3 and 4 correspond to text divisions in the various modes of INDEX as indicated in Figure 7.

	PROS	MILT	PLAY	VPLAY	POET	SPOKE
1	volume	volume	act	act	*	series
2	chapter	chapter	scene	scene	*	session
3	paragraph	paragraph	paragraph	paragraph	stanza	speaker
4	sentence	line	sentence	line	line	sentence

Fig. 7

If the processing mode for INDEX was POET, 3 or 4 must be specified on parameter card 1 and the first two components of the n-tuple described below must be 1.

The second parameter card indicates the text division to be selected. It contains n-tuples or pairs of n-tuples (where n is the level indicated on parameter card 1) whose coordinates indicate the desired division. For example, if 4 were written after "SELECT= " on parameter card 1, and the processing mode of INDEX is PROS, the 4-tuple (1,3,5,14) would select the 14th sentence of the 5th paragraph of the 3rd chapter of the 1st volume for processing by THESR. A pair of n-tuples selects the portion of text inclusively bound by the indicated divisions. For example, ((1,3,5,3), (1,3,5,14)) selects the 3rd to the

14th sentences, inclusive, of the 5th paragraph of the 3rd chapter of the 1st volume.

The index numbers must be enclosed in parentheses and separated by commas as shown. A pair of n-tuples must be enclosed in parentheses and separated by a commas, as shown, and the text division indicated by the first n-tuple of the pair must preceed that indicated by the second. As many n-tuples (single or pairs) as desired may appear in parameter card 2 but must be separated by commas. Additional cards may be used if necessary but an n-tuple or pair of n-tuples should not continue from one card to the next. For example, if PROS is the processing mode and 3 is the level indicated on parameter card 1, then

```
(1,3,5), ((1,6,7),(1,8,5)),(2,5,10)
```

selects the 5th paragraph of the 3rd chapter of volume 1, the 7th paragraph of the 6th chapter of volume 1 through the 5th paragraph of the 8th chapter of volume 1 inclusive, and the 10th paragraph of the 5th chapter of volume 2 for processing by THESR.

VII.B. Job Deck and Running Suggestions for SELECT

The input deck for running SELECT is illustrated as Job Deck 6. The parameters and control cards have all been discussed in previous sections. The processing time may be estimated as .02 hours per 10000 text words. The optional LIMITS control card may be omitted for .10

hours or less of estimated processing time (NORMAL job). If the job is classified as LONG or X-LONG according to the estimated processing time, then the LIMITS card must be included in the job deck. Since processor time is the only reason for classifying the job as LONG or X-LONG, the 'Job Resource Form' will be trivial or unnecessary.

```
Card Column:   1      8        16

               $      IDENT    project number, name
               $      SELECT   SELECT
(optional)     $      LIMITS   time
INPUT TAPE     $      TAPE     01,X1DD,,tape name,,tape label,,IN
OUTPUT TAPE    $      TAPE     02,X2DD,,tape name,,tape label,,ØUT
               $      INCØDE   IBMF
               Parameter Card 1
               additional parameter cards
               $      ENDJØB
               ***EØF
```

Job Deck 6 (SELECT)

VIII. THESR

VIII.A. Function and Input

THESR uses the output from SUFFIX or SELECT and a user supplied 'thesaurus' to build the text specific multilevel thesaurus described in Section II ('specific' to that portion of the text passed to THESR by SUFFIX or SELECT). As was the case with the input data for INDEX, the thesaurus is entered into the system from punched computer cards.

The 'thesaurus' consists of a set of 'primary words' and, associated with each primary word, a set of 'associate words' semantically related to it. It is typed in column 1-36 of computer cards, one card for each associate word in the thesaurus. Each card contains a pair of words, the associate word left justified in columns 19-36 and the primary word to which it is associated left justified in columns 1-18. The cards are ordered so that those containing the same primary word are adjacent and the primary words are in alphabetical order. The order of the associate words within a set of cards bearing the same primary words is immaterial (cf. the sample thesaurus in X.D.3.).

VIII.B. Parameter Card

The job deck for THESR must contain the following parameter card:

$$\text{LEVEL=}\begin{Bmatrix}2\\3\\4\\5\end{Bmatrix}\text{,SAVE=}\begin{Bmatrix}\text{YES}\\\text{NO}\end{Bmatrix}\text{,THRESHOLD= number,FORMATS=NO}$$

The four instructions must appear in the indicated order and be separated by commas. Spacing is immaterial as blanks are ignored. A positive integer is to be typed following 'THRESHOLD= '. The number following 'LEVEL= ' specifies the depth to which the input thesaurus is to be extended. The meaning of 'depth of extension' should be clear from the thesauri (1) and (2) in III.A. which are two and four level thesauri respectively. If LEVEL=2 is specified, no additional linking is made and the output produced by THESR is a text specific (if THRESHOLD=1) version of the input thesaurus. SAVE=YES causes a text specific version of the input thesaurus, together with linking information with which it could be extended to a multilevel thesaurus, to be saved on tape or some other medium. The reader who intends to use only VIA programs discussed in this manual should type SAVE=NO. The number following 'THRESHOLD= ' specifies the minimum matchcount frequency required of root nodes in the output tree. If THRESHOLD=1, the output tree is a text specific thesaurus. FORMATS=NO causes the output to be printed in the standard format described in Section VIII.D. The user may supply his own output format by replacing FORMATS=NO with FORMATS=YES and including format cards in the job deck (see Lewis, 1971).

VIII.C. Job Decks and Running Suggestions for THESR

Data contained on a magnetic tape file produced by SUFFIX or SELECT must be ordered by matchcount before it can be introduced to THESR as input. Job Deck 7 illustrates this procedure. This is not the same sort procedure used before PREFIX or SUFFIX. Like the other procedure, however, it is considered a NORMAL job and should be run as a separate job in the VIA sequence. All control cards have already been described. For reference, see Section V.B. or Section VI.C.

```
Card Column:  1   8        16
____________________________________________________________________________

              $   IDENT    project number,name
              $   SELECT   2632-SEDELØW/SØRTM
 INPUT TAPE   $   TAPE     SA,X1DD,,tape number,,tape label,IN
 OUTPUT TAPE  $   TAPE     SZ,X2DD,,tape number,,tape label,ØUT
              $   167PK    S1,X3R,R,R0001,SCRATCH,PRIVATE,0/number of links
              $   ENDJØB
              ***EØF
____________________________________________________________________________
```

Job Deck 7 (SØRTM)

The input deck for THESR appears as Job Deck 8. Parameter cards and thesaurus deck input have been described previously, as have all of the control card types. A reasonable processing time estimate is not yet available for THESR. Either a large input text file or a large input thesaurus will cause a significant increase in processing time, however, the relationships are uncertain. Classification of the job is based only on processing time. The optional LIMITS control card must be included if the THESR job is classified as LONG or X-LONG. The THESR description for a 'Job Resource Form" includes one activity, and one or two magnetic tapes, depending on whether the SAVE parameter was specified as "YES", plus, of course, the total estimated processing time.

```
Card Column:   1      8       16

               $      IDENT   project number, name
               $      SELECT  2632-SEDELØW/THESR
(optional)     $      LIMITS  time
INPUT TAPE     $      TAPE    02,X2DD,,tape number,,tape label,IN
(optional      $      TAPE    03,X3DD,,tape number,,tape label,ØUT
 OUTPUT TAPE)
               $      INCØDE  IBMF
                      Parameter Card 1
                      Header and Formats if specified on card 1
               $      DATA    01,IBMF,CØPYD
                      Thesaurus input deck
               $      ENDCØPY
               $      ENDJØB
               ***EØF
```

Job Deck 8 (THESR)

VIII.D. Description of Output (cf. the sample output from THESR in Section X.D.4.)

The printed output from THESR is the threshold specified, multi-level thesaurus whose threshold and depth of extension is specified on the parameter card. The output has the tree structure of the 3-level tree in Figure 8.

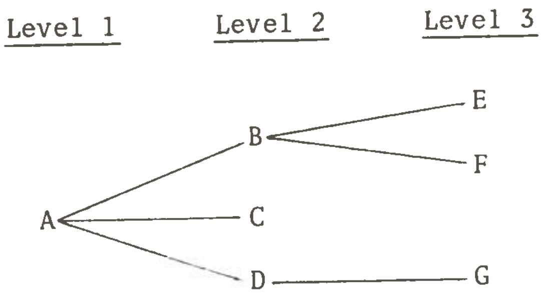


Fig. 8

The points labeled A,B,C,D,E,F,G in the tree in Figure 8 are called nodes. E,F,C and G are terminal nodes; A,B,C and D nonterminal; and A is a root node.

In the output from THESR, all nodes have matchcount frequency ≥ 1. In addition, all root nodes have matchcount frequency ≥ N where N is the threshold value specified on the parameter card. Branching can emanate only from those nodes which match primary words in the input thesaurus. An associate word with matchcount ≥ 1 which does not match a primary word will appear in the output tree as a terminal node. A node which matches a primary word is not allowed to branch to the next level if it matches a node which already appears in, or anywhere in the output above, the path from that node to its root node (the path from 'E' to its root node is the sequence E,B,A in Fig. 8). A node can be a terminal node, therefore, for one of three reasons:

1) it is a node on the kth level where k is the number following 'LEVEL= ' on the parameter card;

2) it is an associate word which does not 'match' a primary word;

3) it is a primary word which has already appeared in the tree.

Listed below each node in the output tree and enclosed in parenthesis are text words which match but are not identical to the node. If the node is identical to a word appearing in the text it is marked with an *. Also given in the output are the matchcounts and matchcount frequencies for each node in the tree.

IX. SAMPLER

IX. A. Capabilities

The SAMPLER program is designed to produce printed "samples" of data from VIA-generated magnetic tape files. Although it may be used to print an entire data file, it is primarily intended to provide the user with a means of checking intermediate data in the VIA sequence. For this reason, "samples" may begin at any point in the data file and contain any specified number of records. (Each word or punctuation symbol token with its associated indexing information is referred to as one "record" in the data file.) Multiple samples may be obtained from a data file, but only one data file may be processed at a time. The data will not be altered in any way by SAMPLER.

IX.B. Parameter and Title Cards

IX.B.1. Parameter Cards

A data file produced by any VIA program except THESR will be accepted for processing, so the file "type" must be identified for SAMPLER. This is accomplished by naming the primary VIA program which generated it (INDEX, PREFIX, or SUFFIX). SORT and SELECT do not alter the data records, therefore do not change the file type. For example, an alphabetically ordered file which is ready to be introduced to the PREFIX program is still considered to be type "INDEX". SAMPLER headings will be printed for any of the available processing modes.

File type and processing mode are therefore identified on the first parameter card as follows:

$$\text{FILE TYPE=}\begin{Bmatrix}\text{INDEX}\\ \text{PREFIX}\\ \text{SUFFIX}\end{Bmatrix}\text{,MODE=}\begin{Bmatrix}\text{PROS}\\ \text{PLAY}\\ \text{SPOKE}\\ \text{POET}\\ \text{VPLAY}\\ \text{MILT}\end{Bmatrix}$$

Only one type and one mode may be specified. Spacing is irrelevant, but order is unalterable, and "equals" signs may not be omitted.

Parameter cards which follow the first card each contain specifications for one data "sample" which is to be printed. A "SAMPLE" parameter card is specified as follows:

SAMPLE BEGIN=*first location*,SIZE=*number of records*

"BEGIN" refers to the record location within the file which contains the first word or punctuation symbol token to be printed. This is the same as the "linear number" for that token only in the case of a text-ordered INDEX data file. "SIZE" refers to the number of sequential records to be printed for the sample. Again, spacing is irrelevant, but the order must not be changed and "equals" signs must appear. If the word "ALL" appears rather than a number to specify sample size, all records in the data file will be printed until the end of the file is encountered.

Multiple samples are obtained by using one "SAMPLE" card for each printed sample desired. The SAMPLER program will not "back up," therefore samples should be ordered sequentially and not allowed to overlap. If a "SAMPLE" card is encountered which specifies a beginning location smaller than the current location (previous sample size added to its beginning location), the program will stop, and none of the remaining samples will be processed. For example,

```
FILE TYPE = PREFIX, MØDE = PLAY
SAMPLE BEGIN = 1, SIZE = 100
SAMPLE BEGIN = 201, SIZE = 100
```

will cause SAMPLER to print headings for a "PREFIX" file in PLAY mode, and two samples will be printed. The first printed sample will contain the first 100 records of the data file. The second will contain another 100 sequential records beginning with record location 201. If, however, the two "SAMPLE" cards appear in reverse order, only one sample will be printed (100 records beginning with record number 201).

Similarly, the "SAMPLE" cards

```
SAMPLE BEGIN = 1, SIZE = 100
SAMPLE BEGIN = 50, SIZE = 100
```

will cause SAMPLER to ignore the second sample because of overlapping. That is, when the second "SAMPLE" card is encountered, the beginning location specified (50) is smaller than the current location (101).

IX.B.2. Title Card

A title card should be included for each SAMPLER run. Information appearing on this card will be printed "as is" with the sample heading on each page. It must immediately precede the parameter cards in the job deck; omission or misplacement of this card will cause program failure. If no title is desired, a blank card should be inserted in its place.

IX.C. Printed Output

The printed sample headings will be produced according to processing mode and file type, and will include the user-supplied title described above. Samples will be numbered sequentially, and each labeled with its associated parameters (type and mode). Paging will occur within each sample, listing 50 lines of data per page (one line per record).

IX.D. Job Deck and Running Suggestions

The input deck for a SAMPLER run is illustrated as Job Deck 9.

```
Card Column:  1    8        16
_____________________________________________________________________

              $    IDENT    project number,name
              $    SELECT   2632-SEDELØW/SAMPLER
(optional)    $    LIMITS   time
              $    TAPE     02,X2DD,,tape number,,tape label,IN
              Title Card
              Parameter Cards
(optional)    $    TAPE     06,X6SD,,tape number,,tape label,ØUT
(optional)    $    SELECT   2632-SEDELØW/LIST
              $    ENDJØB
              ***EØF
_____________________________________________________________________
```

Job Deck 9 (SAMPLER)

Parameter cards are those described in Section IX.B.; control cards have been discussed in sections dealing with other VIA programs. The optional control cards are necessary only if more than 5000 lines of printed output is expected. In this case, the optional TAPE and SELECT card must both appear. The job is then classified as LONG or X-LONG according to the number of lines of output expected (as described in Section IV.C.2.). If more than 40,000 records are accessed (printed or skipped), the optional LIMITS card should be included, estimating .026 hours per 10,000 records.

X. EXAMPLE

X. A. INDEX

X.A.1. JOB DECK FOR INDEX

```
$       IDENT   0000,L3-LEWIS
$       SELECT  2632-SEDELOW/INDEX
$       INCODE  IBMF
PROC=PROS,PRINT=YES,STAGE=NO,DELIM=$
          INDEX OUTPUT -- GREAT BLUE HERON
$       LIMITS  02
$       TAPE    21,X2DD,,11111,,INDEX,OUT
$       DATA    20,IBMF,COPYD
THE GREAT BLUE HERON , LARGEST OF THE DARK HERONS , IS COMMON IN FRESH
.
.
.
THE LOUISIANA OFTEN WADES IN DEEP WATER .
$       ENDCOPY
$       ENDJOB
***EOF
```

X.A.2. SAMPLE INPUT TEXT

```
THE GREAT BLUE HERON , LARGEST OF THE DARK HERONS , IS COMMON IN FRESH
AND SALT WATER MARSHES . ITS HEAD IS LARGELY WHITE , UNDERPARTS BLUISH
TO BLACK .. IT IS COMMONLY SEEN FISHING IN SHALLOW WATERS WHERE FISH AN-
D OTHER WATER ANIMALS ARE ABUNDANT . $$$ IT USUALLY FREQUENTS SOUTHEAST-
ERN LAKES AND STREAMS WHERE ITS HUGE STICK NEST IS A FAMILIAR SIGHT .
$$$$ IT IS BIGGER THAN ITS UNCOMMON COUSIN , THE LOUISIANA HERON , AN
IRREGULAR VISITOR TO THE MISSISSIPPI RIVER DELTA , RECOGNIZED BY ITS WE-
LL-KNOWN CALL . $$$$
THE LOUISIANA OFTEN WADES IN DEEP WATER .
```

X.A.3. Sample Printed Output

```
PROGRAM INDEX                 INDEX OUTPUT -- GREAT BLUE HERON
VERSION - MARCH 21, 1971.


THE FOLLOWING OPTIONS HAVE BEEN SPECIFIED OR ASSUMED- TYPE OF PROCESSING --- PROS

PRINTING             --- YES,

SEQUENCE CHECKING    --- NO,

STAGE DIRECTIONS     --- NO,

DELIMITER USED       --- $
```

```
          INDEY CUTPUT -- GRFAT BLUE HEPON
                           (VOLUME--CHAPTER--PARAGRAPH--SENTENCE--WORD)

  THF                      (   1--   1--   1--   1--    1)
  GREAT                    (   1--   1--   1--   1--    2)
  BLUE                     (   1--   1--   1--   1--    3)
  HEPON                    (   1--   1--   1--   1--    4)
  ,                        (   1--   1--   1--   1--    5)
  LARGEST                  (   1--   1--   1--   1--    6)
  OF                       (   1--   1--   1--   1--    7)
  THE                      (   1--   1--   1--   1--    8)
  DARK                     (   1--   1--   1--   1--    9)
  HERONS                   (   1--   1--   1--   1--   10)
  ,                        (   1--   1--   1--   1--   11)
  IS                       (   1--   1--   1--   1--   12)
  COMMON                   (   1--   1--   1--   1--   13)
  IN                       (   1--   1--   1--   1--   14)
  FRESH                    (   1--   1--   1--   1--   15)
  AND                      (   1--   1--   1--   1--   16)
  SALT                     (   1--   1--   1--   1--   17)
  WATEP                    (   1--   1--   1--   1--   18)
  MAPSHFS                  (   1--   1--   1--   1--   19)
  .                        (   1--   1--   1--   1--   20)
  ITS                      (   1--   1--   1--   2--    1)
  HEAD                     (   1--   1--   1--   2--    2)
  IS                       (   1--   1--   1--   2--    3)
  LARGELY                  (   1--   1--   1--   2--    4)
  WHITF                    (   1--   1--   1--   2--    5)
  ,                        (   1--   1--   1--   2--    6)
  UNDERPARTS               (   1--   1--   1--   2--    7)
  BLUISH                   (   1--   1--   1--   2--    8)
  TO                       (   1--   1--   1--   2--    9)
  BLACK                    (   1--   1--   1--   2--   10)
  .                        (   1--   1--   1--   2--   11)
  IT                       (   1--   1--   2--   1--    1)
  IS                       (   1--   1--   2--   1--    2)
  COMMONLY                 (   1--   1--   2--   1--    3)
  SEEN                     (   1--   1--   2--   1--    4)
  FISHING                  (   1--   1--   2--   1--    5)
  IN                       (   1--   1--   2--   1--    6)
  SHALLOW                  (   1--   1--   2--   1--    7)
  WATERS                   (   1--   1--   2--   1--    8)
  WHERE                    (   1--   1--   2--   1--    9)
  FISH                     (   1--   1--   2--   1--   10)
  AND                      (   1--   1--   2--   1--   11)
  OTHEP                    (   1--   1--   2--   1--   12)
  WATER                    (   1--   1--   2--   1--   13)
  ANIMALS                  (   1--   1--   2--   1--   14)
  ARE                      (   1--   1--   2--   1--   15)
  ABUNDANT                 (   1--   1--   2--   1--   16)
  .                        (   1--   1--   2--   1--   17)
  IT                       (   1--   2--   1--   1--    1)
  USUALLY                  (   1--   2--   1--   1--    2)
  FREQUENTS                (   1--   2--   1--   1--    3)
  SOUTHEASTEPN             (   1--   2--   1--   1--    4)
  LAKES                    (   1--   2--   1--   1--    5)
  AND                      (   1--   2--   1--   1--    6)
  STREAMS                  (   1--   2--   1--   1--    7)
  WHERE                    (   1--   2--   1--   1--    8)
```

```
                INDEX OUTPUT -- GREAT BLUE HERON
                                (VOLUME--CHAPTER--PARAGRAPH--SENTENCE--WORD)

ITS                             (   1--    2--    1--    1--    9)
HUGE                            (   1--    2--    1--    1--   10)
STICK                           (   1--    2--    1--    1--   11)
NEST                            (   1--    2--    1--    1--   12)
IS                              (   1--    2--    1--    1--   13)
A                               (   1--    2--    1--    1--   14)
FAMILIAR                        (   1--    2--    1--    1--   15)
SIGHT                           (   1--    2--    1--    1--   16)
.                               (   1--    2--    1--    1--   17)
IT                              (   2--    1--    1--    1--    1)
IS                              (   2--    1--    1--    1--    2)
BIGGER                          (   2--    1--    1--    1--    3)
THAN                            (   2--    1--    1--    1--    4)
ITS                             (   2--    1--    1--    1--    5)
UNCOMMON                        (   2--    1--    1--    1--    6)
COUSIN                          (   2--    1--    1--    1--    7)
,                               (   2--    1--    1--    1--    8)
THE                             (   2--    1--    1--    1--    9)
LOUISIANA                       (   2--    1--    1--    1--   10)
HERON                           (   2--    1--    1--    1--   11)
,                               (   2--    1--    1--    1--   12)
AN                              (   2--    1--    1--    1--   13)
IRREGULAR                       (   2--    1--    1--    1--   14)
VISITOR                         (   2--    1--    1--    1--   15)
TO                              (   2--    1--    1--    1--   16)
THE                             (   2--    1--    1--    1--   17)
MISSISSIPPI                     (   2--    1--    1--    1--   18)
RIVER                           (   2--    1--    1--    1--   19)
DELTA                           (   2--    1--    1--    1--   20)
,                               (   2--    1--    1--    1--   21)
RECOGNIZED                      (   2--    1--    1--    1--   22)
BY                              (   2--    1--    1--    1--   23)
ITS                             (   2--    1--    1--    1--   24)
WELL-KNOWN                      (   2--    1--    1--    1--   25)
CALL                            (   2--    1--    1--    1--   26)
.                               (   2--    1--    1--    1--   27)
THE                             (   3--    1--    1--    1--    1)
LOUISIANA                       (   3--    1--    1--    1--    2)
OFTEN                           (   3--    1--    1--    1--    3)
WADES                           (   3--    1--    1--    1--    4)
IN                              (   3--    1--    1--    1--    5)
DEEP                            (   3--    1--    1--    1--    6)
WATER                           (   3--    1--    1--    1--    7)
.                               (   3--    1--    1--    1--    8)
```

```
PROGRAM INDEX                 INDEX OUTPUT -- GREAT BLUE HERON
VERSION - MARCH 21, 1971.


     9     CARDS INPUT

   126     RECORDS OUTPUT
                                       ***** END OF JOB *****
```

X.A.4. JOB DECK FOR SAMPLER

```
$      IDENT   0000,L3-LEWIS
$      SELECT  2632-SEDELOW/SAMPLER
$      INCODE  IBMF
         INDEX SAMPLER -- GREAT BLUE HERON
  FILE TYPE = SUFFIX, MODE = FRUS
  SAMPLE BEGIN = 1, SIZE = 50
$      TAPE    J2,X2DD,,11111,,INDEX,IN
$      ENDJOB
***EOF
```

X.A.5. Sample Output from SAMPLER

```
SAMPLE.......       1
FILE TYPE... INDEX                    INDEX SAMPLER -- GREAT BLUE HERON
MODE........  PROS


LINEAR NO   VOL  CHAPT   PARA   SENT   WIS   PAGE   LEN   WORD

        1     1      1      1      1     1      0     3   THE
        2     1      1      1      1     2      0     5   GREAT
        3     1      1      1      1     3      0     4   BLUE
        4     1      1      1      1     4      0     5   HERON
        5     1      1      1      1     5      0     1   ,
        6     1      1      1      1     6      0     7   LARGEST
        7     1      1      1      1     7      0     2   OF
        8     1      1      1      1     8      0     3   THE
        9     1      1      1      1     9      0     4   DARK
       10     1      1      1      1    10      0     6   HERONS
       11     1      1      1      1    11      0     1   ,
       12     1      1      1      1    12      0     2   IS
       13     1      1      1      1    13      0     6   COMMON
       14     1      1      1      1    14      0     2   IN
       15     1      1      1      1    15      0     5   FRESH
       16     1      1      1      1    16      0     3   AND
       17     1      1      1      1    17      0     4   SALT
       18     1      1      1      1    18      0     5   WATER
       19     1      1      1      1    19      0     7   MARSHES
       20     1      1      1      1    20      0     1   .
       21     1      1      1      2     1      0     3   ITS
       22     1      1      1      2     2      0     4   HEAD
       23     1      1      1      2     3      0     2   IS
       24     1      1      1      2     4      0     7   LARGELY
       25     1      1      1      2     5      0     5   WHITE
       26     1      1      1      2     6      0     1   ,
       27     1      1      1      2     7      0    10   UNDERPARTS
       28     1      1      1      2     8      0     6   BLUISH
       29     1      1      1      2     9      0     2   TO
       30     1      1      1      2    10      0     5   BLACK
       31     1      1      1      2    11      0     1   .
       32     1      1      2      1     1      0     2   IT
       33     1      1      2      1     2      0     2   IS
       34     1      1      2      1     3      0     8   COMMONLY
       35     1      1      2      1     4      0     4   SEEN
       36     1      1      2      1     5      0     7   FISHING
       37     1      1      2      1     6      0     2   IN
       38     1      1      2      1     7      0     7   SHALLOW
       39     1      1      2      1     8      0     6   WATERS
       40     1      1      2      1     9      0     5   WHERE
       41     1      1      2      1    10      0     4   FISH
       42     1      1      2      1    11      0     3   AND
       43     1      1      2      1    12      0     5   OTHER
       44     1      1      2      1    13      0     5   WATER
       45     1      1      2      1    14      0     7   ANIMALS
       46     1      1      2      1    15      0     3   ARE
       47     1      1      2      1    16      0     8   ABUNDANT
       48     1      1      2      1    17      0     1   .
       49     1      2      1      1     1      0     2   IT
       50     1      2      1      1     2      0     7   USUALLY
```

```
NORMAL TERMINATION
  1 SAMPLES PROCESSED
```

X.B. PREFIX

X.B.1. JOB DECK FOR SORT

```
$       IDENT    0000,L3-LEWIS
$       SELECT   2632-SEDELOW/SORT
$       TAPE     SA,X1DD,,11111,,INDEX,IN
$       TAPE     SZ,X2DD,,11112,,ISORT,OUT
$       167PK    S1,X3R,R,R0001,SCRATCH,PRIVATE,0/1
$       ENDJOB
***EOF
```

X.B.2. JOB DECK FOR PREFIX

```
$       IDENT    0000,L3-LEWIS
$       SELECT   2632-SEDELOW/PREFIX
$       TAPE     15,X15DD,,11112,,ISORT,IN
$       TAPE     20,X20DD,,11113,,PREFIX,OUT
$       ENDJOB
***EOF
```

X.8.3. JOB DECK FOR SAMPLER

```
$       IDENT    0000,L3-LEWIS
$       SELECT   2632-SFDELOW/SAMPLER
$       INCODE   IBMF
          PREFIX SAMPLER -- GREAT BLUE HERON
  FILE TYPE = PREFIX, MODE = PROS
  SAMPLE BEGIN = 1, SIZE = 50
$       TAPE     02,X2DD,,11113,,PREFIX,IN
$       ENDJOB
***EOF
```

X.B.4. Sample Output from SAMPLER

```
SAMPLE......      1
FILE TYPE...PREFIX                    PREFIX SAMPLER -- GREAT BLUE HERON
MODE........    PROS


                                                          PREF    WORD
LINEAR NO   VOL  CHAPT    PARA    SENT     WIS    PAGE    LEN     LEN     PREFIX   WORD

      62     1      2       1       1       14      0       0       1               A
      47     1      1       2       1       16      0       0       8               ABUNDANT
      78     2      1       1       1       13      0       0       2               AN
      16     1      1       1       1       16      0       0       3               AND
      42     1      1       2       1       11      0       0       3               AND
      54     1      2       1       1        6      0       0       3               AND
      45     1      1       2       1       14      0       0       7               ANIMALS
      46     1      1       2       1       15      0       0       3               ARE
      68     2      1       1       1        3      0       0       6               BIGGER
      30     1      1       1       2       10      0       0       5               BLACK
      28     1      1       1       2        8      0       0       6               BLUISH
       3     1      1       1       1        3      0       0       4               BLUE
      88     2      1       1       1       23      0       0       2               BY
      91     2      1       1       1       26      0       0       4               CALL
      13     1      1       1       1       13      0       0       6               COMMON
      34     1      1       2       1        3      0       0       8               COMMONLY
      72     2      1       1       1        7      0       0       6               COUSIN
       9     1      1       1       1        9      0       0       4               DARK
      98     3      1       1       1        6      0       0       4               DEEP
      85     2      1       1       1       20      0       0       5               DELTA
      63     1      2       1       1       15      0       0       8               FAMILIAR
      41     1      1       2       1       10      0       0       4               FISH
      36     1      1       2       1        5      0       0       7               FISHING
      15     1      1       1       1       15      0       0       5               FRESH
      51     1      2       1       1        3      0       0       9               FREQUENTS
       2     1      1       1       1        2      0       0       5               GREAT
      22     1      1       1       2        2      0       0       4               HEAD
      76     2      1       1       1       11      0       0       5               HERON
       4     1      1       1       1        4      0       0       5               HERON
      10     1      1       1       1       10      0       0       6               HERONS
      58     1      2       1       1       10      0       0       4               HUGE
      97     3      1       1       1        5      0       0       2               IN
      14     1      1       1       1       14      0       0       2               IN
      37     1      1       2       1        6      0       0       2               IN
      79     2      1       1       1       14      0       2       7       IR      REGULAR
      61     1      2       1       1       13      0       0       2               IS
      67     2      1       1       1        2      0       0       2               IS
      12     1      1       1       1       12      0       0       2               IS
      23     1      1       1       2        3      0       0       2               IS
      33     1      1       2       1        2      0       0       2               IS
      32     1      1       2       1        1      0       0       2               IT
      66     2      1       1       1        1      0       0       2               IT
      49     1      2       1       1        1      0       0       2               IT
      57     1      2       1       1        9      0       0       3               ITS
      70     2      1       1       1        5      0       0       3               ITS
      89     2      1       1       1       24      0       0       3               ITS
      21     1      1       1       2        1      0       0       3               ITS
      48     1      1       2       1       17      0       0       1               .
      92     2      1       1       1       27      0       0       1               .
      65     1      2       1       1       17      0       0       1               .
```

```
NORMAL TERMINATION
  1 SAMPLES PROCESSED
```

X.C. SUFFIX

X.C.1. JOB DECK FOR SORT

```
$       IDENT    0020,L3-LEWIS
$       SELECT   2632-SEDELOW/SORT
$       TAPE     SA,X1DD,,11113,,PREFIX,IN
$       TAPE     SZ,X2DD,,11114,,PSORT,OUT
$       167PK    S1,X3R,P,R0001,SCRATCH,PRIVATE,0/10
$       ENDJOB
***EOF
```

X.C.2. JOB DECK FOR SUFFIX

```
$       IDENT    0J00,L3-LEWIS
$       SELECT   2632-SEDELOW/SUFFIX
$       INCODE   IBMF
FUNCT=YES,PUNCT=YES
PRINT=YESJ005800E8
$       SELECT   2632-SEDELOW/PARAMS
$       INCODE   IBMF
          SUFFIX OUTPUT -- GREAT BLUE HERON
$       TAPE     15,X15DD,,11114,,PSORT,IN
$       TAPE     13,X13DD,,11115,,SUFFIX,OUT
$       ENDJOB
***EOF
```

X.C.3. Sample Printed Output

```
                             PARAMETER SPECIFICATIONS

FUNCT SET TO YES
PUNCT SET TO YES
PRINT SET TO YES
OUTPUT PAGE LENGTH IS     58
FUNCT REPORT PAGE LENGTH IS     58
```

```
                         SUFFIX OUTPUT -- GREAT BLUE HERON                                              PAGE      1

  FREQ OF  MATCH                                      VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD
   OCCUR   COUNT  PREFIX   WORD                       NUM  NUMB NUMB NUMB NUMB  NUM NUMB NUMB NUMB NUMB  NUM NUMB NUMB NUMB NUMB

(    1       1)

     1       1             ABUNDANT                  ( 1    1    2    1    15)(

(    1       2)

     1       2             ANIMALS                   ( 1    1    2    1    14)(

(    1       3)

     1       3             BIGGER                    ( 2    1    1    1     3)(

(    1       4)

     1       4             BLACK                     ( 1    1    1    2    13)(

(    2       5)

     1       5             BLUISH                    ( 1    1    1    2     8)(
     1       5             BLUE                      ( 1    1    1    1     3)(

(    1       6)

     1       6             CALL                      ( 2    1    1    1    26)(

(    3       7)

     1       7             COMMON                    ( 1    1    1    1    13)(
     1       7 UN          COMMON                    ( 2    1    1    1     6)(
     1       7             COMMONLY                  ( 1    1    2    1     3)(

(    1       8)

     1       8             COUSIN                    ( 2    1    1    1     7)(

(    1       9)

     1       9             DARK                      ( 1    1    1    1     9)(
                         SUFFIX OUTPUT -- GREAT BLUE HERON

  FREQ OF  MATCH                                  VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD
   OCCUR   COUNT  PREFIX   WORD                   NUM  NUMB NUMB NUMB NUMB  NUM NUMB NUMB NUMB NUMB  NUM NUMB NUMB NUMB NUMB


 (    1       10)

      1       10           DEEP                 (  3     1    1    1    6)(

 (    1       11)

      1       11           DELTA                (  2     1    1    1   20)(

 (    1       12)

      1       12           FAMILIAR             (  1     2    1    1   15)(

 (    2       13)

      1       13           FISH                 (  1     1    2    1   1[illegible])(
      1       13           FISHING              (  1     1    2    1    5)(

 (    1       14)
 (    1       15)

      1       14           FRESH                (  1     1    1    1   15)(
      1       15           FREQUENTS            (  1     2    1    1    3)(

 (    1       16)

      1       16           GREAT                (  1     1    1    1    2)(

 (    1       17)

      1       17           HEAD                 (  1     1    1    2    2)(

 (    3       18)

      1       18           HERON                (  1     1    1    1    4)(
      1       18           HERONS               (  1     1    1    1   10)(
      1       18           HERON                (  2     1    1    1   11)(
```

                     SUFFIX OUTPUT -- GREAT BLUE HERON

 FPEQ OF MATCH   VOL CHAP PARA SENT WORD VOL CHAP PARA SENT WORD VOL CHAP PARA SENT WORD
  OCCUR   COUNT PREFIX   WORD   NUM NUM3 NUMB NUMB NUM3 NUM NUMB NUMB NUMB NUMB NUM NUMB NUMB NUMB NUM3

(1   19)

     1   19   HUGE   (1   2   1   1   10)(

(1   20)

     1   20 WELL-   KNOWN   (2   1   1   1   25)(

(1   21)

     1   21   LAKES   (1   2   1   1   5)(

(1   22)
(1   23)

     1   22   LARGEST   (1   1   1   1   5)(
     1   23   LARGELY   (1   1   1   2   4)(

(2   24)

     2   24   LOUISIANA   (2   1   1   1   10)(3   1   1   1   2)(

(1   25)

     1   25   MARSHES   (1   1   1   1   19)(

(1   26)

     1   26   MISSISSIPPI   (2   1   1   1   18)(

(1   27)

     1   27   NEST   (1   2   1   1   12)(

(1   28)

     1   28   RECOGNIZED   (2   1   1   1   22)(
```
```

```
                         SUFFIX OUTPUT -- GREAT BLUE HERON

 FREQ OF  MATCH                                     VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD
  OCCUR   COUNT  PREFIX   WORD                      NUM  NUMB NUMB NUMB NUMB NUM  NUMB NUMB NUMB NUMB NUM  NUMB NUMB NUMB NUMB

(    1       29)

     1       29 IR        REGULAR                   ( 2    1    1    1   14)(

(    1       30)

     1       30           RIVER                     ( 2    1    1    1   19)(

(    1       31)

     1       31           SALT                      ( 1    1    1    1   17)(

(    1       32)

     1       32           SEEN                      ( 1    1    2    1    4)(

(    1       33)

     1       33           SHALLOW                   ( 1    1    2    1    7)(

(    1       34)

     1       34           SIGHT                     ( 1    2    1    1   16)(

(    1       35)

     1       35           SOUTHEASTERN              ( 1    2    1    1    4)(

(    1       36)

     1       36           STICK                     ( 1    2    1    1   11)(

(    1       37)

     1       37           STREAMS                   ( 1    2    1    1    7)(
                         SUFFIX OUTPUT -- GREAT BLUE HERON

 FREQ OF  MATCH                                      VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD VOL  CHAP PARA SENT WORD
  OCCUR   COUNT  PREFIX   WORD                       NUM  NUMB NUMB NUMB NUMB  NUM NUMB NUMB NUMB NUMB  NUM NUMB NUMB NUMB NUMB

(    1        38)

     1        38          UNDERPARTS                ( 1    1    1    2    7)(

(    1        39)

     1        39          USUALLY                   ( 1    2    1    1    2)(

(    1        40)

     1        40          VISITOR                   ( 2    1    1    1   15)(

(    1        41)

     1        41          WADES                     ( 3    1    1    1    4)(

(    4        42)

     3        42          WATER                     ( 1    1    2    1   13)( 1    1    1    1   18)( 3    1    1    1    7)
     1        42          WATERS                    ( 1    1    2    1    8)(

(    1        43)

     1        43          WHITE                     ( 1    1    1    2    5)(
          PUNCTUATION AND FUNCTION WORD LISTINGS

                              FUNCTION WORDS AND PUNCTUATION --- SUMMARY

                    THERE WERE    12 PUNCTUATION RECORDS OUTPUT (MATCH COUNT = 99999)

                                             AND

                    THERE WERE    35 FUNCTION RECORDS OUTPUT (MATCH COUNT = 99998)

    WORD               TYPE-COUNT


  A                          1
  AN                         1
  AND                        3
  ARE                        1
  BY                         1
  IN                         3
  IS                         5
  IT                         3
  ITS                        4
  .                          6
  OF                         1
  OFTEN                      1
  OTHER                      1
  THAN                       1
  THE                        5
  TO                         2
  WHERE                      2
  ,                          6
```

X.C.4. JOB DECK FOR SAMPLER

```
$       IDENT    0000,L3-LEWIS
$       SELECT   2632-SEDELOW/SAMPLER
$       INCODE   IBMF
          SUFFIX SAMPLER -- GREAT BLUE HERON
  FILE TYPE = SUFFIX, MODE = PROS
  SAMPLE BEGIN = 1, SIZE = 50
$       TAPE     02,X2DD,,11115,,SUFFIX,IN
$       ENDJOB
***EOF
```

X.C.5. Sample Output from THESR

```
SAMPLE......      1
FILE TYPE...SUFFIX               SUFFIX SAMPLER -- GREAT BLUE HERON
MODE........   PFOS


                                                          MATC  TYPE  PREF  WORD
LINEAR NO  VOL  CHAPT   PARA   SENT    WIS   PAGE  MATCNT  FREQ  FREQ  LEN   LEN   PREFIX   WORD

     47     1     1      2      1      16     0      1      1     1     0     8            ABUNDANT
     45     1     1      2      1      14     0      2      1     1     0     7            ANIMALS
     68     2     1      1      1       3     0      3      1     1     0     6            BIGGER
     30     1     1      1      2      10     0      4      1     1     0     5            BLACK
     28     1     1      1      2       8     0      5      2     1     0     6            BLUISH
      3     1     1      1      1       3     0      5      2     1     0     4            BLUE
     91     2     1      1      1      26     0      6      1     1     0     4            CALL
     13     1     1      1      1      13     0      7      3     1     0     6            COMMON
     71     2     1      1      1       6     0      7      3     1     2     6     UN     COMMON
     34     1     1      2      1       3     0      7      3     1     0     8            COMMONLY
     72     2     1      1      1       7     0      8      1     1     0     6            COUSIN
      9     1     1      1      1       9     0      9      1     1     0     4            DARK
     98     3     1      1      1       6     0     10      1     1     0     4            DEEP
     85     2     1      1      1      20     0     11      1     1     0     5            DELTA
     63     1     2      1      1      15     0     12      1     1     0     8            FAMILIAR
     41     1     1      2      1      10     0     13      2     1     0     4            FISH
     36     1     1      2      1       5     0     13      2     1     0     7            FISHING
     15     1     1      1      1      15     0     14      1     1     0     5            FRESH
     51     1     2      1      1       3     0     15      1     1     0     9            FREQUENTS
      2     1     1      1      1       2     0     16      1     1     0     5            GREAT
     22     1     1      1      2       2     0     17      1     1     0     4            HEAD
      4     1     1      1      1       4     0     18      3     1     0     5            HERON
     10     1     1      1      1      10     0     18      3     1     0     6            HERONS
     76     2     1      1      1      11     0     18      3     1     0     5            HERON
     58     1     2      1      1      10     0     19      1     1     0     4            HUGE
     90     2     1      1      1      25     0     20      1     1     5     5     WELL-  KNOWN
     53     1     2      1      1       5     0     21      1     1     0     5            LAKES
      6     1     1      1      1       6     0     22      1     1     0     7            LARGEST
     24     1     1      1      2       4     0     23      1     1     0     7            LARGELY
     75     2     1      1      1      10     0     24      2     2     0     9            LOUISIANA
     94     3     1      1      1       2     0     24      2     2     0     9            LOUISIANA
     19     1     1      1      1      19     0     25      1     1     0     7            MARSHES
     83     2     1      1      1      18     0     26      1     1     0    11            MISSISSIPPI
     60     1     2      1      1      12     0     27      1     1     0     4            NEST
     87     2     1      1      1      22     0     28      1     1     0    10            RECOGNIZED
     79     2     1      1      1      14     0     29      1     1     2     7     IR     REGULAR
     84     2     1      1      1      19     0     30      1     1     0     5            PIVER
     17     1     1      1      1      17     0     31      1     1     0     4            SALT
     35     1     1      2      1       4     0     32      1     1     0     4            SEEN
     38     1     1      2      1       7     0     33      1     1     0     7            SHALLOW
     64     1     2      1      1      16     0     34      1     1     0     5            SIGHT
     52     1     2      1      1       4     0     35      1     1     0    12            SOUTHEASTERN
     59     1     2      1      1      11     0     36      1     1     0     5            STICK
     55     1     2      1      1       7     0     37      1     1     0     7            STREAMS
     27     1     1      1      2       7     0     38      1     1     0    10            UNDERPARTS
     50     1     2      1      1       2     0     39      1     1     0     7            USUALLY
     80     2     1      1      1      15     0     40      1     1     0     7            VISITOR
     96     3     1      1      1       4     0     41      1     1     0     5            WADES
     44     1     1      2      1      13     0     42      4     3     0     5            WATER
     18     1     1      1      1      18     0     42      4     3     0     5            WATER
```

```
NORMAL TERMINATION
  1 SAMFLES PROCESSED
```

X.D. THESR

X.D.1. JOB DECK FOR SORT

```
$       IDENT    0000,L3-LEWIS
$       SELECT   2632-SEDELOW/SORTM
$       TAPE     SA,X1DD,,11115,,SUFFIX,IN
$       TAPE     SZ,X2DD,,11116,,SSORT,OUT
$       167PK    S1,X3R,R,R0001,SCRATCH,PRIVATE,0/10
$       ENDJOB
***EOF
```

X.D.2. JOB DECK FOR THESR

```
$       IDENT    0000,L3-LEWIS
$       SELECT   2632-SEDELOW/THESR
$       TAPE     02,X2DD,,11116,,SSORT,IN
$       INCODE   IBMF
LEVEL=5,SAVE=NO,THRESHOLD=1,FORMATS=NO
$       DATA     01,IBMF,COPYD
BIG                 ENORMOUS
.
.
.
WATER               RAIN
$       ENDCOPY
$       ENDJOB
***EOF
```

X.D.3. SAMPLE INPUT THESAURUS

BIG	ENORMOUS
BIG	GREAT
BIG	HUGE
BIG	LARGE
COLORATION	DARK
COLORATION	LIGHT
COLORATION	BLACK
COLORATION	WHITE
COLORATION	BLUE
COLORATION	RED
COMMON	ABUNDANT
COMMON	FREQUENT
COMMON	NUMEROUS
COMMON	USUAL
COMMON	PREVALENT
FAMILIAR	PREVALENT
FAMILIAR	RECOGNIZED
FAMILIAR	WELL-KNOWN
GREAT	GRAND
GREAT	LARGE
GREAT	SIZEABLE
GREAT	TALL
HEAD	CHIEF
HEAD	FRONT
HEAD	TOP
LARGE	BIG
LARGE	HUGE
LARGE	GIGANTIC
STREAM	RIVER
STREAM	CREEK
STREAM	GULLY
TALK	RELATE
TALK	TELL
TALK	COMMUNICATE
USUAL	FAMILIAR
USUAL	NORMAL
USUAL	COMMON
USUAL	NATURAL
USUAL	REGULAR
VIEW	SEE
VIEW	WATCH
WATER	STREAM
WATER	LAKE
WATER	OCEAN
WATER	RIVER
WATER	RAIN

X.D.4. Sample Printed Output

```
MATCNT  FREQ  LEVEL1          LEVEL2          LEVEL3          LEVEL4          LEVEL5


   7      7   *COMMON
             (UNCOMMON       )
             (COMMONLY       )

   1      1                  *ABUNDANT

  39      1                   USUAL
                             (USUALLY        )

  12      1                                  *FAMILIAR

  28      1                                                  *RECOGNIZED

   7      3                                  *COMMON
                                             (UNCOMMON       )
                                             (COMMONLY       )

  29      1                                   REGULAR
                                             (IRREGULAR      )

  12      1   *FAMILIAR

  28      1                  *RECOGNIZED

  16      1   *GREAT

  17      1   *HEAD

  37      1    STREAM
             (STREAMS        )

  36      1                  *RIVER

  39      1    USUAL
             (USUALLY        )

  12      1                  *FAMILIAR

  28      1                                  *RECOGNIZED

   7      3                  *COMMON
                             (UNCOMMON       )
                             (COMMONLY       )

   1      1                                  *ABUNDANT

  39      1                                   USUAL
                                             (USUALLY        )

  29      1                   REGULAR
                             (IRREGULAR      )
```

```
37       1                   STPFAM
                            (STREAMS           )

30       1                                     *RIVER

21       1                   LAKE
                            (LAKES             )

30       1                  *RIVER
```

APPENDIX A

Prefix Data List

APPENDIX A
PREFIX DATA LIST

```
A           INCLUD                           (NOT)
'ABED'               'AGLEAM'                  'AROUSE'              'ASYLLABIC'
'ABLAZE'             'AGLIMMER'                'ASEETHE'             'ASYMMETR'
'ABLOOM'             'AGLINT'                  'ASEXUAL'             'ASYNCHRON'
'ABLOW'              'AGLISTEN'                'ASHIMMER'            'ASYNTACTIC'
'ABLUSH'             'AGLITTER'                'ASHINE'              'ATHEIS'
'ABOIL'              'AGLOW'                   'ASHIVER'             'ATHIRST'
'ACENTRIC'           'AGROUND'                 'ASHORE'              'ATHRILL'
'ACHROM'             'AHISTORIC'               'ASIDE'               'ATILT'
'ACRITICAL'          'AHOLD'                   'ASKEW'               'ATINGLE'
'ACYCLIC'            'AHORSE'                  'ASLANT'              'ATIPTOE'
'ADANCE'             'AHUM'                    'ASLEEP'              'ATOP'
'ADANGLE'            'AHUNT'                   'ASLOPE'              'ATREMBLE'
'ADREAM'             'AKIN'                    'ASOCIAL'             'ATWIST'
'ADRIFT'             'ALIGHT'                  'ASPHERICAL'          'ATWITTER'
'AFAR'               'ALIKE'                   'ASPRAWL'             'ATYPI'
'AFIELD'             'ALIT'                    'ASPREAD'             'AVOUCH'
'AFIRE'              'ALONE'                   'ASQUINT'             'AVOW'
'AFLAME'             'ALOUD'                   'ASTARE'              'AWAKE'
'AFLOAT'             'AMASS'                   'ASTATIC'             'AWASH'
'AFLOW'              'AMID'                    'ASTIR'               'AWEARY'
'AFLUTTER'           'AMORAL'                  'ASTRADDLE'           'AWHEEL'
'AFOOT'              'APERIODIC'               'ASTRAY'              'AWHIRL'
'AFOUL'              'APHONIC'                 'ASWARM'              'AWING'
'AGAPE'              'APLENTY'                 'ASWAY'               'AWOKE'
'AGAZE'              'ARIPPLE'                 'ASWIRL'              'AWORK'
'AGLARE'             'ARISE'

AB          INCLUD           AWAY FROM
'ABAXIAL'            'ABERRANT'                'ABNEGATE'            'ABNORM'

AC          INCLUD           VAR.OF 'AD'
'ACCEDE'             'ACCOMPT'                 'ACCREDIT'            'ACCURS'
'ACCLAIM'            'ACCOUNT'                 'ACCRESCENT'          'ACCUSTOM'
'ACCLIMAT'           'ACCOUPL'                 'ACCUMULAT'           'ACQUIESC'
'ACCOMPANY'

AD          INCLUD           AWAY FROM
'ADJOIN'             'ADMIX'

AERO        EXCLUD                           AIR
'AEROBATIC'          'AEROGRAM'                'AEROMANC'            'AEROSOL'
'AERODROME'          'AEROLIT'                 'AERONAUT'

AFORE       EXCLUD           BEFORE
'AFOREHAND'

AFTER       EXCLUD           AFTER
'AFTER '             'AFTERGAME'               'AFTERMATH'           'AFTERWARD'
'AFTERCLAP'

AG          INCLUD           TOWARD (TENDENCY,DIRECTION,ADDITION)
'AGGLOMERATE'        'AGGRIEVE'
```

APPENDIX A
PREFIX DATA LIST

```
AL            INCLUD                                VAR OF AD,TOWARD TENDENCY,DIRECTION,ADD
'ALLUR'

ALL-          EXCLUD                ALL

ALLO          INCLUD                                OTHER
'ALLOGRAPH'

ALTI          EXCLUD                                HIGH
'ALTIMETRY'                'ALTITUDE'

AMBI          INCLUD                BOTH
'AMBIDEXT'                 'AMBILATERAL'

AMPHI         INCLUD                                TWO,BOTH,ON BOTH SIDES
'AMPHITHEATER'

AN            INCLUD                NOT,WITHOUT,LACKING,VAR.OF 'AD',VAR.OF 'ANA'←UP,←
'ANALPHABETIC'

ANDRO         INCLUD
'ANDROCENTRIC'             'ANDROPHOBIA'

ANEMO         INCLUD                                WIND
'ANEMOGRA'                 'ANEMOMET'

ANGLO         EXCLUD                                ENGLISH
'ANGLOPHIL'                'ANGLOPHOB'

ANT           INCLUD                                VAR OF 'ANTI' AGAINST
'ANTACID'                  'ANTARCTIC'

ANTE          INCLUD                BEFORE
'ANTE-CHRISTIAN'           'ANTE-WAR'               'ANTEMERIDIAN'         'ANTEPRANDIAL'
'ANTE-DAWN'                'ANTECHAMBER'            'ANTEMUNDANE'          'ANTEPROHIBITION'
'ANTE-MARRIAGE'            'ANTEDAT'                'ANTENATAL'            'ANTEROOM'
'ANTE-SPRING'              'ANTEDILUVIA'            'ANTENUMBER'           'ANTETYPE'
'ANTE-TASTE'               'ANTEHISTORIC'           'ANTENUPTIAL'

ANTHROPO      INCLUD                HUMAN
'ANTHPOPOCENTRI'           'ANTHROPOGEN'            'ANTHROPOGEOGRAPH'

ANTI          EXCLUD                AGAINST,OPPOSITE OF
'ANTIBODY'                 'ANTIDOTE'               'ANTIPATH'             'ANTIQU'
'ANTIC '                   'ANTINOMY'               'ANTIPOD'              'ANTITYP'
'ANTICIPA'

AP            INCLUD                VAR.OF AD,VAR.OF APO ← AWAY,DIFFERENT,FROM
'APPEND '                  'APPLY'                  'APPOS'                'APPRESS'
'APPERTAIN'

AR            INCLUD                                VAR OF AD BEFORE 'R'
'ARREAR'
```

APPENDIX A
PREFIX DATA LIST

```
ARCH          INCLUD                           (CHIEF)
'ARCH-ENEMY'             'ARCH-TRAITOR'           'ARCHDEACON'             'ARCHHERE'
'ARCH-FOE'               'ARCH-VERSIFIER'         'ARCHDIOCESE'            'ARCHPRESBYTER'
'ARCH-HERE'              'ARCH-VILLIAN'           'ARCHDU'                 'ARCHPRIEST'
'ARCH-LIAR'              'ARCHANGEL'              'ARCHENEMY'              'ARCHSEE'
'ARCH-OPPONENT'          'ARCHBISHOP'             'ARCHFIEND'              'ARCHVILLIAN'
'ARCH-POET'              'ARCHCHANCEL'

ARCHE         INCLUD                           (PRIMITIVE)
'ARCHETYP'

ARCHI         INCLUD                           (CHIEF)
'ARCHIDIACONAL'          'ARCHIEPISCOPA'

AT            INCLUD                           VAR OF AD
'ATTEMPER'               'ATTRACT'                'ATTRIBUT'               'ATTUN'

ATMO          INCLUD                           AIR
'ATMOSPHER'

AUDIO         EXCLUD              AUDITORY
'AUDIOGEN'               'AUDIOLOG'               'AUDION'                 'AUDIOPHIL'

AUTO          EXCLUD                           (SELF,SAME)
'AUTOLOG'                'AUTONOM'                'AUTOCLAVE'              'AUTOGRAPH'
'AUTOMAT'                'AUTONYM'                'AUTOCHTHON'             'AUTOPSY'
'AUTOMETRY'

BACK          EXCLUD              BACK
'BACK '                  'BACKING'                'BACKSLIDE'              'BACKSWEPT'
'BACKE'                  'BACKLOG'                'BACKSTAB'               'BACKWARD'

BE            INCLUD              COVER,TO MAKE+,TO DUB,PROVIDED WITH,NO MEANING
'BE-NIGHTMARED'          'BEDRIVEL'               'BELITTL'                'BESLOBBER'
'BEBOOTED'               'BEFLAG'                 'BEMEAN'                 'BESPANGL'
'BECHARM'                'BEFLEA'                 'BEMEDALLED'             'BESPECTACLED'
'BECLOUD'                'BEFLOWER'               'BEMIRE'                 'BESPREAD'
'BECRAWL'                'BEFOG'                  'BEMOAN'                 'BESTRADDL'
'BECRIPPL'               'BEFOOL'                 'BEMOCK'                 'BESTRID'
'BEDABBL'                'BEFOUL'                 'BENIGHT'                'BESTROD'
'BEDASH'                 'BEFRIEND'               'BEPAINT'                'BESTUD'
'BEDAUB'                 'BEGEM'                  'BERASCAL'               'BETHINK'
'BEDAVID'                'BEGLAMOUR'              'BERHYM'                 'BETHOUGHT'
'BEDAZZL'                'BEGLAR'                 'BERIBANDED'             'BEWAIL'
'BEDEVIL'                'BEGRIM'                 'BERIBBONED'             'BEWEEP'
'BEDEW'                  'BEHEAD'                 'BERIM'                  'BEWHISKERED'
'BEDIAMONDED'            'BEJEWEL'                'BEROGUE'                'BEWIGGED'
'BEDIM'                  'BEJESUIT'               'BESCRIBBL'              'BEWITCH'
'BEDRABBL'               'BEKNAVE'                'BESEIG'                 'BEWRITE'
'BEDRAGGL'               'BELATE'

BI            EXCLUD              TWO,TWICE,VAR.OF 'BIO'
'BIAS'                   'BIFARIOUS'              'BIMESTER'               'BIRTH'
'BIB '                   'BIFID'                  'BIN '                   'BIS '
```

APPENDIX A
PREFIX DATA LIST

```
'BIBLE'              'BIG '                'BINAL'              'BISCUIT'
'BIBLIO'             'BIGAM'               'BINARY'             'BISECT'
'BIBULOUS'           'BIGOT'               'BINAURAL'           'BISHOP'
'BICIPITAL'          'BIJOU'               'BIND'               'BIT '
'BID '               'BILF'                'BINOCLE'            'BITCH'
'BIDDABL'            'BILIOUS'             'BINOCULAR'          'BITE'
'BIDDEN'             'BILK'                'BIO'                'BITING'
'BIDDING'            'BILL'                'BIPOD'              'BITTEN'
'BIDDY'              'BILLET'              'BIRCH'              'BITTER'
'BIDE'               'BILLION'             'BIRD'               'BITUM'
'BIENNIAL'           'BILLOW'              'BIRR'               'BIZARRE'
'BIER'

BIBLIO      INCLUD          BOOK,BIBLE
'BIBLIOFILM'         'BIBLIOMANI'

BIN         INCLUD          TWO,TWO AT A TIME
'BINAURAL'           'BINOCULAR'

BOOK        EXCLUD                     BOOKIE,BOOKING,BOOKISH
'BOOK '              'BOOKI'              'BOOKLET'            'BOOKS '

BY          EXCLUD                     (ACCESSORY,PAST,SUBORDINATE,BY THE SIDE
'BY-BLOW'            'BYE'                'BYTE'               'BYWORD'
'BY-WORD'            'BYRNE'

BYE         INCLUD          VAR.OF 'BY'
'BYELAW'

CENTRI      EXCLUD                     (CENTER)
'CENTRIC'            'CENTRIFUGAL'        'CENTRIOLE'

CHRONO      INCLUD                     TIME
'CHRONOGRAPH'        'CHRONOMET'          'CHRONOSCOP'

CIS         INCLUD                     (NEAR SIDE OF)
'CISATLANTIC'        'CISLUNAR'

CO          INCLUD                     VAR OF COM,IN ASSOC WITH
'CO-ORDIN'           'COAXIAL'            'COEX'               'COOPERAT'
'CO-SIGN'            'CODEFEND'           'COFUNCTION'         'COORDIN'
'CO-STAR'            'COEDIT'             'COHEIR'             'COPARTNER'
'CO-WORKER'          'COEDUCAT'           'COINCIDEN'          'COTEMPOR'
'COACT'              'COEQUAL'            'COINSUR'            'COTENANT'
'COADJUT'            'COETERN'            'COMATE'             'COTERMINOUS'
'COADVENTUR'

COL         INCLUD                     VAR OF COM,WITH
'COLLABORA'          'COLLATERAL'         'COLLEAGUE'          'COLLOCAT'
'COLLAPS'

COM         INCLUD                     WITH,TOGETHER,IN ASSOC
'COMMEASUR'          'COMMUTA'            'COMMPATERN'         'COMPOSSIBLE'
'COMMEMOR'           'COMMUTUAL'          'COMPATRIOT'         'COMPROMIS'
'COMMINGLE'          'COMPACT'            'COMPEER'
```

APPENDIX A
PREFIX DATA LIST

```
CON            INCLUD                           VAR OF COM
'CONCAV'                  'CONFEDERA'               'CONJOIN'                'CONSTRAIN'
'CONCENTRIC'              'CONFIGUR'                'CONJUNCTION'            'CONSTRICT'
'CONCORPORAT'             'CONFORM'                 'CONJUNCTU'              'CONTEMPOR'
'CONCOURSE'               'CONFRONT'                'CONSEQUEN'              'CONTORTION'
'CONDENS'                 'CONGENIAL'               'CONSOLIDAT'             'CONTRACT '
'CONDESCEN'               'CONGENITAL'

COUNTER        EXCLUD                           (OPPOSITE)
'COUNTERCHANGE'           'COUNTERMAN'              'COUNTERPART'            'COUNTERWORD'
'COUNTERFEIT'

DE             INCLUD                           SEPARAT,PRIVATION,REMOV,DESCENT,REVERSA
'DE-EMPHASI'              'DEDUCT'                  'DEHYDRAT'               'DEMORALI'
'DAERAT'                  'DEFAC'                   'DEICE'                  'DEMOUNT'
'DEBASE'                  'DEFANG'                  'DELAMINA'               'DENA'
'DEBRIFF'                 'DEFEATUR'                'DELEGALIZ'              'DENOMINAT'
'DEBUG'                   'DEFEND'                  'DELEGATION'             'DENOT'
'DECAMP'                  'DEFLOWER'                'DELIMIT'                'DENUMERA'
'DECANT '                 'DEFOLIAT'                'DELINEAT'               'DEODOR'
'DECAPITAT'               'DEFORM'                  'DELIST'                 'DEPART '
'DECENTER'                'DEFRAUD'                 'DELOCAL'                'DEPARTED '
'DECENTR'                 'DEFROST'                 'DELOUS'                 'DEPARTING '
'DECERTIF'                'DEFUS'                   'DEMAGNETI'              'DEPEOPLE'
'DECLASS'                 'DEGENERA'                'DEMARK'                 'DEPERSON'
'DECOD'                   'DEGLACIA'                'DEMATERIALI'            'DEPICTUR'
'DECOLONI'                'DEGLAMORIZ'              'DEMEAN'                 'DEPLAN'
'DECOM'                   'DEGRAD'                  'DEMERIT'                'DEPLOY'
'DECON'                   'DEGUM'                   'DEMILITAR'              'DEPOL'
'DECRESC'                 'DEHORN'                  'DEMOBIL'                'DEPOP'
'DECRY'                   'DEHUM'                   'DEMODULAT'              'DEPORT'
'DECURV'

DECA           INCLUD                           TEN
'DECAGRAM'                'DECALITER'               'DECAMETER'

DECI           INCLUD                           TENTH
'DECIGRAM'                'DECILITER'               'DECIMETER'

DEMI           INCLUD                           (HALF)
'DEMIBLOND'               'DEMIGOD'

DI             INCLUD                           TWO,DOUBLE
'DIATOMIC'                'DICHROM'                 'DISYLLAB'               'DITHEIS'

DIS            EXCLUD                           APART,AWAY,UTTERLY,PNR
'DISABUS'                 'DISH '                   'DISPLAY'                'DISSOLVI '
'DISAFFECT'               'ISHCOVER'                'DISPOSA'                'DISSONAN'
'DISAST'                  'DISHES '                 'DISPOSE'                'DISTAFF'
'DISBURS'                 'DISHEVEL'                'DISPOSI'                'DISTAIN'
'DISC '                   'DISHFUL'                 'DISPOSUR'               'DISTAL'
'DISCER'                  'DISHRAG'                 'DISPREAL'               'DISTAN'
'DISCI'                   'DISHTOWEL'               'DISPUT'                 'DISTEN'
'DISCLOSE'                'DISHWA'                  'DISRUPT'                'DISTI'
```

APPFNDIX A
PREFIX DATA LIST

```
'DISCORD'                 'DISK'                 'DISSECT '              'DISTORT'
'DISCREET'                'DISMAL'               'DISSECTED'             'DISTRACT'
'DISCREPAN'               'DISMAY'               'DISSEMINAT'            'DISTRAUGHT'
'DISCREJ'                 'DISMISS'              'DISSEN'                'DISTRESS'
'DISCRIM'                 'DISPARA'              'DISSERT'               'DISTRICT'
'DISCUS '                 'DISPATCH'             'DISSIIDEN'             'DISTURB'
'DISCUSS'                 'DISPEL'               'DISSIPAT'              'DISUAD'
'DISDAIN'                 'DISPFN'               'DISSOLUT'              'DISUASIVE'
'DISEAS'                  'DISPER'               'DISSOLVE '             'DISYLLAB'
'DISGRUNT'                'DISPIRIT'

DOWN          EXCLUD                           (DOWN)
'DOWN '                   'DOWNRIGHT'              'DOWNTIME'            'DOWNWARD'
'DOWN-TO-EARTH'           'DOWNSTAGE'              'DOWNTOWN'            'DOWNY'
'DOWNPAYMENT'

E             INCLUD                           VAR OF 'EX' UTTERLY,ETC.
'EDUCE '                  'ELUCIDAT'               'EMERGING '           'EVAL'
'EDUCT'                   'EMASCULAT'              'EMIGRA'              'EVAPO'
'ELABOR'                  'EMERGE '                'ENUMERA'             'EVISCERAT'
'ELAPS'                   'EMERGED'                'ERADIAT'             'EVOC'
'ELOCUTION'               'EMERGENT '              'ERUPT'               'EVOK'
'ELOPE'

EM            INCLUD                           ENCLOS,PUT INTO OR ON,GIVE THE QUALITY,
'EMBALM'                  'EMBLAZ'                 'EMBOWEL'             'EMPANEL'
'EMBANK'                  'EMBOD'                  'EMBRACE '            'EMPLAC'
'EMBATTL'                 'EMBOLDEN'               'EMBRITTL'            'EMPLOY '
'EMBED'                   'EMBOSOM'                'EMBROIDER'           'EMPOISON'
'EMBITTER'                'EMBOW'                  'EMBUS'               'EMPOWER'

EN            INCLUD                           IN, OR VB FORM. OR TRANSITIVE
'ENABL'                   'ENDEAR'                 'ENKINDL'             'ENSUR'
'ENACT'                   'ENDUR'                  'ENLAC'               'ENTANGL'
'ENAMOR'                  'ENFAC'                  'ENLARG'              'ENTHRON'
'ENCAG'                   'ENFEEBL'                'ENLIGHT'             'ENTITL'
'ENCAMP'                  'ENFOLD'                 'ENLIST'              'ENTOMB'
'ENCAPSUL'                'ENFORC'                 'ENLIV'               'ENTRAIN'
'ENCAS'                   'ENFRANCHIS'             'ENMESH'              'ENTRANC'
'ENCHAIN'                 'ENGAG'                  'ENRAPT'              'ENTRAP'
'ENCLASP'                 'ENGENDER'               'ENREGISTER'          'ENTRENCH'
'ENCLOS'                  'ENGIRD'                 'ENRICH'              'ENTRUST'
'ENCOD'                   'ENGORG'                 'ENROLL'              'ENTWI'
'ENCOMPASS'               'ENGRAIN'                'ENSAMPL'             'ENVISAG'
'ENCOURAG'                'ENGRAV'                 'ENSHRIN'             'ENVISION'
'ENCRUST'                 'ENHEARTEN'              'ENSHROUD'            'ENWRAP'
'ENCYST'                  'ENJOIN'                 'ENSLAV'              'ENWREATH'
'ENDANGER'                'ENJOY'                  'ENSNAR'

EPI           EXCLUD                           (AT,BEFORE,AFTER)
'EPIC '                   'EPIGRA'                 'EPISCO'              'EPITHELI'
'EPICURE'                 'EPILFP'                 'EPISO'               'EPITHET'
'EPIDERM'                 'EPILOG'                 'EPISTLE'             'EPITOME'
'EPIGENE'                 'EPIPHENOMEN'            'EPITAPH'
```

APPENDIX A
PREFIX DATA LIST

```
ERE           INCLUD                               (BEFORE,-ARCHAIC-)
'ERELONG'               'ERENOW'                       'EREWHILE'

EX            INCLUD                               EX
'EX-'                   'EXCHANG'                      'EXCURS'                 'EXPORT'
'EXCENTRIC'             'EXCURRENT'

EXTRA         EXCLUD                               OUTSIDE,ADDITIONAL,MORE THAN USUAL,SUPE
'EXTRACT'               'EXTRAMURAL'                   'EXTRAPOLAT'             'EXTRAVER'
'EXTRAD'                'EXTRANEOUS'                   'EXTRAVAGAN'

FARM          EXCLUD                               (FARM)
'FARME'                 'FARMI'

FAT           INCLUD                               (FAT)
'FAT-FACED'             'FATFREE'                      'FATHEAD'

FOR           INCLUD             AWAY,OFF,EXTREMELY,WRONGLY,NEGATIV OR PRIVATIV FO
'FORBAD'                'FORDO'                        'FORSAKE'                'FORSWEAR'
'FORBEAR '              'FORFEND'                      'FORSOO'                 'FORSWOR'
'FORBID '               'FORGIV'                       'FORSPENT'               'FORWENT'
'FORBOR'                'FORGO '

FORE          EXCLUD             BEFORE←←,FRONT,SUPERIOR
'FORE-AND'              'FOREDO '                      'FOREIGN'                'FORESTATION'
'FOREBODING'            'FOREDOING'                    'FORENSIC'               'FORESTER'
'FORECAST '             'FOREDONE'                     'FOREST '                'FORESTRY'
'FORECASTS '            'FOREGO '                      'FORESTED '              'FOREVER'
'FOREDID'

GEO           INCLUD             THE EARTH
'GEOCENTRIC'            'GEOGRAPHIC'                   'GEOPHYSIC'

GOAL          INCLUD                               (GOAL)
'GOALK'                 'GOALTEND'

GUIDE         EXCLUD                               (GUIDE)
'GUIDED'

HAIR          EXCLUD                               (HAIR)
'HAIR '                 'HAIRDO'                       'HAIRSPLIT'              'HAIRY'
'HAIRBRA'               'HAIRLES'

HALF          INCLUD                               HALF
'HALF-AND-HALF'         'HALF-HEARTED'                 'HALF-TRACK'             'HALFWAY'
'HALF-BLOOD'

HEMI          INCLUD                               HALF
'HEMISPHER'

HETERO        INCLUD                               DIFFERENT,OTHER
'HETEROCHROM'           'HETEROSEX'
```

APPENDIX A
PREFIX DATA LIST

```
HEXA        INCLUD                          SIX
'HEXAMETER'           'HEXANGULAR'             'HEXASYLLABL'

HIND        INCLUD                          REAR,PAST
'HINDQUARTER'         'HINDSIGHT'

HOMO        INCLUD                          SAME
'HOMOCENTRIC'         'HOMOCHRO'               'HOMOSEX'

HUMAN       INCLUD                          (HUMAN)
'HUMAN-INTER'         'HUMANHOOD'              'HUMANKIND'           'HUMANMIND'
'HUMANHEART'

HYDRO       EXCLUD            WATER,HYDROGEN
'HYDROCHLORIC'        'HYDROGRAPHY'            'HYDROLY'             'HYDROUS'
'HYDROGEN'            'HYDROLOGY'              'HYDROPHOB'

HYLO        INCLUD                          WOOD,MATTER
'HYLOTHEIS'

HYPER       EXCLUD                          OVER,(SEM.DIF.EXCESS)
'HYPERBO'             'HYPERURBAN'

HYPNO       INCLUD            SLEEP,HYPNOSIS
'HYPNOANALY'          'HYPNOG'                 'HYPNOTHERAP'

HYPO        EXCLUD                          UNDER,LESS,LOW
'HYPOTHESIS'          'HYPOTHETC'

ICONO       INCLUD                          IMAGE,LIKENESS
'ICONOGRAPH'

IL          INCLUD            VAR.OF IN,NOT,VB.FORM .
'ILLAUDAB'            'ILLIBERAL'              'ILLIMIT'             'ILLOGIC'
'ILLEGAL'             'ILLICIT'                'ILLITERA'            'ILLUMIN'
'ILLEGI'

IM          INCLUD            VAR.OF IN,NOT,ETC.
'IMBALANC'            'IMMOR'                  'IMPERC'              'IMPOWER'
'IMBALM'              'IMMOTI'                 'IMPERF'              'IMPRACT'
'IMBARK'              'IMMOV'                  'IMPERIL'             'IMPREC'
'IMBITTER'            'IMMUSIC'                'IMPERISHAB'          'IMPREGNAB'
'IMBO'                'IMMUTA'                 'IMPERM'              'IMPRINT'
'IMBRANGL'            'IMPARADI'               'IMPERSON'            'IMPRISON'
'IMBRUT'              'IMPARIT'                'IMPERT'              'IMPROB'
'IMMACULAT'           'IMPARK'                 'IMPLAC'              'IMPROMPT'
'IMMERG'              'IMPARTIB'               'IMPLANT'             'IMPROPER'
'IMMESH'              'IMPASSAB'               'IMPLAUS'             'IMPROPRIET'
'IMMIGRA'             'IMPASSION'              'IMPOLI'              'IMPROVIDEN'
'IMMINGL'             'IMPATIEN'               'IMPOND'              'IMPRUDEN'
'IMMISCIB'            'IMPEND'                 'IMPOSS'              'IMPUISSAN'
'IMMIX'               'IMPENET'                'IMPOT'               'IMPUS'
'IMMOBIL'             'IMPENIT'                'IMPOVER'             'IMPUTRESCIB'
'IMMOD'
```

APPENDIX A
PREFIX DATA LIST

```
IN          EXCLUD          UN,NOT,IN(TO),VB.FORMATIVE    TRANSITIVE
'IN-AND'                'INDENT'              'INHAL'               'INTEG'
'INADVERT'              'INDEX'               'INHE'                'INTELL'
'INAM'                  'INDIA'               'INHIBIT'             'INTEN'
'INANE'                 'INDIC'               'INI'                 'INTER'
'INANI'                 'INDIGEN'             'INJECT'              'INTESTIN'
'INARM'                 'INDIGNA'             'INJUN'               'INTHRALL'
'INASMUCH'              'INDIGO'              'INJUR'               'INTIM'
'INAUGURA'              'INDIT'               'INK'                 'INTIN'
'INCANDESC'             'INDIVIDUA'           'INNE'                'INTO '
'INCANT'                'INDO-'               'INNO'                'INTOXICA'
'INCAR'                 'INDOCH'              'INOC'                'INTRA-'
'INCEN'                 'INDOL'               'INQUIR'              'INTRAM'
'INCEP'                 'INDU'                'INQUIS'              'INTRAS'
'INCES'                 'INEBRI'              'INROAD'              'INTRAV'
'INCH'                  'INERT'               'INSCR'               'INTREPID'
'INCID'                 'INFAM'               'INSECT'              'INTRI'
'INCIN'                 'INFANT'              'INSERT'              'INTRO'
'INCIP'                 'INFECT'              'INSIGNE'             'INTRUD'
'INCIS'                 'INFER'               'INSIGNIA'            'INTRUSI'
'INCIT'                 'INFEST'              'INSINUAT'            'INTUIT'
'INCLE'                 'INFIDEL'             'INSIP'               'INUNDA'
'INCLI'                 'INFIRMAR'            'INSIST'              'INUR'
'INCLU'                 'INFIRMI'             'INSOFAR'             'INVAD'
'INCOG'                 'INFIX'               'INSOLA'              'INVAS'
'INCREAS'               'INFLAT'              'INSOLE'              'INVECT'
'INCREM'                'INFLECT'             'INSOMNIA'            'INVEI'
'INCRESC'               'INFLICT'             'INSOMUCH'            'INVENT'
'INCRET'                'INFORM '             'INSPECT'             'INVERSE'
'INCRIMIN'              'INFORMANT '          'INSPIR'              'INVERT '
'INCUB'                 'INFORMAT '           'INSTAL'              'INVERTED '
'INCULCAT'              'INFORMED '           'INSTAN'              'INVERTI'
'INCULPATE '            'INFORMER '           'INSTI'               'INVET'
'INCULT'                'INFORMING '          'INSTRU'              'INVID'
'INCUMBRA'              'INFRA'               'INSUL'               'INVIG'
'INCUMBREN'             'INGEST'              'INSURA'              'INVIT'
'INCUR '                'INGO'                'INSURE'              'INVOC'
'INCURRE'               'INGRATIAT'           'INSURG'              'INVOIC'
'INCURS'                'INGRED'              'INSURRECT'           'INVOK'
'INDEMNI'               'INHABIT'             'INTAK'               'INVOLV'

INFRA       EXCLUD          BELOW
'INFRACT'               'INFRANGIB'

INTER       EXCLUD          AMONG,BETWEEN,MUTUALLY,DURING,ETC.
'INTERCED'              'INTERJECT'           'INTERNED'            'INTERSECT'
'INTERCEPT'             'INTERLUDE'           'INTERNEE'            'INTERSPERS'
'INTERCESS'             'INTERMEDIAT'         'INTERNING'           'INTERSTIC'
'INTERCOM '             'INTERMENT'           'INTERNIST'           'INTERSTIT'
'INTERDICT'             'INTERMINAB'          'INTERNMENT'          'INTERVAL '
'INTEREST'              'INTERMISSI'          'INTERPEL'            'INTERVALS '
'INTERFER'              'INTERMIT'            'INTERPOL '           'INTERVEN'
'INTERIM '              'INTERMURAL'          'INTERPOLAT'          'INTERVIEW'
'INTERIOR'              'INTERN '             'INTERPRET'           'INTERVOLV'
```

APPENDIX A
PREFIX DATA LIST

```
'INTERJACEN'          'INTERNAL'            'INTERR'

IR          EXCLUD            IN,NOT,VB.FORMATIVE   TRANSITIVE
'IRA'                 'IRISH '              'IRREMEAB'            'IRRIGUOUS'
'IRE'                 'IRO'                 'IRRIGA'              'IRRITA'

KEY         INCLUD            KEY(LOCK),CENTRAL IMPORTANCE
'KEYHO'               'KEYNOTE'             'KEYSMI'              'KEYSTROK'
'KEYMAN'              'KEYPUNC'             'KEYSTONE'            'KEYWORD'

LITHO       INCLUD                      STONE
'LITHOGRAPH'          'LITHOPRINT'          'LITHOSPHER'

MACRO       EXCLUD            LARGE.LONG.EXCESSIVE,NO WORDS-SOME ARE RARE

MAL         INCLUD            BAD,WRONG,ILL, FR.
'MALADAPT'            'MALADMINIS'          'MALCONTENT'          'MALF'
'MALADJUST'           'MALAPPORTION'

META        EXCLUD                      AFTER,AWAY,BEYOND,BEHIND
'METABOLI'            'METAMER'             'METAMOR'             'METAPHOR'
'METAL'

MICRO       EXCLUD            SMALL,ENLARGING SOMETHING SMALL
'MICROBE '

MID         INCLUD            MIDDLE,BETWEEN
'MIDA'                'MIDL'                'MIDRASH'             'MIDSTREAM'
'MIDB'                'MIDMO'               'MIDSECTION'          'MIDSUMMER'
'MIDC'                'MIDN'                'MIDSHIP '            'MIDT'
'MIDDAY'              'MIDPOINT'            'MIDSHIPS '           'MIDW'

MIS         EXCLUD            ILL,MISTAKEN,WRONG
'MISCE'               'MISER'               'MISPRIS'             'MISTAK'
'MISCHA'              'MISGIV'              'MISS '               'MISTER'
'MISCHIE'             'MISHAP'              'MISSED '             'MISTLE'
'MISCI'               'MISHMASH'            'MISSI'               'MISTOOK'
'MISCREAN'            'MISNOMER'            'MISSY'               'MISTRESS'
'MISE '               'MISO'                'MIST '               'MISTY'

MON         INCLUD            ALONE,SINGLE,ONE,VAR.OF 'MONO'
'MONAURAL'

MONO        EXCLUD            ALONE,SINGLE,ONE,
'MONOLITH'            'MONOPOL'

MULTI       EXCLUD            MANY
'MULTIFARIOUS'        'MULTIPAR'            'MULTIPLI'            'MULTITUD'
'MULTIFID'            'MULTIPLE'            'MULTIPLY'

NEO         EXCLUD                      NEW,RECENT
'NEOLITH'             'NEOLOG'              'NEON '               'NEOPHYT'

NO          INCLUD            NO
'NO-'                 'NOBODY'              'NOWAY'               'NOWHERE'
```

APPENDIX A
PREFIX DATA LIST

```
NON            EXCLUD             NOT,'LACKING+',NOT NECESSAIRILY 'REVERSE'
'NONAGE '                 'NONCOM '                'NONESUCH'               'NONPLUS'
'NONCE \'                 'NONDESCRIPT  '          'NONPAREIL'              'NONSUCH'
'NONCHALAN'               'NONE '

OB             INCLUD                                 TOWARD,ON,OVER,AGAINST
'OBLIGAT'                 'OBLONG'                 'OBNOX'                  'OBVER'

OFF            INCLUD             OFF
'OFF-'                    'OFFCAST'                'OFFPRINT'               'OFFTAK'
'OFFBEAT'                 'OFFHAND'                'OFFS'

OUT            EXCLUD             OUT+TRANS.VB.GOING BEYOND,SURPASSING,OUTDOING
'OUT-OF'                  'OUTFIT'                 'OUTRAG'                 'OUTSIDE'
'OUTAG'                   'OUTLIER'                'OUTSET '                'OUTWARD'
'OUTER'

OVER           EXCLUD             OVER A LIMIT
'OVER '                   'OVERLAP'                'OVERSEER'               'OVERT '

PAN            INCLUD                                 ALL,GENERAL
'PANSOPHISM'              'PANTHEISM'              'PANTROPIC'

PAPA           INCLUD             PARACHUT,GUARD AGAINST,BESID,NEAR,AMISS,+IMP.ALTE
'PARABOMB'                'PARAMEDIC '             'PARAPHRAS'              'PARASOL'
'PARACHUT'                'PARAMAGNET'             'PARAPSYCHO'             'PARATROOP'
'PARAGLIDER'              'PARAMILIT'              'PARARESCU'

PAY            INCLUD             TO PAY ETC.
'PAYDAY'                  'PAYLOAD'                'PAYMASTER'

PEP            EXCLUD             THROUGH,UTTERLY,VERY,THOROUGHLY
'PERAMBULAT'              'PERFECT '               'PERHAPS '               'PERSON '
'PERDUR'                  'PERFERVID'              'PERMUTAT'               'PERSUA'

PERI           INCLUD             ABOUT,AROUND,BEYOND
'PERISCOP'

POLY           INCLUD             MULTIPLE,MUCH,MANY
'POLYANG'                 'POLYGENE'               'POLYSYLLAB'             'POLYTON'
'POLYCHROM'               'POLYGRAPH'              'POLYTECHNIC'            'POLYTYP'
'POLYETHNIC'              'POLYPHON'               'POLYTHEIS'

POST           EXCLUD             BEHIND,AFTER,MAIL
'POSTAGE'                 'POSTER'                 'POSTING'                'POSTU'
'POSTAL'                  'POSTIC'                 'POSTPONE'

PRE            EXCLUD             BEFORE,PRIOR TO,EARLY,IN FRONT OF
'PREACH'                  'PREFACE'                'PREP '                  'PRESENT'
'PREAMBL'                 'PREFAT'                 'PREPARA'                'PRESERV'
'PRECARI'                 'PREFE'                  'PREPARE '               'PRESID'
'PRECAT'                  'PREFIX'                 'PREPARED'               'PRESS'
'PRECE'                   'PREGNAN'                'PREPENSE'               'PREST '
'PRECI'                   'PREHENS'                'PREPOSITION'            'PRESTI'
```

APPENDIX A
PREFIX DATA LIST

```
'PRECL'                    'PREJUDIC'                 'PREPOSSES'                'PRESUM'
'PRECOG'                   'PRELA'                    'PREPOSTEROUS'             'PRETEN'
'PRECUR'                   'PRELIMIN'                 'PREPUCE'                  'PRETER'
'PREDAGIOUS'               'PRELU'                    'PREROG'                   'PRETT'
'PREDATOR'                 'PREMIE'                   'PRESBYTER'                'PREVA'
'PPEDECESS'                'PREMISE'                  'PRESCIND'                 'PREVEN'
'PREDIC'                   'PREMIUM'                  'PRESCRIBE'                'PREVIOUS'
'PREEMPT'                  'PREMON'                   'PRESCRIPT'                'PREXY'
'PREEN '                   'PRENTICE'                 'PRESENC'                  'PREY'
'PREFAB '

PRETER       EXCLUD              BEYOND,MORE THAN,BY,PAST
'PRETERI'                  'PRETERMIT'

PRO          EXCLUD              FOR (A CAUSE,ETC.)
'PRO-SAT'                  'PROFF'                    'PROMOT'                   'PROSAI'
'PROBAB'                   'PROFICI'                  'PROMPT'                   'PRORAT'
'PROBAT'                   'PROFIL'                   'PROMULG'                  'PROSCRIB'
'PROBE'                    'PROFIT'                   'PRONE '                   'PROSE'
'PROBITY'                  'PROFLIG'                  'PRONG'                    'PROSOD'
'PROBLEM'                  'PROFOUND'                 'PRONOUNC'                 'PROSPECT'
'PROCED'                   'PROFU'                    'PRONTO'                   'PROSPER'
'PROCEED'                  'PROGEN'                   'PRONUN'                   'PROSTHE'
'PROCESS'                  'PROGN'                    'PROOF'                    'PROSTITUT'
'PROCLAIM'                 'PROGRAM'                  'PROPAG'                   'PROSTRAT'
'PROCLAM'                  'PROGRESS'                 'PROPEL'                   'PROTAG'
'PROCLIVITY'               'PROHIB'                   'PROPEN'                   'PROTEAN'
'PROCRAST'                 'PROJECT'                  'PROPERTY'                 'PROTECT'
'PROCREANT'                'PROLE'                    'PROPH'                    'PROTEG'
'PROCTO'                   'PROLIF'                   'PROPI'                    'PROTEIN'
'PROCUR'                   'PROLIX'                   'PROPJET'                  'PROTEST'
'PROD '                    'PROLOG'                   'PROPON'                   'PROTO'
'PRODDF'                   'PROLONGAT'                'PROPOS'                   'PROTRACT'
'PRODIG'                   'PROMENAD'                 'PROPOUND'                 'PROTRU'
'PRODU'                    'PROMINEN'                 'PROPRIET'                 'PROTU'
'PROF '                    'PROMIS'                   'PROPRIO'                  'PROUD '
'PROFAN'                   'PROMON'                   'PROPULS'                  'PROVINC'
'PROFESS'

PROTO        INCLUD              FIRST,FOREMOST,EARLIST FORM OF,MAYB SHOULD 0 IF S
'PROTOHUMAN'               'PROTOLINGU'               'PROTOPLASM'               'PROTOTYP'
'PROTOLANGUAGE'

PSEUDO       EXCLUD              FALSE,PRETENDED
'PSEUDOMORPH'              'PSEUDONYM'

QUASI        EXCLUD                                RESEMBLING,SEEMING

RE           EXCLUD              BACKWARD,AGAIN
'REACH'                    'RED-'                     'RELIE'                    'REQUIS'
'REACTA'                   'REDB'                     'RELIGIO'                  'REREMOUSE'
'READ '                    'REDD'                     'RELIQUE'                  'RESCI'
'READER'                   'REDEEM'                   'RELISH'                   'RESCU'
'READI'                    'REDEMPT'                  'RELUC'                    'RESEARCH'
'READY'                    'REDIN'                    'REMAIN'                   'RESEMBL'
```

APPENDIX A
PREFIX DATA LIST

```
'REAL  '
'REALIS'
'REALIT'
'REALIZ'
'REALLY'
'REALM'
'REALPOLITIK'
'REALTY'
'REAM  '
'REAP  '
'REAPER'
'REAPI'
'REAR  '
'REARMOST'
'REARWARD'
'REASON'
'REBATE'
'REBEL'
'REBUFF'
'REBUK'
'REBUT  '
'REBUTTAL'
'RECALCITRA'
'RECANT'
'RECED'
'RECEI'
'RECENSION'
'RECENT  '
'RECEPT'
'RECESS'
'RECIDIV'
'RECIP'
'RECIS'
'RECIT'
'RECK'
'RECLAM'
'RECLIN'
'RECLUS'
'RECOGNI'
'RECOIL'
'RECOLLECT'
'RECOMPENSE'
'RECOMPENSI'
'RECON  '
'RECONCIL'
'RECONDITE  '
'RECONNAISANCE'
'RECONNOIT'
'RECORD'
'RECOUP'
'RECOURS'
'RECOVER'
'RECREANT'
'RECRIMINAT'
'RECRUIT'
'REDN'
'REDOLEN'
'REDOUBT'
'REDRESS'
'PEDS'
'REDUC'
'REED  '
'REEF'
'REEK'
'REEL  '
'REELING'
'REEVE'
'REFECT'
'REFER'
'REFINE'
'REFLAT'
'REFLECT'
'REFLEX'
'REFORMATORY'
'REFORMIST'
'REFRACT'
'REFRAIN'
'REFRI'
'REFUG'
'REFUS'
'REFUT'
'REGAL'
'REGARD'
'REGENCY'
'REGENT'
'REGICID'
'REGIM'
'REGION'
'REGIST'
'REGRESS'
'REGRET'
'REGULA'
'REGURGITAT'
'REHASH'
'REHEARS'
'REICH'
'REIGN'
'REIMBURS'
'REIN  '
'REINS  '
'REITERANT'
'REJECT'
'REJOIC'
'REJOINDER'
'KEYBO'
'REJUVEN'
'RELAT'
'RELAX'
'RELAY'
'RELEAS'
'REMAND'
'REMARK'
'REMEDIAB'
'REMEDIAL'
'REMEDILESS'
'REMEDY'
'REMEMB'
'REMIND'
'REMINISC'
'REMISS'
'REMIT'
'REMNAN'
'REMONSTRA'
'REMORS'
'REMOTE'
'REMOVAL'
'REMUNERAT'
'REND'
'RENEGAD'
'RENEGE'
'RENEWAL'
'RENOUNC'
'RENOVAT'
'RENOWN'
'RENT'
'RENUNC'
'REPAIR'
'REPARA'
'REPARTE'
'REPAST'
'REPEA'
'REPEL'
'REPERTO'
'REPETIT'
'REPLENISH'
'REPLET'
'REPLI'
'REPLICA'
'REPLY'
'REPORT'
'REPOSAL'
'REPOSE'
'REPOSITORY'
'REPREHEN'
'REPRESENT'
'REPRESS'
'REPRIEV'
'REPRIMAND'
'REPRIS'
'REPROACH'
'REPROBAT'
'REPROOF'
'REPROV'
'REPTIL'
'REPUBLIC'
'RESENT'
'RESERV'
'RESID'
'RESIGN'
'RESILE'
'RESILI'
'RESIN'
'RESIST'
'RESOLU'
'RESOLV'
'RESONA'
'RESORT'
'RESOURC'
'RESPECT'
'RESPIR'
'RESPIT'
'RESPLEND'
'RESPON'
'REST  '
'RESTED  '
'RESTAUR'
'RESTFUL'
'RESTITUT'
'RESTIVE'
'RESTLESS'
'RESTORAT'
'RESTRAIN
'RESUM'
'RESURG'
'RESURR'
'RESUSC'
'RETAI'
'RETAL'
'RETARD'
'RETCH'
'RETENTI'
'RETIC'
'RETICEN'
'RETIN'
'RETIR'
'RETORT  '
'RETRACT'
'RETREAT'
'RETRIBUT'
'RETRIEV'
'RETRO'
'RETRU'
'RETURN'
'REVEAL'
'REVEL'
'REVEN'
'REVER'
'REVIS'
'REVIV'
'REVOC'
```

APPENDIX A
PREFIX DATA LIST

```
'RFCT'                'RELEG'               'REPUGN'              'REVOK'
'RECUM'               'RELENT'              'REPULS'              'REVOL'
'RECUP'               'RFLEVANT'            'REPUT'               'REVULS'
'RECUR '              'RELIAB'              'REQUEST'             'REWARD'
'RECURP'              'RELIAN'              'REQUIE'              'REX'
'RECURS'              'RELIC '              'REQUIR'              'REY'
'RED '                'RELICS '

RETRO        EXCLUD                         BACKWARD,BEHIND
'RFTROCED'            'PETROGRESS'          'RETROSPECT'          'RETROVERSION'
'RETROGPADE'

SELF         INCLUD           COMB.FORM OF 'SELF'
'SELF-'               'SELFSAME'

SEMI         EXCLUD           HALF
'SEMINA'              'SEMINI'              'SEMITIC'

SIDE         EXCLUD           SIDE
'SIDEBURN'            'SIDER'               'SIDES '              'SIDEWINDER'
'SIDFKICK'            'SIDE '               'SIDESPLIT'

STEP         EXCLUD           STEP
'STEPH'               'STEPP'               'STEPS '              'STEPWISE'

SUB          EXCLUD           BELOW,SLIGHTLY,(NOTION OF ASSISTANCE)
'SUBALTERN'           'SUBMERG'             'SUBSIST'             'SUBTILI'
'SUBDUCT'             'SUBMERS'             'SUBSTANC'            'SUBTILLY'
'SUBDUE'              'SUBMISS'             'SUBSTANT'            'SUBTLE'
'SUBER'               'SUBMIT'              'SUBSTITUT'           'SUBTRA'
'SUBJECT'             'SUBORN'              'SUBSUM'              'SUBURB '
'SUBJUNCT'            'SUBSCRIB'            'SUBTEND'             'SUBURBIA '
'SUBLIMAT'            'SUBSCRIPT'           'SUBTER'              'SUBURBS '
'SUBLIME'             'SUBSID'              'SUBTILE'             'SUBVE'

SUBTER       INCLUD           UNDER,BELOW
'SUBTERNATURAL'

SUPER        EXCLUD           ABOVE,BEYOND,TO AN ESPEC. HIGH DEGREE
'SUPER-DUPER'         'SUPERFACECT'         'SUPERLATIV'          'SUPERSCRIBE'
'SUPERABLE'           'SUPERFIC'            'SUPERNAL'            'SUPERSCRIPTION'
'SUPERANNUAT'         'SUPERFLUITY'         'SUPERORDINAT'        'SUPERSED'
'SUPERB '             'SUPERFLUOUS'         'SUPERPOSE '          'SUPERSESSION'
'SUPERCARGO'          'SUPERFUS'            'SUPERPOSED '         'SUPERSTIT'
'SUPERCIL'            'SUPERINTEND'         'SUPERPOSING '        'SUPERVENE'
'SUPEREGO'            'SUPERIOR'            'SUPERPOSITION'       'SUPERVIS'
'SUPERETTE'

SUPRA        EXCLUD           VAR.OF 'SUPER' EMPHASIZING POSITION

SUR          INCLUD           VAR.OF 'SUPER',VAR.OF 'SUB'
'SURCEAS'             'SURFACE'             'SURPASS'             'SURREAL'
'SURCHARG'            'SURMOUNT'            'SURPLUS'             'SURROUND'
'SURCOAT'             'SURNAM'              'SURPRINT'            'SURTAX'
```

APPENDIX A
PREFIX DATA LIST

```
SYM            INCLUD              WITH,TOGETHER
'SYMMETRIC'

SYN            INCLUD              WITH,TOGETHER,IN ASSOC.(WITH)
'SYNAESTHESIA'          'SYNECOLOGY'            'SYNESTHESIA'           'SYNGENESIS'

SYNCHRO        EXCLUD                                 SYNCHRONOUS
'SYNCHRONAL'            'SYNCHRONISM'           'SYNCHRONIZ'            'SYNCHRONOU'
'SYNCHRONISE'

TAX            INCLUD              ORDERING,DIRECTION,TAX
'TAX-'                  'TAXGATHER'             'TAXP'

TAXI           INCLUD              TAXI(CAB),VAR.OF TAXO
'TAXIMETER'             'TAXIPLANE'             'TAXIWAY'

TETRA          EXCLUD                                 FOUR
'TETRAD'                'TETRAHED'              'TETRAMER'

THOROUGH       EXCLUD              THOROUGH,THROUGH
'THOROUGHFARE'

THROUGH        INCLUD              THROUGH
'THROUGHPUT'            'THROUGHWAY'

TRANS          EXCLUD              ACCROSS,BEYOND,THROUGH
'TRANSCEIVER'           'TRANSFER'              'TRANSLITERAT'          'TRANSPIR'
'TRANSCEND'             'TRANSFORMER'           'TRANSMISS'             'TRANSPORT'
'TRANSCIEN'             'TRANSGRESS'            'TRANSMIT'              'TRANSPOS'
'TRANSCRIPTION'         'TRANSIST'              'TRANSOM'               'TRANSUD'
'TRANSDUCE'             'TRANSIT'               'TRANSPAREN'            'TRANSV'
'TRANSECT'              'TRANSLAT'

TRI            INCLUD              THREE
'TRI-STATE'             'TRIFORM'               'TRIMOTOR'              'TRIPLANE'
'TRIANGLE'              'TRILINGUAL'            'TRINITRO'              'TRISYLLAB'
'TRICHROMAT'            'TRIMETALLI'            'TRIPEDAL'              'TRIWEEK'
'TRICYCL'               'TRIMONTHLY'

TROPO          INCLUD              TURN,TURNING
'TROPOSPHER'

ULTRA          EXCLUD                                 BEYOND USUAL,EXCESSIVE
'ULTRAISM'

UN             EXCLUD              UN,NOT,LACKING IN,ONE
'UNANIM'                'UNCT'                  'UNGUENT'               'UNID'
'UNCANNY'               'UNDER'                 'UNGUL'                 'UNIF'
'UNCHANY'               'UNDIES'                'UNIAXIAL'              'UNIL'
'UNCLE '                'UNDULA'                'UNIC'                  'UNION'

UNDER          EXCLUD              UNDER,ONE
'UNDER-THE'             'UNDERNEATH'            'UNDERSTAND'            'UNDERTOOK'
'UNDERLING'             'UNDEROGATORY'          'UNDERTAK'
```

APPENDIX A
PREFIX DATA LIST

```
UNI         INCLUD           ONE
'UNIAXIAL'           'UNIDIRECT'          'UNILING'            'UNISEX'
'UNICAMERAL'         'UNIFORM'            'UNILO'              'UNIVERSE '
'UNICYCL'            'UNILATERAL'         'UNIP'

UP          EXCLUD           UP
'UP-AND'             'UPBRINGING'         'UPON'               'UPSHOT'
'UP-TO'              'UPHEAVAL'           'UPP'                'UPWARD'
'UPBRAID'            'UPHOLSTER'          'UPSET'

VICE        EXCLUD                        DEPUTY
'VICELES'            'VICEN'

WELL        EXCLUD                        GOOD
'WELL '              'WELL-HEEL'          'WELL-SPRING'        'WELLAWAY'
'WELL-FAVOR'         'WELL-OFF'           'WELL-TO-DO'         'WELLS '
'WELL-FIX'           'WELL-OIL'           'WELL-TURN'

WITH        INCLUD           COMBINING FORM OF WITH,SEPARATIVE OR OPPOSING FOR
'WITHDRAW '          'WITHDREW'           'WITHIN'             'WITHSTAND'
'WITHDRAWING '       'WITHOLD'            'WITHOUT'

XYLO        INCLUD                        WOOD
'XYLOGRAPH'          'XYLOPHON'

YESTER      EXCLUD                        PRECEDING

ZYGO        INCLUD                        SCI. UNION,CONNECT
'ZYGOGENESIS'        'ZYGOSPORE'
```

APPENDIX B

Suffix Data List

APPENDIX B
SUFFIX DATA LIST

```
<BLANK> EE
     LETTER  T
<BLANK> ABILITY
<BLANK> ABLE
     CAP
     CONSIDER
     PORT
<BLANK> AIN
     BEG
     PLANT
<BLANK> AGE
     MESS
<BLANK> AL
     CAN
     FIN
     FORM
     INFORM
     JACKAL
     LATERAL
     LEG
     METAL
     MINER
     PERSONAL
     PET
     PHYSICAL
     ROY
     SAND
     SEVERAL
     SIGN
     SPIN
     VEST
<BLANK> ALLY
     FIN
<BLANK> AMENT
<BLANK> AN
     LETTER E
     LETTER O
     CRIME
     UNCLE
<BLANK> ANCE
     FIN
     IMPORT
<BLANK> ANT
     IMPORTANT
     PAGE
     PROTESTANT
<BLANK> ARD
<BLANK> ARDMENT
<BLANK> ARY
     BOUNDARY
     CAN
     ELEMENT
     HUNGARY
     LITER
     SECRETARY
<BLANK> ARM
<BLANK> ATED
<BLANK> ATIC
<BLANK> ATION
     FOUNDATION
     ROT
<BLANK> ATIZE
<BLANK> C
<BLANK> D
     LETTER E
     AIRE
     FEE
     SEE
     SUITE
<BLANK> DOM
<BLANK> E
     LETTER H
     LETTER L
     LETTER O
     LETTER S
     LETTER X
     MORALE
     CLOTHE
<BLANK> ED
     CARED
     CUB
     FADED
     FIN
     FIRED
     FOUNDED
     HAT
     MATED
     MET
     PAST
     PENN
     RAGGED
     RAT
     RUG
     SCRAPED
     SING
     SPARED
     STARED
     STRIPED
     TAMED
     TWINED
     UNITED
<BLANK> EN
     BARREN
     HAST
     HEATH
     LIST
     POLLEN
     RIP
     SEAM
     STR
<BLANK> ENCE
     OFF
     PRESS
     SENT
<BLANK> ENCIE
<BLANK> ENED
     LIST
<BLANK> ENT
     MISSENT
     PRESENT
     RIP
     ROD
<BLANK> ER
     ARCHER
     BARBER
     BAT
     BIT
     BOOKER
     CAREER
     CENT
     COCKER
     CORN
     CUSTOM
     ENGINEER
     FIN
     FLOWER
     FORM
     HAM
     HUNG
     INN
     LAD
     LET
     LIT
     MAN
     MAST
     MATTER
     METER
     MOTH
     MUST
     NUMB
     OFF
     PET
     PROP
     QUART
     RANGER
     RUB
     SCRAPER
     SETTER
     SHOW
     SHOULD
     SHUT
     SLIP
     SOLD
     SPRINGER
     STAG
     SUM
     SWEATER
     TOW
     TWINER
     WICK
<BLANK> ERN
<BLANK> EST
     DIG
     EARN
     FIN
     FOR
<BLANK> FALL
<BLANK> FARE
<BLANK> FIELD
<BLANK> FUL
<BLANK> HOOD
<BLANK> IA
     GARDEN
     VIRGIN
<BLANK> IAL
     BUR
<BLANK> IAN
     PHYSIC
<BLANK> IATION
<BLANK> IC
     ANT
     CLASS
     CUB
     SONIC
     TOP
<BLANK> ICAL
     CLASS
     LOG
     PERIOD
<BLANK> ING
     BEAR
     BOOK
     CAR
     CLOTHING
     EARRING
     EVEN
     FAD
     FIR
     HERRING
     INN
     MAT
     RANG
     RIDING
     SCRAPING
     TAM
     TICK
     TWINING
     UNITING
<BLANK> ION
     BILL
```

APPENDIX B
SUFFIX DATA LIST

```
          CONTENT
          LEG
          LOT
          MILL
          MISS
          PASS
          PORT
          PROCESSION
          STALL
<BLANK> ISH
          FIN
          FLOURISH
          SPAN
<BLANK> ISM
<BLANK> IST
          ASS
          CELL
          PHYSIC
<BLANK> ISTIC
<BLANK> ITE
<BLANK> ITIE
<BLANK> ITION
          COAL
          PART
<BLANK> ITIONAL
<BLANK> ITY
          AUTHORITY
          DIVERSITY
          SEVERITY
          VANITY
<BLANK> IVE
<BLANK> IZATION
          ORGAN
<BLANK> IZED
          ORGAN
<BLANK> IZE
          ORGAN
<BLANK> IZING
          ORGAN
<BLANK> KING
          TALL
          THIN
          WIN
<BLANK> L
          LETTER A
          IDEAL
          SEA
<BLANK> LEDGE
<BLANK> LESS
          SHIFTLESS
          WIRE
<BLANK> LIKE
<BLANK> -LIKE
<BLANK> LINESS
<BLANK> LIZATION
<BLANK> LY
          EARLY
          HARDLY
          HOMELY
          PEAR
          SING
          STATE
<BLANK> MAKER
<BLANK> MAN
          AIRMAN
          BUSH
          GENTLEMAN
<BLANK> MEN
          GENTLEMEN
          MINUTE
<BLANK> MENT
          APART
          BASE
          DECREE
          PIG
          STATEMENT
<BLANK> MOST
<BLANK> N
<BLANK> OU
          ANGLER
          ARCEER
          BADGER
          LETTER E
<BLANK> R
<BLANK> POWER
<BLANK> OUT
          BEER
          GORGE
          LETTER A
          LETTER E
          LETTER W
          BROWN
          CROW
          DOZE
          FLOW
          HEAVE
          LIEN
          LAW
          LINE
          OWE
          THEN
          TOW
<BLANK> NED
          BUR
          CROW
          EAR
          LEAVE
          PATTER
          WAR
<BLANK> NESS
<BLANK> OLOGY
<BLANK> OR
          FACT
          MAY
          PASTOR
          POT
          TAIL
          TRACT
          TENOR
          CALL
          BROKER
          BUTTE
          CASTE
          CRATE
          DOVER
          EVE
          HOME
          LIVER
          OFFICER
          OLIVE
          PIE
          PRIME
          SKIE
<BLANK> RE
          CENT
          LUST
          STATU
<BLANK> REN
          BARREN
<BLANK> RENCE
<BLANK> RY
          ARCHE
          COUNT
          EVE
          FIE
          HEN
          HUNG
          LIVE
          MAR
          NURSE
          SENT
          SURGERY
<BLANK> SE
          BROW
          DEN
          FALL
          TEASE
<BLANK> SHIP
          AIRSHIP
<BLANK> SIDE
          PRESIDE
<BLANK> SLY
<BLANK> SMEN
<BLANK> ST
          LETTER E
          FORE
<BLANK> STORM
<BLANK> T
          EVEN
          FEE
          FREE
          HEAR
          NIGH
          PAIN
          PLANT
          SEA
          SHE
          SIGH
          THOUGH
<BLANK> TENED
<BLANK> TH
          BREADTH
          DEARTH
          EARTH
          FIR
          FOR
          HEAR
          NOR
          YOU
<BLANK> TIME
          MEANTIME
<BLANK> TOP
<BLANK> TURE
          CAP
          TEMPERA
<BLANK> TY
          CASUAL
          COMMUNITY
          PROPERTY
          SIX
<BLANK> UAL
<BLANK> URE
          ASS
          END
          FEAT
          FIG
          MAN
          PASTURE
          POST
<BLANK> URAL
<BLANK> WARD
          FOR
          WOOD
<BLANK> WIDE
<BLANK> Y
          ALL
          BATTER
          BILL
```

APPENDIX B
SUFFIX DATA LIST

```
    BUG
    BUR
    BUS
    CARRY
    COOK
    COUNTY
    DOWNY
    EARL
    EVERY
    FACTOR
    FAIR
    FORT
    GULL
    HARD
    HAST
    LAD
    LIVER
    MAN
    MAR
    MOB
    MUST
    PARTY
    PEAR
    PENNY
    PIT
    PLAN
    POPY
    READY
    ROMANY
    PUB
    SLIM
    SLIPPER
    SPIN
    TINY
A       ESE
A       IAN
A       ISM
A       ON
    COMMON
ABLE    E
    CAPE
    PROBE
    SUITE
ACTION  Y
AIN     ENANCE
AIN     ISH
AL      E
    CANE
    CHORE
    CORE
    FINE
    MORE
    PARTIAL
    PRACTICAL
    SEVERE

    SPECIAL
AL      ED
    PRACTICAL
AL      IAN
AL      IST
AL      ST
AL      U
AL      Y
AME     OME
AMENT   ED
AN      E
    CUBAN
    MILAN
    RICE
AN      EN
AN      IN
AN      O
ANCE    ED
ANCE    OME
AND     E
    COME
AND     ISH
ANT     E
    TRUE
ARY     E
    CANE
    SALARY
    SECRETE
ARY     ITY
AT      IT
ATE     E
    PRIMATE
ATE     IC
ATION   E
ATION   ED
    FOUNDATION
ATION   ING
AY      EGIAN
BE      PTION
BED     PTION
BILITY  TION
BY      EST
C       E
C       SM
CAL     ST
    TYPIST
CATION  ED
CE      T
    FLEET
    FORCE
    GREECE
    INSTANCE
    PEAT
    PRINT
    SEAT

    SPIT
CE      TIAL
CE      TIST
CE      TLY
CE      X
CH      LAND
CIE     TIC
    POLITIC
CITU    X
CTION   GUISH
CTION   GUISHED
CY      T
    CURRENCY
CY      TE
CY      TIC
    POLICY
D       LY
    CLOSED
    CHILD
    CURLY
    FOLD
    HOLLY
D       MENT
    BASED
D       NT
    AGED
    MEAD
    MISSENT
    PAID
    SAID
    SPED
    SPENT
D       R
    LETTER E
    ARCHER
    BARBER
    BANNER
    DEER
    FLOWER
    FORMER
    LIVER
    MANNER
    MATTER
    NUMBED
    RUBBER
    SHED
    SHOWER
    TOWER
D       RAL
D       SE
    WORD
D       T
    BEAD
    BOLD
    CARD

    CONTENT
    COLT
    DIET
    FEET
    FOOT
    FORGET
    FORT
    GRAND
    HATCHED
    HEAD
    HEARD
    HOOD
    HORNET
    MARKED
    MEAD
    MOLD
    MOUND
    PLOD
    TENT
    THREAD
    WARD
    WILT
DED     SION
    TENSION
DED     T
    CART
    HEAT
    HOOT
    MOLD
    TENDED
DING    T
E       DOM
E       IAL
E       IC
    BASE
    COMIC
    CUBE
    LOGE
    LYRE
    MUSE
    OPERATE
    POLITE
    SLAVIC
    STATE
    TUNIC
E       ICAL
E       ICE
    MALE
    POLE
E       IERY
E       IFIER
E       IFY
    MODE
E       INAL
```

APPENDIX B
SUFFIX DATA LIST

```
E         ING
     BARRING
     CASTE
     DAME
     HUGE
     SKIE
     THEE
     TIDE
     TWINE
E         ION
     DEFINATE
     MILE
     MILLE
     NOTE
     PASSION
     STATE
     VERSE
E         ISH
     FINE
     MOORE
E         IT
     COME
     CUBE
     LIME
     MERE
     SPIRE
E         ITION
E         ITY
     CAVE
     NATIVITY
     POLITY
E         IVE
E         IVITY
E         OU
     CURIOU
     SERIOU
E         PTION
E         T
     BUST
     CARE
     CASE
     DART
     EASE
     FACT
     FORE
     HOSE
     LIFT
     MINT
     MOLE
     MUSE
     PACT
     PARE
     PLANT
     PORE
     POST
```

```
     PRODUCT
     RIFT
     SALE
     SCENT
     SHORE
     SHOT
     SORT
     SPIE
     SPORE
     SQUIRE
     STARE
     SURFACT
     TILE
     VASE
     WARE
E         TEN
     FATE
     MOLE
     SHORE
E         TH
     FIFE
     WORE
E         TION
     CAPE
     FACE
     PARTITION
     PORE
E         TRIL
E         URE
     CREATURE
     MANE
     PASTURE
     PRESSE
     STATE
E         Y
     BUSE
     HEAVE
EAK       OKEN
ED        FIED
ED        IBLY
ED        ING
     MATTER
ED        ION
     MILLED
     MISSED
     NOTED
     PASSION
     STATION
ED        ISON
ED        ITION
ED        ITIONAL
ED        ITURE
ED        IVE
     EXECUTED
     PASSED
```

```
ED        MENT
ED        OR
     MINOR
     PASTOR
ED        SE
ED        SIVE
ED        SIVELY
ED        T
     CART
     DART
     FACED
     LEAST
     MINED
     PLANNED
     POST
     STARED
     TRACT
ED        TION
     FACED
ED        TURE
     POSED
ED        URE
     ENDED
     PASTED
EDED      SSFULL
ED        Y
     CITY
     COOKY
     COUNTY
     MANY
     PENNY
     SCRUBBY
     STORY
     TREATY
ELL       OLD
EN        INAL
EN        ING
     FASTEN
     LINING
     LISTING
ENT       Y
     STUDY
EP.       PT
EP        PTH
ER        EST
     FLOWER
ER        IAL
ER        ING
     FORMER
     INNER
     LETTING
     LIVER
     MATTER
     TREATY
     SHOWER
```

```
     SHUTTER
     SLIPPER
     SPRINGING
     SWEATER
     TOWER
ER        LY
     HARDLY
     SUPPER
ER        PTION
ER        RE
ER        RAL
ER        RIC
     COUNTRY
ER        RY
     COOKER
ER        Y
     HUNGER
     FAIRY
     COUNTY
     MANY
     READY
     SHOWER
ERY       TIC
ETIC      Y
EW        OW
EY        ISH
F         VE
     SERF
F         VOU
FE        VE
     STRIFE
FIC       ST
FIED      TED
FIED      TY
FYING     TY
GUARD     TY
I         U
IAL       IST
IAL       Y
     PARTY
IAN       Y
IC        ISM
IC        OE
IC        OGEN
IC        Y
     ITALY
     TERRIFY
ICAL      OLOGY
ICAL      Y
ICALLY    Y
ICATION   YING
ICE       SE
ID        Y
     PLAY
```

APPENDIX B
SUFFIX DATA LIST

```
IE        YING
IED       ICATIO
IED       Y
IE        Y
IER       Y
IES       Y
IEST      Y
IFUL      Y
IGHT      OUGHT
ILY       Y
IN        UN
INESS     Y
ING       ION
ING       MENT
    COMMENT
ING       TH
ING       UNG
ING       Y
    BILLY
    COOKING
    COUNTY
    KINDLY
    READY
    STORING
    TREATY
ION       IVE
ION       IVELY
ION       OR
    MENTOR
```

```
    MOTION
    VISION
ION       ORY
ION       UAL
ISM       IST
IST       O
IST       Y
ISTIC     Y
IVE       URE
L         TING
    HALL
    FOOL
    LOCAL
LAND      LE
LE        ILITIE
LE        ILITY
LE        ULAR
    PARTICLE
LY        NESS
M         T
    LETTER S
N         T
    BEAT
    CANNON
    CONCERT
    DART
    GREET
    LOON
    MEAT
```

```
    SHORN
    SHUN
    SPAN
    SPIN
    SPURN
    STRAIN
    WARN
N         TICAL
NG        ON
NG        TE
    WRITE
    WROTE
    FLUTE
NT        TE
OM        SSOM
ON        RY
    HURRY
ON        TED
ON        VE
    ACTIVE
    CAPTION
    DERIVATIVE
    DIRECTION
    NATION
    POSITIVE
OR        RAL
OR        MENT
OR        RESS
```

```
OR        RY
OU        Y
OY        UCTION
R         LY
    HOMELY
    LIVER
    SCAR
    WHIRLY
R         ST
    BEAR
    BOAR
    FEAR
    ROAR
    YEAST
R         UR
SH        TURE
SI        TIC
SION      T
    FIST
    MIST
    PAST
SIONARY T
SITY      US
ST        ZE
    BLAST
    ORGANIZE
TH        WARD
TION      ZE
```

APPENDIX C

Function Word List

APPENDIX C
FUNCTION WORD LIST

A
ABOUT
ABOVE
ADO
AFORESAID
AFTER
AGAIN
AGAINST
AH
ALL
ALMOST
ALONE
ALONG
ALREADY
ALSO
ALTHOUGH
ALWAY
ALWAYS
AM
AMONG
AN
AND
ANON
ANOTHER
ANY
ANYBODY
ANYTHING
ANYWHERE
APART
ARE
AS
ASIDE
AT
AWFULLY
BE
BECAUSE
BEEN
BEFORE
BEING
BETWEEN
BOTH
BUT
BY
CAN
CANNOT
CANST
CONCERNING
CONSEQUENTLY
COULD
DID
DO
DOES
DOING
DONE
DON'T
DONT
DURING
EACH
EITHER
ELSE
ELSEWHERE
ENOUGH
ETC.
EVEN
EVER
EVERMORE
EVERY
EVERYONE
EVERYTHING
EVERYWHERE
EXCEPT
FARTHER
FEW
FOR
FORASMUCH
FOREGOING
FOREVER
FORWARD
FROM
FURTHER
FURTHERMORE
GET
GOT
HAD
HARDLY
HAS
HAVE
HAVING
HE
HENCE
HENCEFORTH
HER
HERE
HEREIN
HERETOFORE
HERSELF
HE'D
HE'LL
HE'S
HES
HIM
HIMSELF
HIS
HITHER
HOW
HOWBEIT
HOWEVER
I
IF
IN
INASMUCH
INDEED
INSIDE
INSOFAR
INSOMUCH
INTO
I'D
I'LL
I'M
IS
IT
IT'D
IT'LL
IT'S
ITS
ITSELF
JUST
LATTER
LEST
LIKE
LIKEWISE
MADE
MAKE
MAKING
MAY
ME
MIGHT
MINE
MOREOVER
MY
MYSELF
NAY
NEITHER
NEVER
NEVERTHELESS
NO
NOBODY
NONE
NOR
NOT

APPENDIX C
FUNCTION WORD LIST

NOTE
NOTHING
NOW
NOWADAYS
NOWHERE
O
OF
OFTEN
OFTENTIMES
OH
ON
ONCE
ONE
ONES
ONLY
ONTO
OR
OTHER
OTHERWISE
OUR
OURS
OURSELVES
OVERMUCH
OWN
PER
PERHAPS
QUITE
RATHER
REALLY
SAME
SELF
SELVES
SHALL
SHALT
SHE
SHE'D
SHE'LL
SHE'S
SHOULD
SHOULDEST
SINCE
SO
SOMEBODY
SOMETHING
SOMETIMES
SOMEWHAT
STILL
SUCH
TAKE
THAN
THAT
THAT'D
THAT'LL
THE
THEE
THEIR
THEIRS
THEM
THEMSELVES
THEN
THENCE
THERE
THEREAFTER
THEREBY
THEREFORE
THEREIN
THEREOF
THEREON
THERETOFORE
THEREWITH
THESE
THEY
THEY'D
THEY'LL
THEY'RE
THINE
THIR
THIS
THITHER
THOSE
THOU
THOUGH
THROUGH
THROUGHOUT
THUS
THY
THYSELF
TO
TOGETHER
TOWARD
TRULY
UNDER
UNDOING
UNLESS
UNTIL
UNTO
UP
UPON
US
VERY
WAS
WASN'T
WE
WELL
WERE
WE'RE
WHAT
WHATEVER
WHEN
WHENCE
WHENEVER
WHERE
WHEREAS
WHEREFORE
WHEREIN
WHEREINSOEVER
WHEREOF
WHEREON
WHEREVER
WHEREWITH
WHETHER
WHICH
WHILE
WHILST
WHITHER
WHO
WHOM
WHOSE
WHY
WILL
WILT
WITH
WITHAL
WITHOUT
WORK
WOULD
YE
YEA
YES
YESES
YET
YOU
YOUR
YOURS
YOURSELF
YOURSELVES
YOU'D
YOU'LL
YOU'RE

APPENDIX D

Punctuation List

APPENDIX D
PUNCTUATION LIST

.	PERIOD
,	COMMA
?	QUESTION MARK
;	SEMI-COLON
#	POUND SIGN
!	EXCLAMATION MARK
(	LEFT PARENTHESIS
)	RIGHT PARENTHESIS
'	APOSTROPHE
-	HYPHEN
:	COLON
--	DASH
...	ELLIPSIS
<	LEFT CARET
>	RIGHT CARET
*	ASTERISK
%	PER CENT SIGN
"	QUOTE SIGN
@	AT-SIGN

APPENDIX E

Iterative Use of THESR

APPENDIX E

Iterative Use of THESR

E. 1. Introduction

A recent update has made it possible to use THESR in an iterative process for "comparing" sequential sections of input text in order to accumulate and retain information from one section to the next. This technique is designed to roughly parallel the process a literary critic would go through in reading a text from beginning to end. For example, in reading the first of ten chapters of a text, a critic would theoretically be starting "fresh." That is, the only information available to him would be that contained in the first chapter. However, in reading the second chapter, the critic would have not only the information in that chapter available to him, but also the information from the first chapter. In this way, he would acquire a more complete picture of the text as he read each succeeding chapter, including the introduction of new themes and reference to old ones for each chapter.

Similarly, THESR could be used to process the same hypothetical text from the example above, using the option designated "ITERATE" along with the previously available "SAVE" and "THRESHØLD" options (with which the reader should already be familiar). The input text, after having been processed by previous VIA routines, should be divided into chapters by SELECT. (This is only for the example; other divisions may be used at any level where applicable.) The first THESR run, to process the first chapter, would, like the critic, start "fresh." However, the linked thesaurus information derived from this run would be written to tape and saved (SAVE=YES option).

Processing of the second chapter would include the introduction of this previously saved information to the program before printing of the linked thesaurus trees. Use of the THRESHØLD option to determine "relevance" of thesaurus information is necessary to complete the analogy. As before, thesaurus words occurring with frequency less than that specified by THRESHØLD for chapter one will not appear as root nodes in the printed output for that chapter. However, with the introduction of information from chapter one for processing with chapter two, the THRESHØLD performs a slightly different function. Each thesaurus word from chapter two which did not reach the THRESHØLD requirement for that chapter is checked against the information from chapter one. If the word occurred in chapter one with a frequency exceeding the THRESHØLD, it will still function as a root node for chapter two output. A message to that effect will be printed, and both frequencies shown on the output. At this time, the information which is saved from one run to the next includes only the greatest frequency encountered for a given thesaurus word, and does not specify from which preceding chapter it comes.

Subsequent chapters would be processed similarly, with the introduction for each run of the most recently updated SAVE tape. Figure E.1 illustrates this process. Of course, the last section may be processed without a SAVE tape.

E. 2. Input and Output

In addition to input and output SAVE tapes, each run in the iterative THESR sequence described above will require the same inputs as a normal

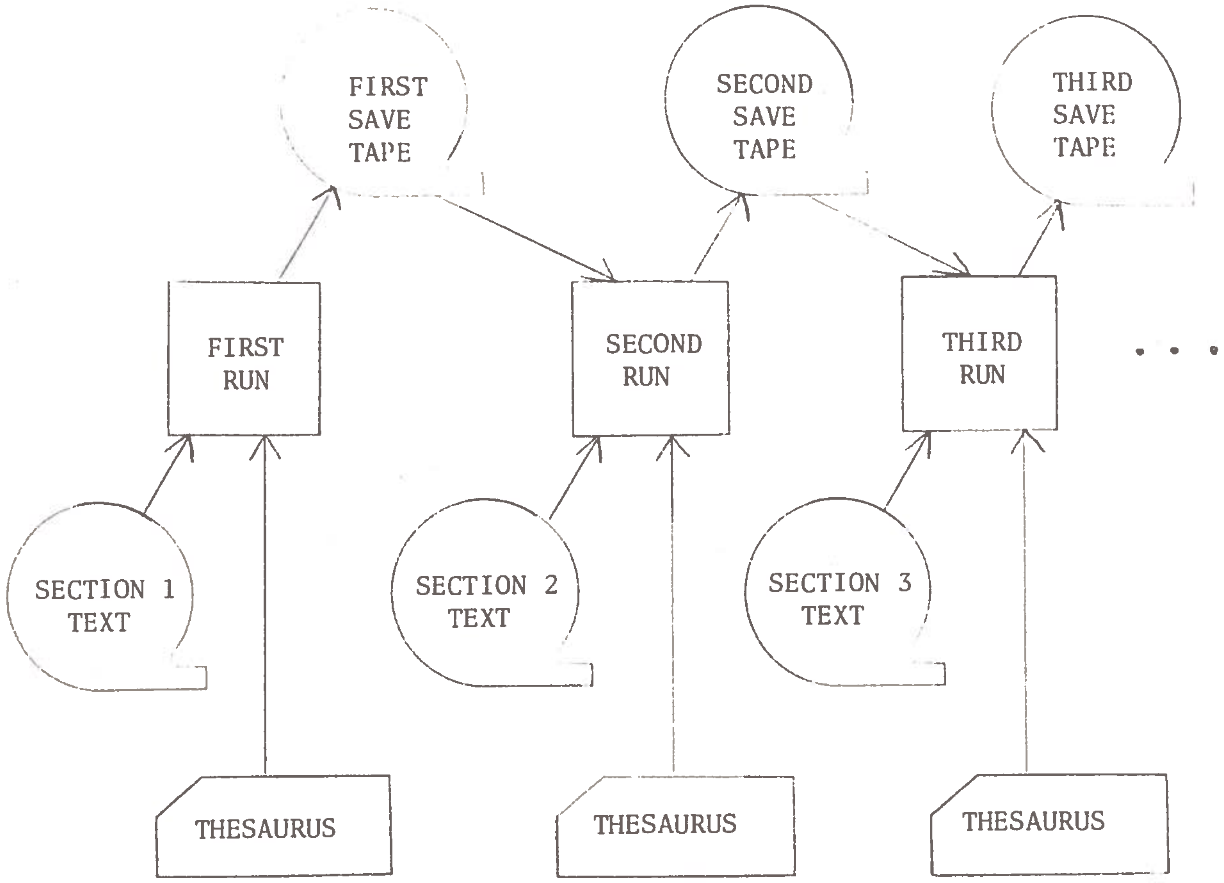


Figure E.1

THESR run. These include the text input tape, the thesaurus, and a parameter card.

The text input has been described as sections of a commonly-indexed text divided by SELECT. For greater flexibility, THESR will also accept text sections which have been separately processed by INDEX, PREFIX, and SUFFIX. No special instructions are indicated for this case.

The input thesaurus will most likely remain the same from one run to the next in an iterative sequence. Again, this is not a necessity, and no special checking procedures are present in the program. Only thesaurus words which match are considered.

The parameter card, illustrated in Figure E.2, has been changed for iterative THESR by the addition of the "ITERATE=YES/NØ" option. This is the last option in sequence, and for purposes of compatibility it may be omitted entirely. ITERATE=NØ is assumed when this is the case.

$$\text{LEVEL=n,SAVE=}\left\{\begin{matrix}\text{YES}\\ \text{NO}\end{matrix}\right\}\text{,THRESHØLD=n,FØRMATS=NØ,ITERATE=}\left\{\begin{matrix}\text{YES}\\ \text{NØ}\end{matrix}\right.$$

Figure E.2

The correspondence of input and output tapes to THESR run parameters can be summarized as follows. For each THESR run, SAVE=YES means an output tape will be created; ITERATE=YES means a previous SAVE tape will be introduced as input. Failure to include tapes in the job deck in exact correspondence to the options of the parameter card will result in either an "abort" (failure of the job to run at all), or in a run which does not produce the desired results.

The THRESHOLD option will most likely retain significance only if it remains constant for the entire iterative sequence. Again no special checking procedures are present because a changing threshold may at some time be desirable and significant.

THESR output for each run remains the same, except for the addition of a message "APPEARS ON SAVED TAPE WITH FREQUENCY = n" for those words which do not exceed the THRESHØLD value in the current section of text, but have previously exceeded that THRESHØLD.

E. 3. Job Deck and Running Suggestions

The job deck for any run in an iterative THESR sequence is illustrated as Job Deck E.3. The only difference from a normal THESR job deck is the addition of an optional input tape for ITERATE=YES (which is in each case an output SAVE tape from the immediately preceding run). The parameter card changes have already been described, as have the relationships of the parameters to the optional input and output tapes.

```
Card Column:        1      8        16

                    $      IDENT    project number,name
                    $      SELECT   2632-SEDELØW/THESR
(optional)          $      LIMITS   time
(text)              $      TAPE     02,X2DD,,tape number,,tape label,IN
(if SAVE=YES)       $      TAPE     03,X3DD,,tape number,,tape label,ØUT
(if ITERATE=YES)    $      TAPE     04,X4DD,,tape number,,tape label,IN
                    $      INCØDE   IBMF
                    Parameter Card
                    $      DATA     01,IBMF,CØPYD
                    Thesaurus Input Deck
                    $      ENDCØPY
                    $      ENDJØB
                    ***EØF
```

Job Deck E.3

The LIMITS card may be included with a time estimate greater than that for a normal THESR run because of the additional processing required by iterative THESR. No time estimates are yet available.

REFERENCES

Joyce, Frank. "Suffix Programmer's Manual," Automated Analysis of Language Style and Structure in Technical and Other Documents, Technical Report No. 1, Contract N00014-70-A-0357-001, Office of Naval Research, University of Kansas, September, 1971, pp. 136-146.

________. "New Utility Routines for Program PREFIX," Automated Language Analysis, Report on research for the period, September 1, 1971-August 31, 1972, Contract N00014-70-A-0357-0001, Office of Naval Research, University of Kansas, pp. 99-105.

Lewis, Peggy. "List-Structure Thesaur," Automated Analysis of Language Style and Structure in Technical and Other Documents, Technical Report No. 1, Contract N00014-70-A-0357-001, Office of Naval Research, University of Kansas, September, 1971, pp. 147-184.

Sedelow, Sally Yeates. Automated Language Analysis, Report on research for the period, September 1, 1971-August 31, 1972, Contract N00014-70-A-0357-0001, Office of Naval Research, University of Kansas. 124 pp.

________. Automated Analysis of Language Style and Structure in Technical and Other Documents, Technical Report No. 1, Contract N00014-70-A-0357-001, Office of Naval Research, University of Kansas, September, 1971. 275 pp.

________. Automated Analysis of Language Style and Structure, Report on research for the period March 1, 1969, to August 31, 1970, Contract N00014-67-A-0321-001, Office of Naval Research, University of North Carolina. 162 pp.

________. Automated Language Analysis, Report on research for the period March 1, 1968, to February 28, 1969, Contract N00014-67-A-0321, Office of Naval Research, University of North Carolina. 286 pp.

V. Professional Activities of Project Personnel

Sally Yeates Sedelow

Publications:

"Networks for Languages and the Humanities," Proceedings of the EDUCOM Fall Conference, October, 1972, pp. 61-65. Co-author with Walter A. Sedelow, Jr.

"Common Themes and Consensus: Report and Discussion of Workshops," Proceedings of the EDUCOM Fall Conference, October, 1972, pp. 151-154.

"Shakespeare Studies and the Computer," in Shakespeare 1971, Proceedings of the World Shakespeare Congress, ed. Clifford Leech and J.M.R. Margeson, University of Toronto Press, 1972, pp. 284-288.

Automated Language Analysis, report on research for the period, September 1, 1971-August 31, 1972, Contract N00014-70-A-0357-0001, Office of Naval Research, University of Kansas. 124 pp.

"Models, Computing, and Stylistics," Current Trends in Stylistics, B. B. Kachru and H.F.W. Stahlke, eds., Linguistic Research, Inc., 1972, pp. 275-286. Co-author with Walter A. Sedelow, Jr.

"Professional Ethics, Standards, and Education," Research Trends in Computational Linguistics, Center for Applied Linguistics, 1972, pp. 87-91.

Papers/Seminars/Addresses/etc.:

"Literary Text Processing," Computational Linguistics Session, National Computer Conference, June, 1973.

"A Schema for Integrating Symbolic Behavior Research in Near Environments," Michigan State University, May, 1973.

"Pattern Recognition in Natural Language Research," Rutgers University, May, 1973.

"Some Research Problems in Computational Linguistics," Michigan State University, October, 1972.

Activities:

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

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

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

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

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

Field Reader of Proposals, U. S. Department of Health, Education, and Welfare, 1966--.

Proposal Evaluation, Canada Council, 1968--.

Proposal Evaluation, National Endowment for the Humanities, 1969--.

Proposal Evaluation, Special Projects Program, NSF, 1970--.

Proposal Evaluation, Division of Social Systems and Human Resources, NSF, 1972--.

Faculty Senate Committee on Scholarly Publications, University of Kansas, 1971--.

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

Resource Faculty Member, Staff, NSF Summer Institute for Computer Science in Social and Behavioral Science Education, University of Colorado, June, 1973.

Invited Participant, Workshop on Computer Perceptions, Attitudes and Literacy, April 30-May 1, 1973, Institute for the Future.

Consultant, College of Human Ecology, Michigan State University, 1973; Seminars for Professional Development of Faculty in Human Ecology, May 14-15, 1973.

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

Chairman, Nominating Committee, Special Interest Committee on Language Analysis and Studies in the Humanities, Association for Computing Machinery, 1973.

Invited Participant, EDUCOM Working Seminar on a National Science Computer Network, December, 1972.

Co-Chairman, Workshop on Languages and Humanities, EDUCOM Fall Conference, 1972.

Reviewer of Papers for FJCC, 1968-1972, and SJCC, 1969-1972.

Walter A. Sedelow, Jr.

Publications:

"Bibliography for a Science of Language," Computer Studies in the Humanities and Verbal Behavior. (Forthcoming, 1973).

Essay Review (on Communicational Analysis and Methodology for Historians, Heller), History and Theory, October, 1973, Vol. XII, No. 3, pp. 358-365.

"A Perspective for World Shakespeare Congress Efforts as to Computers and New Methodologies," Computer Studies in the Humanities and Verbal Behavior, Vol. IV, No. 1, June, 1973.

"Observations on Computer Art: Hardware and Software vs. Aesthetics," Proceedings 7th National Sculpture Conference, National Sculpture Center, Lawrence, Kansas, 1973, pp. 179-181, 191-193.

"The Ce/NCoReL Study," Networks and Disciplines, Interuniversity Communications Council, Princeton, New Jersey, 1973, pp. 20-24.

"Workshop on Languages and Humanities," in Networks and Disciplines, Interuniversity Communications Council, Princeton, New Jersey, 1973, pp. 61-65. Co-chairman with Sally Yeates Sedelow.

Papers/Seminars/Addresses/etc.:

"The Promise and Prospects for Computational Sociolinguistics," Department of Sociology, Michigan State University, October 11, 1972.

"The Ce/NCoReL Project," Department of Computer Science, University of Kansas, 1972.

"Networks (Computer) and Linguistics," Department of Linguistics, University of Kansas, February 20, 1973.

"Computer Networks and Natural Language Research," Department of Applied Mathematics and Computer Science, University of Virginia, March 20, 1973.

"The Ce/NCoReL Study," Plenary Session Paper, EDUCOM 8th Annual Fall Conference, University of Michigan, October 12, 1972.

"Computer Networking for Humantiies Research," International Conference on Computers in the Humanities, University of Minnesota, July 22, 1973.

Activities:

Consultant, College of Human Ecology, Michigan State University, East Lansing, Michigan, February-May, 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--.

Board of Editors, Computer Studies in the Humanities and Verbal Behavior, 1966--.

Series Editor, The Free Press/Macmillan Company, 1968-72.

Advisory Board, Historical Abstracts, 1973-81.

Seminar Leader, Professional Development Series for Faculty in Human Ecology, Michigan State University, May 14-15, 1973.

Resource Faculty Member, Staff, (NSF) Summer Institute on Computer Science in Social and Behavioral Science Education, University of Colorado, Boulder, June 25-29, 1973.

Referee, Social Forces, 1971-72.

Referee, National Science Foundation, (Division of Social Sciences, Program in Sociology), 1973.

Referee, Technical Papers, 1st National Computer Conference and Exposition, 1973.

Member, Advisory Council, Computer Studies Institute, International Academy at Santa Barbara, California, 1972--.

Chairman, Undergraduate Studies Committee, Department of Sociology, University of Kansas, Spring, 1971, and Member, Department Executive Committee; other Departmental Committee responsibilities (Personnel Committee, etc.), 1971--.

Member, University Senate Committee on Lectures and Convocations, University of Kansas, 1972--.

Chairman, University of Kansas Chapter, AAUP Committee on University Government, 1972-73; 1973-74.

Member, Kansas Alpha Phi Beta Kappa Graduate Chapter, Committee on Special Cases, 1972-75.

Principal Investigator, National Science Foundation Study Grant re Possible Center/Network for Computational Research on Language (Ce/NCoReL), Study Grant GJ-28599, 1971-73.

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 66044

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

1 September 1973

7a. TOTAL NO. OF PAGES

7b. NO. OF REFS

8a. CONTRACT OR GRANT NO.

N00014-70-A-0357-0001

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 covers the following topics: the editing of Roget's International Thesaurus, the mathematical modelling of thesauri, and a user's guide to the VIA content analysis programs as implemented at the University of Kansas. Articles in the report concerned with the practical and theoretical issues raised by the effort to edit the Thesaurus include, "The Conversion of Roget's International Thesaurus to an Automated Data Base," by Herbert Harris, "Handling of Bracketed Information," by Scott Taylor, and "Etc. in Roget's International Thesaurus," by Sally Yeates Sedelow. "Abstract Thesauri and Graph Theory Applications to Thesaurus Research," by Robert Bryan explores the mathematical modelling of thesauri. Robert Bryan and Peggy Lewis have provided the user's guide to the VIA programs.

Security Classification

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

Security Classification