Talk given in Darmstadt, March 1 1996

Sally Sedelow

Although the title of this talk is "Thesauri and Formal Concept Analysis," so far as thesauri are concerned the focus will be upon the classic thesaurus, which has been the model for many subsequent thesaural efforts: Roget's International Thesaurus (3rd edition). → 1962

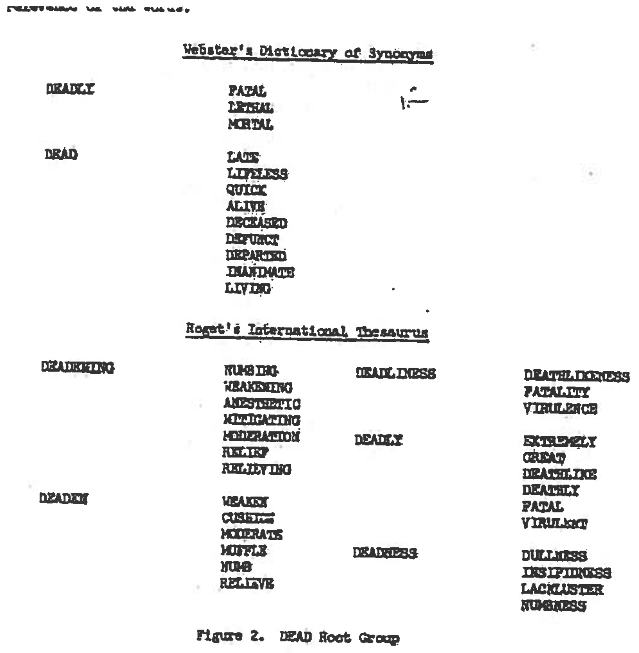
The motivation for my initial involvement (as a professor of literature) with Roget's was a desire to discuss literature with sufficient rigor so that my students would have some sense of a replicable methodology THEY could use for the appreciation of literature. At the time (early 1960's), I was interested primarily in the written text, but not specially in the syntax of that text; rather, it was the semantics, the meaning and the way it was structured that captured my attention.

As luck would have it, in the early 1960's I found myself in a computational setting which prompted me to try to use the computer to push toward greater rigor in the study of literature. A Shakespearean scholar named Caroline Spurgeon had written a multi-volume treatise on chains of images in Shakespeare's plays; chains such as "rotten, disease, decay, death" that one finds, for example, in Hamlet.

I decided to begin my efforts by designing a program to look for such chains of words; obviously, the chains were perceived as connected words, and the relation connecting the words was semantic. Caroline Spurgeon had used her own knowledge of English and of Shakespeare to produce these chains; I wanted a resource other than my own memory so as to automate more of the procedure and, thus, make it more ruleful. Since the resource needed to be based on words placed in structures reflective of semantic relationships, I looked to thesauri and synonym dictionaries for help. Initially, in looking at Hamlet, I simulated an automated look-up procedure using Webster's Dictionary of Synonyms, Roget's, and Brown's List of Scientific Words. The VIA (Verbally-Indexed Associations) program then produced output such as in Figure 1.

I used the results from this system as the basis of a paper given at the World Shakespeare Congress in Vancouver; the scholars felt that the VIA program had turned up the major themes/motifs in the play that had been noticed over the many span of years during which Hamlet had been an object of literary interpretation, but also there were some shifts in emphases which no one had ever discussed in print but which were interesting once pointed out. So here was an early very encouraging validation of the use of such resources but, of course, since I had used a number of lexicons, I could not say which was the most promising for an automated system (as that time – early 1960's – putting such lexicons into computer-accessible form was a major undertaking; hence, I wanted to select just one, at least for starters).

Next, I conducted a rather extensive comparison (more rigorously extensive, I believe than anything hitherto) of Webster's Dictionary of Synonyms and of two thesauri, Roget's International Thesaurus, 3rd ed. and the University Thesaurus. Both of these thesauri are conceptual thesauri, which is to say that there is a hierarchical structure, moving at the top from the most general or abstract to, at the bottom, words more restricted in meaning. Also, as you know, the groupings at the bottom are based upon those words which are most closely related semantically. So, in effect, at the bottom of the hierarchy, you have a kind of dictionary of synonyms; but it is not alphabetical, rather it is located according to the concepts further up the tree. For this comparison, I shifted away from a literary text and looked at a translation of an entire chapter of a work entitled, in translation, Soviet Military Strategy. The search keys included all the words in the same root group as Dead, Decline, and so on up to ten groups. I then looked those words up directly in the alphabetic Webster's Dictionary of Synonyms and I used the indices in the thesauri as guides to the entries there. Figure 2 shows a sample for the root group "Dead" of the word lists which were then submitted to the VIA program.



relevance of the words.

Webster's Dictionary of Synonyms

DEADLY	FATAL LETHAL MORTAL
DEAD	LATE LIFELESS QUICK ALIVE DECEASED DEFUNCT DEPARTED INANIMATE LIVING

Roget's International Thesaurus

DEADENING	NUMBING WEAKENING ANESTHETIC MITIGATING MODERATION RELIEF RELIEVING	DEADLINESS	DEATHLIKENESS FATALITY VIRULENCE
		DEADLY	EXTREMELY GREAT DEATHLIKE DEATHLY FATAL VIRULENT
DEADEN	WEAKEN CUSHION MODERATE MUFFLE NUMB RELIEVE	DEADNESS	DULLNESS INSIPIDNESS LACKLUSTER NUMBNESS

Figure 2. DEAD Root Group

Fig. 2. DEAD Root Group

The difference in number between the words gleaned from the two entries in the Synonym Dictionary and those from the entries from the thesauri is obviously considerable. In fact, the total number of words in the lists under the Dictionary of Synonyms is 12; the total for Roget's International Thesaurus is 268, and for Roget's University Thesaurus, 2452. The outputs also varied considerably; we concluded that the Dictionary of the Synonyms gave us too little information and the University Thesaurus gave us too much, particularly since many of the words seemed at best only remotely related to the search keys. Thus, we decided to use the International Thesaurus, noting that

Our doctoral student Sam Warfel undertook the study and concluded that if the Thesaurus hierarchy were regarded as having six levels (Figure 3), in a large number of cases it is safe to assume that words which occur in the same category at any level are more closely related to each other than to words outside that category, e.g., a word which occurs in 515.3 will be more closely related to a word in 515.4 than to word in 517.2. He also noted, however, that the hierarchical structure did not always show relationships that could be shown, given the information in the Thesaurus (Figure 4).

Warfel then went on to develop an algorithm which assumed an equivalence table of such related categories. This algorithm could, for example, properly analyze the word "prevent" as non-prefixed by determining that the word "prevent" does not occur in any of the categories related to the categories associated with the unprefixed root "vent." Tested against the "control" group from my earlier work, the algorithm correctly paired 8 of the 9 pairs I had identified as correctly matched by the "brute-force" program, correctly excluded 3 which the program had included, and dealt with program pairings about which I was uncertain (good in some contexts, e.g., 17th century texts, but not in others, e.g., 20th century texts) by including 11 and excluding 14. The algorithm also dealt with cases where the identity of the prefix is in question. For example (Figure 5), the word "unideal" could be interpreted by a program as either (un)ideal or (uni)deal. The algorithm correctly paired ideal and (un)ideal and rejected deal and (uni)deal.

Warfel's study thus showed the Thesaurus to be quite a reliable guide to semantic relatedness in English. There were some problems created by the placement of words in the hierarchical tree. For example "weave" and "unweave" occur in different Classes (Space) and (Abstract Relations) and thus are not shown as connected. More recent work, both with Bryan's T-graphs for thesaural representation and exploration and, with the more illuminating representations provided by the concept lattices based upon Formal Contexts, overcomes the relational distortions produced by the tree. We have not used Formal Concept Analysis to look at the issue of automating prefixation, but it is something to think about.

In an earlier day, Walter and I talked to the group here in Darmstadt about Robert Bryan's approach to representing the Thesaurus using T-graphs (Figure 6), and I will not say any more about that particular model now. But again, it was used by our graduate students Archie Patrick, Donna Mooney (in collaboration with John Talburt), and Victor Jacuzzi to show that the Thesaurus can be used to disambiguate among word senses, by using the locations within the Thesaurus of words having more than one meaning and which therefore appear in more than one place in the Thesaurus. Like Formal Concept Analysis, the Bryan model overrides the hierarchical structure of the Thesaurus so as to show relationships scattered throughout the hierarchical tree. The lattice provided by Formal Concept Analysis makes such relationships much more evident to the human user of such analyses, than do the lists of words upon which we had earlier relied. I cannot forbear showing a couple of slides (Figures 7 and 8) concerning the word "concept," first scattered throughout the Thesaurus by the tree structure, and then as ordered by the Formal Concept Analysis lattice. The importance of the disambiguation provided by both the Bryan approach and Formal Concept Analysis cannot be over-emphasized; the challenge is to determine how to make such results effectively available to systems

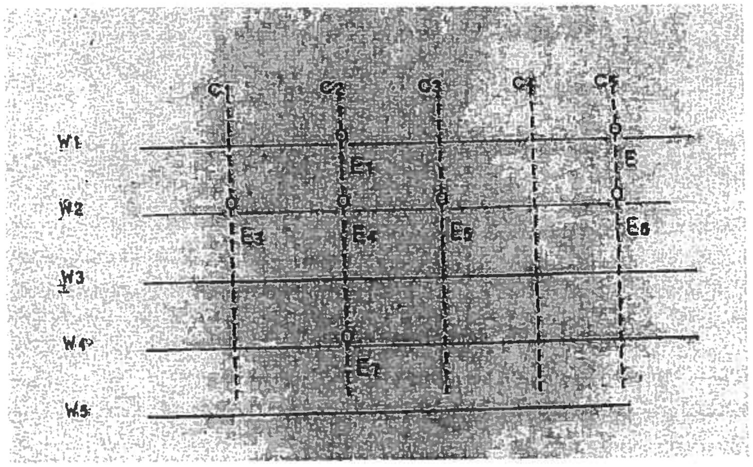


Fig. 6. Bryan's T-graphs: Entries as intersections of Words and Categories in the Thesaurus

used for large information analysis and retrieval applications (searching entire digitized libraries, for example).

I will just briefly mention other "tests" of the Thesaurus: first, a distribution of the so-called Chinese simplicia, as categorized by Karlgren, against categories in the Thesaurus showed semantic gaps conformal with observations made more 'anecdotally' by scholars comparing aspects of Chinese and English; secondly, research by John Brady and Lim Liaw using the Thesaurus to provide a conceptual overview of abstracts of articles in the 1985 SCAMC (Symposium on Computer Applications in Medical Care) Proceedings produced results which were again quite satisfactory (never perfect!); third, a distribution of the Unix Spelling Dictionary against terms occurring in the Thesaurus shows a very high correlation with the grouping of entries in the Thesaurus as to semicolon group, paragraph, category, etc. (That is terms in the dictionary "pile up" in those areas in the Thesaurus which also have large numbers of terms.) A distribution of the Oxford Advanced Learner's Dictionary against the Thesaurus also has produced a very high correlation; fourth, inasmuch as the sentence "Time flies like an arrow" is a classic in discussion of ambiguity in the English language, it is worth noting that the Thesaurus, used by the same GAME program as for the SCAMC abstracts, produces the reading that seems often to come to mind first, i.e., the speed with which times goes by; fifth, Brady again applied the GAME program to a group of text samples from a DARPA TIPSTER task (these were articles having to do with business startups and articles which might be construed by a computer program to be concerned with business startups (because of the presence of ambiguous words), but in fact were about something else altogether). Using the Thesaurus, as it is designed to do, the GAME program appropriately rejected all the misleading samples and accepted all but one of the samples deemed relevant to the topic.

As a final example of testing the Thesaurus against other data bases, I will cite John Old's study, written up in a very nice paper for the Midwest AI group in the

U.S., of three lexical networks based on the word "over." For his study, John used, first, the work of the well-known linguist, George Lakoff and his associate, Claudia Brugmán (1988), secondly the Oxford English Dictionary, and third, the Thesaurus. There is not time to go into John's methodology here, but he concluded that the central sense for Brugman and Lakoff (whose methodology is somewhat difficult to ascertain) is ABOVE+ACROSS, for the OED it is ACROSS TO, and for the Thesaurus, ADDITIONALLY. John did this work prior to our group's fortunate meeting with Professor Dr. Wille and he has subsequently produced a concept lattice for the senses of "over" in the Thesaurus. It may interest you to see Brugman and Lakoff's Radial category network (Figure 9 - the notion of "over" and "across" seems to be conveyed by "vertical" and "extended contact" above "ground;" notice that they also have the senses of "excess," "repetition" and "end" in this representation); now for a look at the representation John produced for the OED (Figure 10, note senses in the upper left-hand corner); next at the Thesaurus in John's representation (Figure 11) and finally the Concept Lattice (Figures 12 and 13); the point I am making here is that the senses in the other two networks are in the Thesaurus and certainly the Concept Lattice sets them out in an accessible way. We were particularly pleased to see that the OED senses (much richer than the Brugman-Lakoff) can be extracted from the Thesaurus.

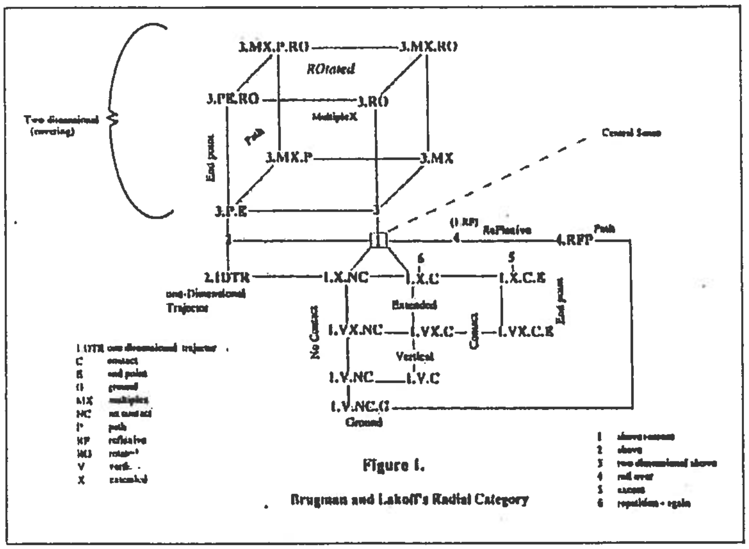


Fig. 9. Brugman and Lakoff's categories for "over" (Old, 1991)

In summary for this section of my presentation, I have indicated the extended range of the Thesaurus when used for a variety of tasks involving quite different semantic domains within the English language. For the types of retrieval/analysis tasks cited, the Thesaurus is good, albeit not perfect, and the fact that the disambiguation using the Thesaurus was completely automated is of major significance. Since the goal of Formal

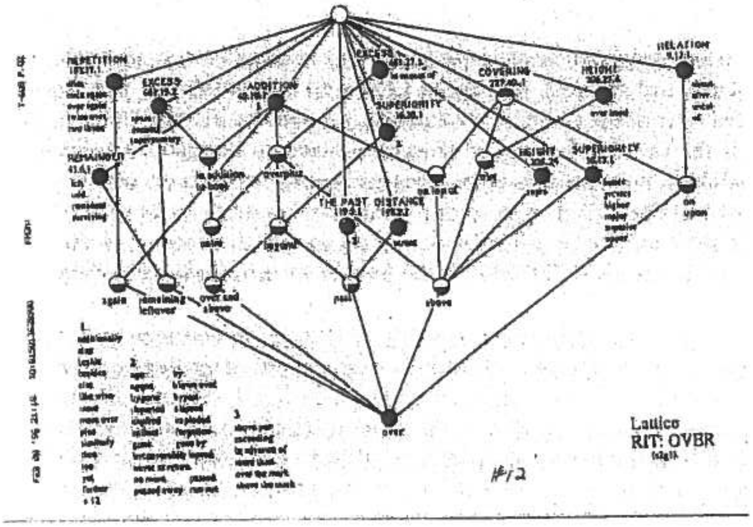


Fig. 12. Concept Lattice for "over"

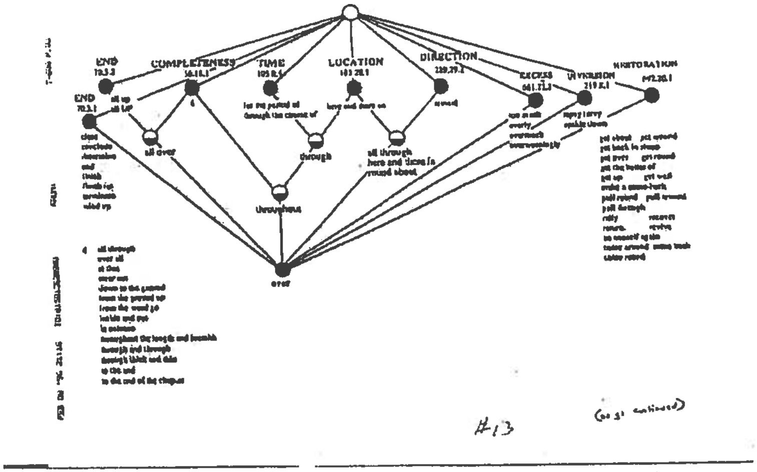


Fig. 13. Concept Lattice for "over"

In this Figure, the concepts from the concept lattice are represented as classes, and the inheritance links are represented as half circles. The classes are rounded rectangles and are divided into three sections. The first section of the class represents the names of the class; as class names, Brady used the corresponding concept names from the Concept Lattice. If the concept has an object generator, Brady includes the object name as part of the class name. The next two sections of the class represent the attributes and services of the class. The attributes contain the information kept about a class and the services are the actions a class can perform. For the Thesaurus, Brady notes that it is convenient to associate nouns with attributes and verbs with services (Figure 16).

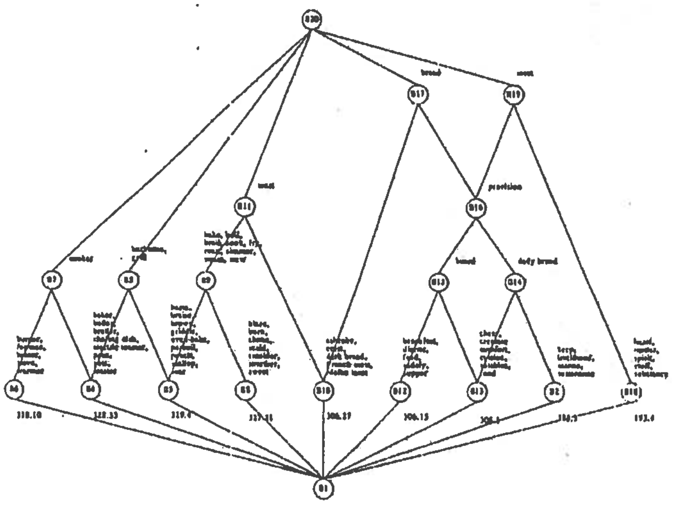


Fig. 14. Concept Lattice of 'toast', 'toaster', and 'bread'

In OOA diagrams, each service should have a defined behavior. Look at the verb "toast" in class B3 in Figure 16. Even though the service "toast" does not explicitly appear in the services section of class B3, class B3 inherits all of the attributes and services from class B11. The service "toast" is included in class B3 with the specific sense of toast as a method of cooking. Since class B3 contains only services, the attribute "toast" from class B11 will need to be overridden as an empty attribute. Along with the inheritance links from Figure 15, Brady included the behavior for the verb "toast" in Figure 16. The OOA notation for a Whole-Part link is a small triangle. The notation for a Message is a thick-lined arrow. Using the OOA methodology, the Whole-Part link may be employed to represent a "uses" relationship. In the case of the verb "toast," the Whole-Part links represent the action of toasting, using class B4 (containing "toaster") and class B16 (containing "bread"). The Message link has been used to show a constructor message sent to the B10 class (containing the noun "toast"). Similar links may be drawn for representing the behavior of the other services in class B3 and class B8.

Brady notes that although he used his own "native-speaker" understanding of the English language to determine where the Whole-Part and Message links should be applied, he believes that further analysis of the configuration and arrangement of the paragraphs within the Categories could help to automate the process of determining the Whole-Part and Message links. For example, Category 329 contains a paragraph for cooking styles, while Category 328 contains a paragraph for heating styles. The two paragraphs may be linked together by the word "cooking." So, Brady notes, while there is no explicit link between B3 - 329.4 (the verb "toast") and B4 - 328.33 (electric toaster), there are word links elsewhere in Categories 328 and 329. Within Category 329, we can informally say that a cooking method (B3) USES a cooking style. Furthermore, we can informally say that a cooking style IS-A heating style. Finally, we can informally say that a cooking device (B4) USES a heating style. Informally, this path traces a link between cooking methods (B3) and cooking devices (B4). As Brady notes, this reasoning is informal and an attempt should be made to formalize the Whole-Part links between the Categories so that automatic identification of those links can occur. It is quite possible that the excellent work Uta Priss has done with WordNet will be helpful here: either by analogy or by importation from WordNet (with the functional arrows reversed, to be true Brady's approach). At any rate if we want to deal with functionality here we have a deficiency in the Thesaurus that requires remedy.

Brady ultimately rejects OOA diagrams, as well as work by a number of other scientists, in favor of an approach by William Cook as a way to provide a more robust representation of behavior or function in the structures in the Thesaurus. Brady rejected several of the other approaches because the compatibility of behavior is imposed by the inheritance hierarchy (top-down), rather than having inheritance built from the compatibility of behavior. Cook argued for the latter approach, noting that it is necessary to build a behaviorally compatible hierarchy because "there is a growing consensus that inheritance is a 'producer's mechanism' (Meyer 1991) that has little to do with a client's use of classes (Cook, 1992, p. 1). Brady then proceeds to define a Toast conformance hierarchy in terms of procedural/functional constraints. Again he stresses that a manual process was used to identify the constraints for each of the Thesaurus' paragraphs used to build the concept lattice in Figure 14 and calls for further research to ensure that the constraints do exist in the Thesaurus and that they could be automatically recognized.

The conformance hierarchy using the Cook notations is shown in Figure 17. Brady notes that examination of the conformance hierarchy in comparison with the compatibility of behavior associated with the original concept lattice shows that several words are not compatible. Inasmuch as HeaterThing is an object and HeatProcessThing is a process, the words "toast," "grill" and "barbecue" in the original concept lattice are used both as nouns and as verbs and would be split across HeaterThing and HeatProcessThing. Since this dual usage causes problems with the compatibility of behavior, Brady labels the occurrences of the words as "toast-N," "toast-V," "grill-N," "grill-V," "barbecue-N," and "barbecue-V" depending on whether the word is used as a noun or as a verb. Brady goes on to state that a conformance hierarchy as a partial order may be used as a multivalued attribute in a Formal Context. The original context used to build the concept lattice may be supplemented with the multivalued attribute representing the partial ordering of the conformance hierarchy. He thus modified his formal context and

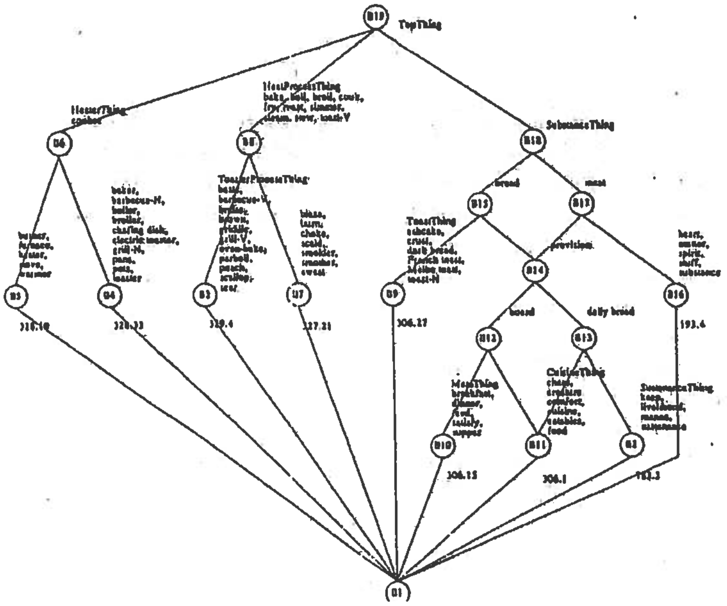


Fig. 18. Concept Lattice using Conformance Hierarchy

have liked, it is high enough to provide the basis of a semi-automatic tool. The tool could provide candidate locations along with the evidence it has compiled for each location." (One can imagine that concept lattices would considerably enhance the meaningfulness of the output for the human investigator. Also, Jacuzzi's algorithm, which produced finer discriminations than Talburt and Mooney's algorithm, might have increased the percentage of successful mappings.) Even with less than ideal results, they were able to use an integrated Roget/Longman Dictionary lexical browser for aero-space terminology which they felt performed well enough so as to provide a "rather nice demonstration of some of the functionalities that are possible with tightly coupled lexical resources. The most obvious use of such a tool would be for people who need to explore a new domain in depth; in this capacity the browser would be an aid to learning." Here is the kind of application for which Roget's 2000, incorporating Formal Concept Analysis, would be a natural.

The exciting thing about this research from our point of view (other than having other researchers use the Thesaurus in a serious way) is the mapping of a dictionary onto the Thesaurus. Even a 63% success rate will greatly enhance the scope of the Thesaurus and could presumably provide an even better structure for the mapping of the remaining 37% as well as entirely different lexicons onto the Thesaurus. So that one could anticipate being able to deal with the vocabularies in specialized domains (e.g., McHale and Crowter's work with aerospace engineering) as well as with the more general-purpose vocabulary in which domain-specific terms find their context.

Intelligence, ed. S. I. Small, G. W. Cottrell and M. K. Tanenhaus, Morgan Kaufmann, San Mateo, California, pp. 477-508.

3. Coad, Peter and Yourdon, Ed. (1990). *Object Oriented Analysis (2nd Edition).*
4. McHale, M. L. and Crowter, I. J. (1994). *Constructing a lexicon from a machine readable dictionary.* Technical report RL-TR-94-178, Rome Laboratory, Griffiss Air Force Base, New York.
5. Miller, George A. (1992) *Nouns in WordNet: A Lexical Inheritance System.* In: Five Papers on WordNet. Available at http://wordnetcode.princeton.edu/5papers.pdf
6. Old, L. John, ?? "over" MAICS
7. Old, L. John, (1991). Image Schemas and Lexicons: A Comparison Between Two Lexical Networks, Dictionary Society of North America. Conference presentation. Available at http://www.johnold.org/LJOLD/papers/DictSocPaper.pdf.
8. Talburt and Mooney