Second Annual Conference of the UW Centre for the New Oxford English Dictionary

Advances in Lexicology

Proceedings of the Conference

November 9-11, 1986
Waterloo, Canada

"Thesaural Knowledge Representation"

Sally Yeates Sedelow and Walter A. Sedelow, Jr.
University of Arkansas at Little Rock
Little Rock, AR 72204

ABSTRACT

The adequacy of Roget's International Thesaurus, 3rd Edition, as a representation of our knowledge, and use, of English semantic space is explored in this paper. A distinction is made between the explicit structure of the Thesaurus and implicit structures in the Thesaurus, with emphasis upon the thesaural property of multilocality of word types and tokens. The implications of this property for the disambiguation of homographs are discussed at large, and then with specific reference to a general mathematical model of thesauri which uses Roget's as an instantiation. Selected components of this model are described in order to show that it is possible to design algorithms to elicit 'intuitively' satisfying implicit semantic structures from the Thesaurus. The paper ends with a brief overview of empirical research which has tested the semantic organization -- explicit and implicit -- of the Thesaurus, and with a statement to the effect that any assertions that the Thesaurus is a poor representation of English semantic organization would be ill-founded and, given depth of analysis, would have to be regarded as counterfactual.

* * * * *

There are two types of information to be derived from modeling the structural properties of thesauri: the one from explicit structure, the other from implicit structure. Although our research has developed an abstract model applicable to thesauri in general, our primary focus has been upon a particular instantiation, Roget's International Thesaurus, 3rd Edition (1962).

The explicit structure of the printed Thesaurus is as follows: One can think of the basic text as comprising 1040 semantic categories, each with a number and a label, e.g., 515. Truth. Each of these numbered categories is divided syntactically and semantically. The semantic sub-categories are numbered consecutively following the decimal point, but are not labeled. The syntactic labels occur with the

N.B.: This research is currently supported by the National Science Foundation (Intelligent Systems); earlier phases were in part supported by the Office of Naval Research.

numbered sub-categories and indicate that all following sub-categories belong to that syntactic category until another syntactic label is given in the sequence. The sub-categories consist, evidently, of the words which are considered semantically related. The words in these sub-categories are further divided into semicolon-delimited sub-sub-categories.

The uppermost level in the hierarchy of levels in the Thesaurus given in the "Synopsis of Categories," which is not part of the basic text but is presented as an outline following the Preface, is divided into eight classes. Each class is divided into several labeled sub-classes indicated by Roman numerals, and each sub-class is divided into labeled sub-sub-classes designated by capital letters. Each sub-sub-class is divided into several of the 1040 categories, which are numbered consecutively throughout the text.

In the course of this research project, initiated in the latter 60's, various explicit hierarchies have been proposed. One, by Martin Dillon and David Wagner (in S. Y. Sedelow, et al., 1970) proposes six levels based on the formal structure presented in the Thesaurus itself. Here is their example, based on an occurrence of the word perfect:

Class Six: Intellect
I. Intellectual Facilities and Properties
L. Conformity to Fact
515. Truth
Adjectives
515.14
(perfect)

One occurrence of the word perfect is to be found in a semicolon group in sub-category 515.14, which is one of several sub-categories of adjectives under category 515. Truth. Category 515. Truth is found in the "Synopsis of Categories" under the letter L. Conformity to Fact, which is in turn a part of Roman numeral I. Intellectual Facilities and Properties, and in turn that is a division of Class Six: Intellect.

Owing to its relevance for both explicit and implicit structure and for the applied utilizations of thesauri in natural-language computing, emphasis should be placed on an important property of the Thesaurus: consider a word (e.g., lead) and all its homographs; let each separate meaning of a word be called a type (e.g., lead [to conduct, etc.] and lead [a mineral] are two separate types; together they constitute homographs). The occurrences of each type are called its tokens. Now, the thesaural property to be

emphasized is the mathematical (e.g., topological or graph-theoretic) implicature of the fact that tokens are not unilocal; they may occur in more than one place and sometimes occur in many; and those locales are not necessarily in the same structural neighborhood within the thesaurus. The mathematical and analytical problem is significantly compounded by the necessity of dealing with the occurrence of multilocality both among the tokens of a given homographic type and among the types, themselves, within any given homograph. The attractive implications of this sticky technical property (or 'problem'), however, include the possibility of the utilization of this same property as an effective and powerful disambiguator in natural language computing.

Remember that Roget's International Thesaurus is a six or seven-tiered (depending upon how lower levels are described) hierarchy. One factor that has impeded an accurate perception of the utilities in employing the Thesaurus in natural-language computing is that the eight categories comprising the top tier are more problematic vis-a-vis a model of semantic space (for English) than, for example, the lower-tiered semi-colon groups. The significance of the Thesaurus seems sometimes to have been evaluated on the basis of the partitioning of English semantic space by the upper tiers without attention to the much less problematic semantic significance of the lower tiers. For an exposition of empirical research calling attention to the much greater descriptive and analytical power, semantically, of the Thesaurus in the lower tiers, see S. Y. Sedelow (1985) and W. A. Sedelow (1985) as well as S. Y. Sedelow, et al., passim, 1965 forward, and S. Y. Sedelow and W. A. Sedelow (1986). For example using the lower tiers of the explicit structure, we have achieved intuitively satisfying results when dealing with the problem of prefixation. Our assumption was that for those words sharing the same stem, where one or more of the words is a prefixed form, occurrence of those words in the same or nearby sections of the Thesaurus probably would identify the prefixing function of the initial character strings. As is of course well known, such identification is important for computer-based stem or root identification programs -- which, in turn, are essential to many NLC (natural language computing) application programs. Although not perfect for this type of performance, the Thesaurus worked quite well in this way. Happily, there are many word pairs in the same 5th or 6th level categories (5th or 6th level depending upon the structure employed) which would be regarded as properly paired; examples include BUY-REBUY, JUNCTION-CONJUNCTION, MOLAR-PREMOLAR. By contrast, and all to the good, PREVENT, for one example, was analyzed as non-prefixed because the word PREVENT does not occur in any of the categories associated with the unprefixed root VENT.

In some, if not all, of the thesauri which would be generated in accordance with the systematic realization of one or more mathematical properties of thesaural models, an interesting result could be the disappearance of the viability of the distinction between explicit and implicit properties of a thesaurus.

To the best of our knowledge, heretofore there has been no effort to constitute -- and to explore the possibilities of such a constituting -- one or more possible thesauri designed to exemplify one or more of the mathematical properties of thesauri. It is not surprising that such an option has not been explored since, in some significant measure, the option becomes apparent only through exploration of properties of a mathematical model of thesauri. And, so far as we have been able to establish, no thesaurus has ever been looked at as a mathematical construct.

As soon as one has built a (thick) mathematical model of a thesaurus, it then becomes technically manageable to consider possible thesauri systematically. In the case at the extreme, combinatoric possibilities entailed in the model would be run through and examined in an orderly and comprehensive way (W.A. Sedelow, 1985). Short of so ambitious -- but nonetheless desirable -- a project, we may elect to realize specific thesaural possibilities in mathematically pure form.

As part of our research effort to date, we have applied a formal characterization of a thesaurus structure to Roget's International Thesaurus, 3rd Edition. Although trivially dependent upon the explicit hierarchy for the words forming the terminals, and less trivially for determination of which tier within the Thesaurus will form the categories, a principal component of this model's attractiveness derives from its ability to capitalize upon the multilocal occurrence of types and tokens within the Thesaurus. Thus, this model provides a way (or ways) of getting at the implicit structure of the Thesaurus by looking at connectivity patterns cutting across explicit hierarchial branches. Among other aspects of the model, description of connectivity in terms of chains, from the most general, Type 1, through the most restricted, Type 10, has led to some promising research results. Specifically, restricted Type 9 and Type 10 chains, working within the syntactic categories, provide sufficient power of resolution so as to prove strong disambiguators among homographs within a text.

The initial, basic model is the work of Robert Bryan, Computer Science, San Francisco State University (S. Y.

Thesaural Representation

Sedelow, et al., 1973), and we are indebted to both Jerzy Gryzmala-Busse (Computer Science, University of Kansas) and John Talburt (Computer Science, University of Arkansas/Little Rock) for further analysis and description of components of the model. Talburt's examples have been particularly helpful and, with his permission, we use them within this exposition.

Bryan considers Roget's International Thesaurus as an instantiation of an abstract thesaurus, under at least one and probably several interpretations, in terms of the definable elements present in the Thesaurus. He defines a thesaurus as follows:

Let T = (E, W, C)

Thus, in the T-Graph in Figure 1 below

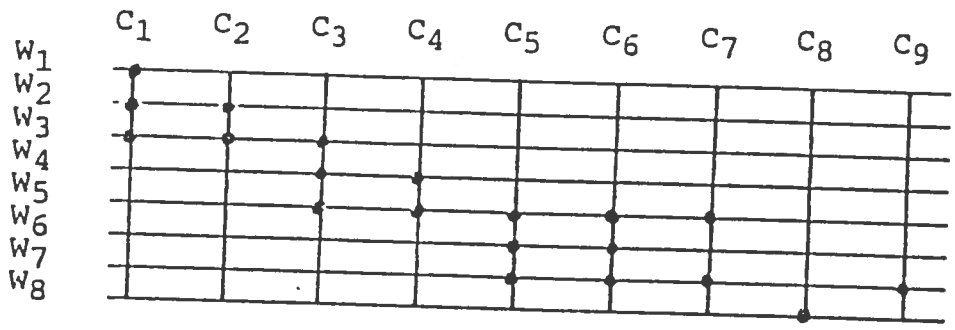

Figure 1

E would be the set of entries $\{e_{11}, e_{21}, e_{22}, e_{31}, e_{32}, e_{33}, \ldots e_{88}\}$. W would comprise the words $\{w_1, \ldots w_8\}$. You see that by "word," Bryan means the type, rather than the tokens of any type. As we have already noted, Roget's International Thesaurus of course has tokens of types; but through increasingly restricting the chains in his model, Bryan is able to disambiguate among homographs representing different types. Finally, to refer again to Figure 1, C would be the set of categories $\{C_1, \ldots C_9\}$.

Using these three basic elements, Bryan builds up further definitions, including that of M, which stands for Molecule, which is W U C, or all the words and categories denoted by the entries.

In this sample microthesaurus, the number of entries |E| is twenty, the number of words |W| is eight, and the number of categories |C| is nine.

For a discussion of range, let us assume a set of entries, E_1 as follows: $\{e_{11}, e_{21}, e_{22}, e_{43}, e_{48}\}$. The range of the molecule for the set of entries, E_1, is the set consisting of $\{w_1, c_1, w_2, c_2, w_4, c_3, w_5\}$. The range of words

would be the set $\{w_1,w_2,w_4,w_5\}$. The range of categories would be the set $\{c_1,c_2,c_3\}$.

Next, Bryan takes up the notion of chains within a thesaurus. An example of an e-chain based on Figure 1 could be e_{32},e_{55},e_{57}. In other words, the items in an e-chain can be anything from the set of entries in the thesaurus.

An example of an m-chain would be: c_1,w_3,c_4; clearly, entries in an m-chain can be any molecules, including both categories and words.

A c-chain might be: c_3,c_5,c_6, and a w-chain: w_2,w_4,w_6. Thus, a c-chain can consist of any of the categories, and a w-chain, of any of the words.

Given these general definitions, Bryan then proceeds to define ten types of chains, E^1 through E^{10}, moving from the most unrestricted to the most restricted. In order to understand these definitions, you must bear in mind that a link is an ordered pair, so that direction of movement through the link is significant, whereas a block is a link, or connector in which the direction along the link does not matter; therefore, for a block, order is not significant. Again, these examples will be based upon Figure 1 above.

An E^1 or $\langle E\rangle_1$ chain, the most unrestricted, is any chain over E, e.g., e_{79},e_{55},e_{21}. With the possible exception of individuals who enjoy dipping into texts, including lexicons, almost at random, this chain is of no interest.

An example of an $\langle E\rangle_2$ chain would be: $e_{11},e_{21},e_{22},e_{22},e_{21},e_{31},e_{21},e_{22},e_{32}$. The restriction placed on this chain is that the connections between entries are uniform, which is to say that consecutive elements are in the same word of the same category. From a computational point of view it is unattractively possible to repeat endlessly the same entry, the same link or the same block, or any combination of these possibilities.

An $\langle E\rangle_3$ chain has the additional restriction that connections between entries are not symmetric; this restriction prevents consecutive repetition of the same entry. It is, however, still possible to repeat the same links, blocks, or both. An example of an $\langle E\rangle_3$ chain is: $e_{11},e_{21},e_{22},e_{21},e_{31},e_{21},e_{22},e_{32}$.

With the $\langle E\rangle_4$ chain, the restriction of finiteness is introduced by prohibiting the repetition of links (remember that links are ordered pairs). Thus, although it is still possible to traverse a connection between entries in both directions, it is no longer possible to circle back through

such a connection. An example of an $\langle E \rangle_4$ chain is: e_{32}, $e_{33}, e_{43}, e_{33}, e_{32}$.

The $\langle E \rangle_5$ chain adds the additional constraint that blocks are not repeated. Hence, once a connection has been traversed in <u>either</u> direction, it may not be crossed again. An $\langle E \rangle_5$ chain could look as follows: $e_{43}, e_{44}, e_{54}, e_{53}, e_{43}$.

In an $\langle E \rangle_6$ chain, in addition to the restrictions placed on chain types $\langle E \rangle_2$ through $\langle E \rangle_5$, the connections must be pairwise distinct, which is to say that no element may be repeated. Thus an $\langle E \rangle_6$ chain might be $e_{11}, e_{21}, e_{31}, e_{32}$.

For the discussion of types $\langle E \rangle_7$ through $\langle E \rangle_{10}$ chains, it is necessary to understand the <u>induced</u> chain. Take, as an example, the following $\langle E \rangle_3$ chain: $e_{11}, e_{21}, e_{22}, e_{32}, e_{31}, e_{21}, e_{22}$. Traversal of the connection between e_{11} and e_{21} <u>induces</u> category c_1. Likewise, traversal of the connection between e_{21} and e_{22} induces word w_1. Figure 2 shows the relationship between the entries in this chain and the induced categories and words:

$$e_{11}, \underset{c_1}{|} \quad e_{21}, \underset{w_2}{|} \quad e_{22}, \underset{c_2}{|} \quad e_{32}, \underset{w_3}{|} \quad e_{31}, \underset{c_1}{|} \quad e_{21}, \underset{w_2}{|} \quad e_{22}$$

Figure 2

In summary, the induced m-chain from the sample $\langle E \rangle_3$ chain is: $c_1, w_2, c_2, w_3, c_1, w_2$. The induced c-chain is: c_1, c_2, c_1. The induced w-chain is: w_2, w_3, w_2.

Bearing this information in mind, we can now proceed to define an $\langle E \rangle_7$ chain as an $\langle E \rangle_6$ chain but with the additional restriction that the <u>induced</u> chain is either non-word-repeating <u>or</u> non-category-repeating. Thus the $\langle E \rangle_6$ chain: $e_{11}, e_{21}, e_{31}, e_{32}$ induces the following m-chain: c_1, c_1, w_3. Since the induced m-chain is non-w-repeating, this chain qualifies as an $\langle E \rangle_7$ chain.

As one might expect, the type 8 E-chain prohibits the repetition of both words and categories in the induced m-chain. An example would be the following: $e_{11},e_{31},e_{33},e_{43},e_{44}.e_{54}$, which induces the non-repeating m-chain c_1,w_3,c_3,w_4,c_4. You will note that a trace of the chain of entries produces a turn at the end of each connector, so that the induced m-chain alternates words and categories.

For the discussion of types 9 and 10 E-chains, the notion of strength must be introduced. A connector e_1,e_{i+1} is strong if $|r(e_i) \cap r(e_{i+1})| > 1$ (remember that entries range over words and categories). So that in the sample $\langle E \rangle_8$ chain above, the connector $\langle e_{11},e_{31} \rangle$ is weak, $\langle e_{31},e_{33} \rangle$ is weak, $\langle e_{33},e_{43} \rangle$ is weak, but $\langle e_{43},e_{44} \rangle$ is strong, as is $\langle c_{44},e_{54} \rangle$. Looking at Figure 1, we see that there are at least two -- in this case exactly two -- parallel lines between words or categories. Another way of describing strength is to say that a strong link exists where at least two categories contain more than one word in common, or at least two words contain more than one category in common.

We are now ready to define an $\langle E \rangle_9$ chain as an $\langle E \rangle_8$ chain with the additional restriction that <u>either</u> every c-connector is strong <u>or</u> every w-connector is strong. Thus the $\langle E \rangle_9$ chain: $e_{44},e_{54},e_{55},e_{65}$ induces c_4,w_5,c_5. The induced c_4 and c_5 connectors are strong while the w_5 connector is weak; thus this chain satisfies the requirements for an $\langle E \rangle_9$ chain. This type of $\langle E \rangle_9$ chain, for which the parallel connectors are in the vertical plane of the T-graph, is said by Bryan to be word-strong. An $\langle E \rangle_9$ chain for which the parallel connectors are in the horizontal plane is said to be category-strong.

The most restricted chain defined by Bryan, the $\langle E \rangle_{10}$ chain, is an $\langle E \rangle_8$ chain where <u>all</u> the connectors are strong. As an example, the chain: $e_{55},e_{75},e_{76},e_{66},e_{65}$ induces c_5,w_7,c_6,w_6, all of which are strong.

Homographs are defined as follows: $\langle e_i,e_j \rangle$ are homographs if and only if, first, $\langle e_i,e_j \rangle$ is a word-connector (i.e., on the horizontal rather than the vertical plane of the T-graph), and second, there does <u>not</u> exist a type-10 chain connecting e_i and e_j (i.e., there is not a connector in parallel to e_i and e_j). If we again look at Figure 1, we observe that e_{31} and e_{32} are <u>not</u> homographs since $\langle e_{31}, e_{33} \rangle$ is itself a strong connector; in contrast, e_{32} and e_{33} <u>are</u> homographs because they do not form a strong connection. Hence, e_{32} and e_{33} represent two different word types, whereas e_{31} and e_{33} are tokens of a single type. Intuitively it makes sense to suppose that strong connections represent greater semantic clustering than do

weak connections; therefore, in the abstract, Bryan's approach to homography would seem to be on the right track and, in fact, empirical results (e.g., discussed below) based on his definitions support his approach.

Two more concepts -- Star and Neighborhood -- should be mentioned. A mathematical formulation is provided for chains radiating out from a given entry, thus forming a Star. A 'pure' Star allows chains to be of length r or less, and $\bar{\text{Star}}$ demands that chains be of exactly length r. More formally:

$$S_r^n(e)=\{\langle\varepsilon\rangle\varepsilon E^n \mid \langle e\rangle|\leq r \text{ and } \langle e\rangle \text{ emanates from } e\}$$

$$\bar{S}_r^n(e)=\{\langle\varepsilon\rangle\varepsilon E^n \mid |\langle e\rangle|=r \text{ and } \langle\varepsilon\rangle \text{ emanates from } e\}$$

For example, again using Figure 1 as the reference,

$$S_3^6(e_{21})=\{\langle e_{21}\rangle,\langle e_{21},e_{11}\rangle,\langle e_{21},e_{22}\rangle,\langle e_{21},e_{31}\rangle,\langle e_{21},e_{22},e_{32}\rangle,\langle e_{21},e_{31},e_{32}\rangle,\langle e_{21},e_{31},e_{33}\rangle\}$$

$$\bar{S}_3^6(e_{21})=\{\langle e_{21},e_{22},e_{32}\rangle,\langle e_{21},e_{31},e_{32}\rangle,\langle e_{21},e_{31},e_{33}\rangle\}$$

Informally, Neighborhoods comprise whatever the arms of the Star cover. More formally:

$$N_r^n(e)=U_{\langle\varepsilon\rangle\ \varepsilon S_r^n(e)}\langle\varepsilon\rangle$$

$$\bar{N}_r^n(e)=U_{\langle\varepsilon\rangle\ \varepsilon \bar{S}_r^n(e)}\langle\varepsilon\rangle$$

Using Figure 1 and the Star example above:

$$N_3^6(e_{21}) = \{e_{21},e_{11},e_{22},e_{31},e_{32},e_{33}\}$$

$$\bar{N}_3^6(e_{21}) = \{e_{21},e_{22},e_{32},e_{31},e_{33}\}$$

Although by no means comprehensive, this material from Bryan's model is intended as necessary background for brief summaries of research based on his model. Here, we focus on results from two theses. The first (Dale, 1979) concentrated on type 6 e-chains, showing that they are not sufficiently restricted to serve as effective modellers of semantic space. The second (Patrick, 1985) used the most restricted types 9 and 10 chains, with results suggesting yet again -- as does much of our other empirical

yet again -- as does much of our other empirical investigation of Roget's -- that Roget's can be very useful (see final summary below).

Using nouns from word pairs taken from the similarities test making up part of the Wechsler-Bellevue intelligence test, Dale used type 6 e-chains to grow neighborhoods. It is possible, of course, that the word pairs, themselves, were responsible for the unsatisfactoriness (in terms of the desired 'intuitively' pleasing, to a native speaker, semantic clusters) of the neighborhoods. But our conclusion is that the inability of type 6 e-chains to deal with homographs simply produced associational patterns which are not useful, at least for most applications. For example, the word "jigger" (an archaic word for "bicycle") introduced many words having little relationship to the initiating pair "wagon" and "bicycle" -- unless, of course, one were giving the operators of such vehicles sobriety tests.

Patrick's thesis focused, as noted, upon types 9 and 10 e-chains. His work used a verb sub-thesaurus (of about 21,000 entries) from Roget's created by Christopher Gunn. Patrick observes that:

> Roget's International Thesaurus (R.I.T.) is designed to have the most closely semantically related words in the same semicolon groups. Therefore, the words "love", "affection", "attachment", "devotion", and the prefix and suffix "philo-" and "-philly" are closely related to each other in the English language according to R.I.T. Making a neighborhood of semantically related words from the entry "love" would have to include at least the rest of the words of the semicolon group. This would mean that from the word "love" e-chains would have to emanate from at least the links ("love", "affection"), ("love", "attachment"), ("love", "devotion"), ("love", "philo-") and ("love", "philly"). These links according to definition are semantically strong, and according to the model would be considered in the set of type 8 e-chains, or proper chains. The question is where to go from here to find relatedness. There are two choices: one, to use the hierarchy as a guide; two, to use the duplicate entries in the "word" set to "navigate" through R.I.T. To perform the latter we must be careful not to take a link which is a homograph. The avoidance is accomplished in the abstract model through giving the links between categories strength, that is the degree of overlap of two categories and the entries in each category. If the intersection of entries of two different categories is greater than one,

then there is strong w-link between them. This does get rid of the "navigational" errors caused by homographs (Patrick, 1985).

Patrick proceeded by constructing a Sub-T-graph of verbs, and navigating through them as follows:

> Once the Sub-T-graph is constructed to the level specified, two routines similar in concept are used to trace and mark the Sub-T-graph. First, I provide a description of the way type 10 chains are marked since the process is simpler and will help convey the concept better than if type 9 chains are discussed first.
>
> A type 10 e-chain is one which has strong links, that is strong w-links and strong c-links. To trace these chains and mark them in the abstract data type, the algorithm starts at the first entry placed in the structure, marking the node with a "1". A w-link is taken if the degree of overlap between two categories in which a word has entries is greater than one. If the w-link is taken the entry is marked with a "1". A c-link is then taken from this entry to another entry if the degree of overlap between the two words where the entries belong is greater than "1". This alternating pattern of navigation by taking c-links and w-links is continued until no more advances can be made in the Sub-T-graph.
>
> Then from each entry, other possible links are attempted by skipping over the last-tried entry member of the same molecule to the adjacent entry. That is if a c-link was attempted, the next attempt will be to make another c-link from the adjacent entry following the one previously tried. Thus by navigating through the Sub-T-graph it is possible to extract all words that are members of the type 10 unbounded neighborhood of the original entry.
>
> The next step is to find all other stars for the word where the original entry belongs; this is done by starting the marking and tracing at the next entry on the Sub-T-graph that belongs to the same word as the original entry, and which has not been marked with a "1". We now mark the entries that belong to this next star with a "2" and proceed as before. The entries for each set of stars give us the respective neighborhoods.
>
> The type 9 e-chains are followed using the same algorithm as for type 10 e-chains, except that

> this navigation must be done in two passes. The first pass will mark the strong w-link chains only. That is when taking w-links the links must be strong, but when taking c-links the links need not be strong though they may be. The second pass marks all the chains with strong c-links. This process is repeated as it was for type 10 stars, by marking all possible type 9 stars which encompass all entries of the same word the original entry belonged to. This two-step process was eliminated early in the research, because it was immediately obvious that type 9c neighborhoods were vulnerable to homographs. Therefore, it is not desirable to include these neighborhoods when building semantically connected groups of entries (Patrick, 1985).

Patrick reports that word-strong (type 9-w) neighborhoods sorted-out such homographs as "inspire" meaning to raise the spirits, inspirit, etc., and "inspire" meaning to inhale, breathe in, sniff, etc. Type 10 neighborhoods also dealt with homographs effectively. Intuitively satisfying results, such as the separation of "sauce" meaning to make a sauce or season a dish, from "sauce" as related to insolence, or the separation of "question" meaning to doubt from "question" as in asking a question, emerged quite consistently. Other words such as "powder", "object", "magnetize", etc. were subjected with success to Patrick's procedure.

Type 8 neighborhoods, in Patrick's experience, grow to unmanageable size, but Type 9-w and Type 10 neighborhoods seem not to pose this problem. As Patrick says:

> A large percentage of the entries have unmanageable type 8 neighborhoods, and a restraint must be placed on the growth of these neighborhoods. Type 9 and type 10 neighborhoods, on the other hand, seem quite manageable, and most unbounded type 9 strong w-link neighborhoods seem to emanate no further than about six levels, and most of the time less. All type 9 and 10 neighborhoods seem to emanate no further than about six levels, and most of the time less. All type 9 and 10 neighborhoods are semantically related, and their semantic connectivity is relatively strong. Also semantic connectivity is highly correlated to the actual meaning and sense of the words; this correlation contrasts with the lack of such a connection in most of the deeper levels of type 8 <e> chains. Claudia Dale (Dale, Thesis, 1979) showed that "bicycle" was unfortunately related to "thingamajig" and "jigger" in a type 6

neighborhood, but a type 9 strong w-link neighborhood would never associate the three words (Patrick, 1985).

In conclusion, we point out that we have now tested Roget's International Thesaurus, 3rd Edition, in the following ways: as a guide to content in Hamlet and in translations of Soviet Military Strategy (Sedelow and Sedelow, 1969, and S. Y. Sedelow, et al. 1966, 1967); as a guide to when initial strings of letters in a word are serving as prefixes (Warfel, 1972 and S. Y. Sedelow, 1969); as an instantiation of an abstract model (described in this paper); as a guide -- using a computer algorithm to move from textual context to thesaurus categories -- to classification of medical abstracts (research in progress); and as a guide (for an Interlingual-Communication Support System) to contrasting the creation and partitioning of semantic spaces in pairs of languages (English-Chinese, research in progress). As a result of this work, we feel that a strong case can be made for Roget's International Thesaurus, 3rd Edition as an effective resource for research ranging from the pure to the applied. In fact, our experience leads us to suggest that any assertion to the contrary would have to be construed as simply counterfactual. We hope now that it will be possible to cross-fertilize dictionary and thesaurus research and development, so that the semantic associational patterns forming a thesaurus such as Roget's and the lexical feature patterns provided by dictionaries can be utilized for more effective computer-based natural language analyzers and synthesizers.

References

Bryan, Robert M. (1973), "Abstract Thesauri and Graph Theory Applications to Thesaurus Research," in Sally Yeates Sedelow, ed., Automated Language Analysis, 1972-3. (University of Kansas, Departments of Computer Science and Linguistics; Lawrence); also U.S. Defense Documentation Center, #AD 774-692, 303 pp.

Dale, Claudia J. (1979), Investigations of Nouns in Roget's International Thesaurus, Master of Science Thesis, (Unversity of Kansas, Department of Computer Science; Lawrence).

Dillon, Martin and David Wagner (1970) in Sally Yeates Sedelow, et al, Automated Analysis of Language Style and Structure (University of North Carolina; Chapel Hill); also U.S. Defense Documentation Center #AD 711-643, 162 pp.

Patrick, Archibald B. (1985), An Exploration of an Abstract Thesaurus Instantiation, Master of Science Thesis, (University of Kansas, Department of Computer Science, Lawrence

Roget's International Thesaurus (1962), (Thomas Y. Crowell, New York).

Sedelow, Sally Yeates, et al. (1965), Updating of THESAUR Program, Defense Documentation Center #AD 613-291; (1966) Stylistic Analysis, Second Annual Report #AD 629-789; (1967) Stylistic Analysis, Third Annual Report #AD 651-591; (1968) Automated Language Analysis #AD 666-587; (1969) Automated Language Analysis #AD 691-451; (1970) Automated Analysis of Language Style and Structure #AD 711-643; (1971) Automated Analysis of Language Style and Structure in Technical and Other Documents #AD 735-134; (1972) Automated Language Analysis #AD 752-508; (1973) Automated Language Analysis #AD 774-692; (1974) Automated Language Analysis #ADA002-463.

Sedelow, Sally Yeates (1969), "PREFIX", in Sally Yeates Sedelow, ed.Automated Language Analysis, 1968-1969, (University of North Carolina/Chapel Hill, Departments of English and Computer & Information Science); also [U.S.] Defense Documentation Center, #AD 691-451, pp.286.

Sedelow, Sally Yeates (1985), "Computational Literary Thematic Analysis: The Possibility of a General Solution," in Carol Parkhurst, ed., Proceedings of the 40th ASIS Annual Meeting, 22. 359-362.

Sedelow, Sally Yeates and Walter A. Sedelow, Jr. (1969), "Categories and Procedures for Content Analysis in the Humanities", in George Gerbner et al, edd. The Analysis of Communication Content, (John Wiley & Sons, New York), pp. 487-499.

Sedelow, Sally Yeates and Walter A. Sedelow (1986), "The Lexicon in the Background", Vol. I, No. 2 (In Press), Computers and Translation, ed. Winfred Lehmann.

Sedelow, Walter A. (1985), "Semantics for Humanities Applications: Context and Significance of Semantic 'Stores'", in Carol Parkhurst, ed., Proceedings of the 40th ASIS Annual Meeting, 22. 363-366.

Warfel, Sam (1972), "The Value of a Thesaurus for Prefix Identification", in Sally Yeates Sedelow, ed., Automated Language Analysis, 1971-72, (University of Kansas, Departments of Computer Science & Linguistics, Lawrence); also, [U.S.] Defense Documentation Center, #AD 752-508, 124 pp.