THE SEMANTIC STRUCTURE OF ROGET'S, A WHOLE-LANGUAGE THESAURUS

L. John Old

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Library and Information Science of
Indiana University
November 2003

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Charles H. Davis, Ph.D.

Debora Shaw, Ph.D.

Blaise Cronin, Ph.D.

John K. Kruschke, Ph.D.

November 24, 2003

Acknowledgements

I wish to thank Chuck Davis for his friendship, encouragement and support, without which this dissertation would truly not have been completed.

I wish to thank Ralph Shaw for being my intellectual sounding board, guide and mentor throughout the Ph D. process.

I wish to thank John Kruschke for his perseverance and tolerance over the years.

I wish to thank my academic grandparents, Walter and Sally Sedelow, for starting me on the road to research and the meaning of meaning.

I wish to thank my wife, Uta Priss, for her love and encouragement.

I wish to thank Doug Hofstadter for his encouragement, and for his active support of this research.

I wish to thank Blaise Cronin for bailing me out when the ship was going down.

ABSTRACT

L. John Old

THE SEMANTIC STRUCTURE OF ROGET'S, A WHOLE-LANGUAGE THESAURUS

This study analyzed a database version of Roget's Thesaurus (Roget's International Thesaurus, 3rd Edition, 1962) for frequency and connectivity patterns among the words, senses, and cross-references in order to identify the implicit conceptual structure. Using descriptive statistics, lattices, and information maps, semantic patterns implicit in the data, at both the local and global levels of the structure, were identified.

The explicit organizational structure of the thesaurus is, at the local level, sets of synonyms; and at the global level, a hierarchy of concepts. In contrast, the implicit organization at the local level has the characteristics of dictionary sense definitions (genus and differentiae), and at the global level has the characteristics of a small-world social network. The concept of genus and differentiae provides a model that can be seen to account for the distribution of polysemy within senses and across the Thesaurus. The small-world network model can be seen to account for the incidence of semantic hubs and authorities among cross-references, and conceptual and semantic switching centers among senses and words in the Thesaurus.

Previous work on Roget's Thesaurus calculated chains and equivalence relations algorithmically from senses and words. In that research it was found that there is an inner semantic core of most-densely-connected words and senses. This study expanded on that research identifying the semantic structure of the inner core and relating it to the top most polysemous words in Roget's.

While the largest thesaurus Categories relate to concrete objects such as plants, animals, food, clothing and technology, the most-connected words (in terms of numbers of senses and synonyms) were found to relate to abstract concepts such as motion, agitation and what appear to be concepts related to survival. This observation was supported by frequency counts, and global cross-reference and word connectivity patterns.

Table of Contents

List of Figures

List of Tables

List of Schematic Diagrams

Chapter 1: Introduction

Roget’s Thesaurus is a cultural object, a condensation of the English vocabulary, and a “culturally validated” “whole-language” thesaurus (S. Y. Sedelow, 1991,1993), as opposed to a special-topic thesaurus or controlled vocabulary. It has been described as a synonym dictionary, but it is different from alphabetically organized synonym dictionaries such as Richard Soule‘s 1871 “Dictionary of English Synonyms & Synonymous Expressions” in that these lack the hierarchical conceptual organization of Roget‘s Thesaurus.

Dr. Paul Mark Roget intended his thesaurus as a classification—a “classed catalogue of words … according to the ideas which they express” (P. M. Roget, 1852, Introduction)—into what Gerard Salton describes as “concept classes” (Salton, 1968), and as such, it has been described as a conceptual hierarchy. It compares to George Miller‘s WordNet, a model of the “mental lexicon” (Miller, G., Beckwith, Fellbaum, Gross, Miller, K., and Tengi, 1993) and Doug Lenat‘s Cyc, a network instantiation of “common-sense” knowledge (Lenat et al., 1990; Lenat, 1995).

In part due to the unavailability of a database version of Roget‘s it has not been studied in the way WordNet and CYC have been studied—as a conceptual model. There are few descriptive statistics available on Roget‘s, and there has been no study of the connectivity patterns between concepts or the “inner structure” (W. A. Sedelow, 1990, p.17) outside of word fields (sets of semantically-related words) and semantic neighborhoods (sets of semantically-related words, along with their senses).

Thesauri have been well modeled—defined mathematically and graph-theoretically (for example, Bryan, 1973 and 1974; Taylor, 1974; Kochen, 1965; and Priss, 1996)—but there is no global model that accounts for or describes Roget's conceptual interconnectivity.

Roget's has also been extensively studied as a potential, and actual, natural language processing tool (for example by Yarowsky (1992); Jones (1964); McHale & Crowter (1994); Liddy et al. (1990); Ellman & Tait (1999); Wang, Vandendorpe, & Evens (1985); and W. A. Sedelow Jr. & S. Yeates Sedelow (multiple NSF and Defense Department technical reports from 1965 to 1994)). However it has not been fully studied in the way that Roget intended it to be used—as a classification of ideas. That is what this study contributes.

Overview of the Study

The word "semantic" is derived from the Greek word "sema," meaning, "sign." On the frontispiece to the original edition Roget placed this quotation:

> It is impossible we should thoroughly understand the nature of the SIGNS, unless we first properly consider and arrange the THINGS SIGNIFIED -
> Επεα Πτεροεντα [Winged Words]

Roget interpreted "the signs" to mean words and "the things signified" to mean, as noted earlier, the ideas which the words express (rather than objects of the world—denotata).

Ideas may be broad notions; specific, labeled concepts—or senses; or “image schemas” (Brugman and Lakoff, 1988). The patterns of concepts hidden in Roget’s has implications for understanding at least the culture of English language speakers, and perhaps for understanding broader phenomena such as the history of concepts and language; and an understanding of cognition.

Exploring these patterns involves studying the relations within the hierarchy, and among the lowest level senses and words. To achieve this a database version of the 1962 edition of Roget’s International Thesaurus (RIT), 3rd Edition, edited by Lester Berry, was used primarily.

The external, explicit structure of RIT is explored to provide knowledge and context for the analysis of the hidden, or implicit structure. Figure 1 illustrates schematically one aspect of what the word “hidden” means, as used here. Green nodes represent the classes and categories of the RIT hierarchy. When browsing the body of the RIT text (the Sense Index) only the Entries—instances of words—can be seen. They are represented in Figure 1 by small black nodes.

The relationships between these Entries and Entries found elsewhere in the text are not evident, but they may be implied by the fact that the Entries are all instances of particular words (represented by the red nodes); each Entry representing one sense of one word. Nor is the relationship evident which exists between all words derived from the same

Indo-European root; or that the words which the neighbors of each Entry represent, may all share the same senses elsewhere in the thesaurus.

Nor is the fact made obvious that some words are more pervasive and connected through their relationships with other words; or have only one sense and no synonyms; or exist in a subset of the senses of another word; or have multiple parts-of-speech (can be used as both a noun and a verb, for example); or form parts of chains, clusters or partitions, when the thesaurus is viewed as a whole.

The isolines beneath the words (like weather map isobars) are included to suggest that there is a variable density in the distribution of the words, which forms a topology or landscape. The words, their senses and the relationships among them form a hidden world that will be explored in Chapter 5, and, where relevant, compared to semantic structures found in other sources of data. For example, word association data, other conceptual classifications, and etymological data.

Significance of this Study

The meaning of words, the relationship of words to concepts, and the structure of the "mind" vis-à-vis the meaning of meaning have been discussed since before Aristotle's time. This research may help clarify these topics and have a positive impact on information science for the following areas:

- better document keyword identification and indexing methods

- guidelines for thesaurus construction
- a conceptual inter-lingua for cross-cultural document classification, and automatic categorization of text and documents
- term expansion for information retrieval queries
- term disambiguation for queries and retrieval results.
- classification structures, or systems, based on conceptual patterns

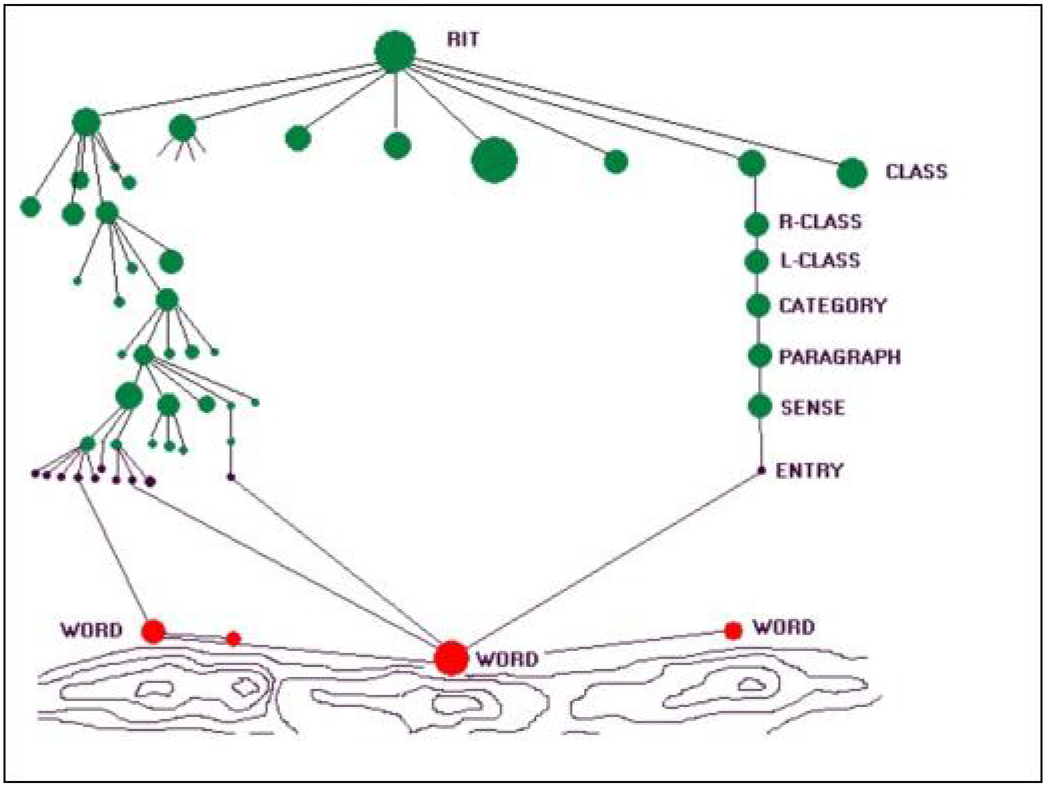


Figure 1. **Explicit structure versus implicit structure of Roget's Thesaurus**

For example without a clear understanding of the patterns of relationships between words and their meanings, term expansion would be a blind process. Assuming search terms are expanded by adding synonyms from Roget's that share senses with a source term, if a source term was found in one of the highly interconnected areas of the Thesaurus, such as TMC-69 (see Section: *Models of, and Research on, the Structure of Roget's Thesaurus*, below), the effect would be to increase the number of search terms enormously. If, in the other extreme, the term for expansion were a monosemous word (only one sense) from a monolexic sense (only one entry) in Roget's, there would be no additional terms derived. By studying the patterns of connectivity in the Thesaurus using techniques such as information cartography, it should be possible to identify ways in which the effectiveness of such applications as information retrieval may be increased.

Outside of information science, any insights into the meaning of meaning, the relationships between words, and how concepts are structured in the mind would also be significant. After more than 40 years of effort, natural language interfaces to computers are still clumsy, largely because semantics, rather than syntax, is still poorly understood. New understandings of semantic structures could lead to better models.

Cognitive science is commonly concerned with the construction of models of human thinking. Results from this study could underpin, or bring into question, existing cognitive models. Even where Roget's model is "wrong" by modern standards it can provide insight into the basis of 19th Century beliefs; and provides a model for comparison and contrast with models from the 21st Century.

Patterns in Roget's Thesaurus change over time as new words are added and the thesaurus structure is reorganized to take account of developments in the real world.

These changes reflect changes in culture, including technology, philosophy, politics, and even morality. The Thesaurus differs from a dictionary in that words are not simply added, along with their definitions, but are classified. Studying the patterns of changes in the classes, as well as the new terms, can give insights into cultural trends, and may even have predictive value—assuming they do not simply reflect the beliefs of a single editor. For example, a trend toward classifying words for "scientist" and more recently, "technologist" as derogatory terms ("egg head," "geek," "nerd") could indicate a sense of threat or helplessness among cultures adapting to rapidly changing technology, compared to more technologically stable cultures.

At the broadest level of analysis the global connectivity patterns and semantic structures that will be shown to be hidden in the thesaurus may reflect patterns found in the ancient language semantics of the hypothesized proto-Indo-European language. This has value to archaeolinguists studying the origins of language. Combined with genetic and archaeological evidence, this could aid in unraveling our common history.

Summary

Roget's Thesaurus is a classification of words by the ideas they express. Dr Roget developed the explicit or external structure, a classification hierarchy of concepts with words of similar meaning, ideas, or concepts at the lowest level. Connectivity patterns between concepts, the conceptual interconnectivity, forms the implicit, hidden or inner structure. It is a conceptual model, and to that extent, potentially a model of the mind. A global model is needed to account for patterns found in the connectivity across the whole thesaurus. This research explores and studies those patterns, comparing them, where it is relevant, to patterns found in other important sources of data that involve connectivity between words or concepts.

Chapter 2: Previous Research

This chapter reviews previous research into the semantics, structure, and connectivity of Roget's Thesaurus; other, closely related thesaurus research; and formal models of thesauri. It begins with biographical information on Dr P. M. Roget, and a history of his Thesaurus as it relates to the semantic content and explicit thesaurus structure.

Roget and his Thesaurus

Peter Mark Roget was born on January 18, 1779, on Threadneedle Street, London, where his father, a Belgian immigrant of Swiss Huguenot extraction, had "oversight" of the French Protestant Church located there (RIT, Biography of Paul Mark Roget). After his father died his mother moved the family to Edinburgh where at fourteen Roget attended medical school and, at age nineteen, graduated as a medical doctor from Edinburgh University. Dr. Roget's practice included periods at the Manchester Infirmary where he helped establish the Manchester Medical School; the Northern Dispensary, which he also helped establish and where he treated patients free for eighteen years; the post of Fullerian Professor of Physiology at the London Institute; an appointment as Examiner of Physiology in the University of London; and ultimately, in 1831, as an elected Fellow of the Royal College of Physicians.

He was made a Fellow of the Royal Society in 1815, and served as secretary of the organization until he retired from the position in 1848. He was also a Fellow of the Geological Society, Member of the Senate of the University of London, and Member of many Literary and Philosophical Societies. He published several treatises, mostly on physiology, but some on electricity, galvanism, magnetism and electromagnetism; wrote in English, French, German, and Latin; founded the Society for the Diffusion of Knowledge; devised the log-log slide rule; designed a pocket chessboard (Dutch, 1962, xviii); and is even credited with inventing cinema:

> …in 1825, came his paper "Explanation of an Optical Deception in the Appearance of the Spokes of a Wheel Seen Through Vertical Apertures," which is regarded as seminal by modern historians of the cinema. (Winchester, 2001, p.2)

Roget also set chess problems for the Illustrated London News; contributed sections totaling 300,000 words to the seventh edition of the *Encyclopaedia Britannica*; and led the commission that studied London's water supply, "recommending the idea of sand filtration - a method that is in use to this day" (Sabbage, 2001, para. 10).

In 1852 he published his "Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition," and spent the next 27 years revising and adding to it. He died in West Malvern, on September 12, 1869, at the age of ninety.

Roget first "projected" his "system of verbal classification" in 1803 (Roget, 1852, p. iii), when he was 24 years old, and had by 1805 completed a "classed catalogue of words" (ibid) in the same principle and in almost the same form as in the First Edition. However it was only after his retirement from his duties as the Secretary of the Royal Society, in the late1840's, that he was able to return to, and give full attention to, this work. It took him three or four years to expand it from a "small scale" (ibid) to its published form.

Roget's goal was to classify ideas so that, by looking up an idea, one could find words to describe it; in contrast to the method used in a dictionary where one looks up a word to find an idea (its *signification*, or what the word means). According to Simon Winchester, author of *The Professor and the Madman*, a history of the *Oxford English Dictionary*, the first book to do this was not Roget's Thesaurus but one by John Trusler. Its portmanteau title was *The Difference Between Words Esteemed Synonymous in the English Language; and the Proper Choice of Them*. It was a book that:

> would lead the inquirer to a particular word if he knew roughly what it was that he wanted to say but had no firm idea of the assemblage of letters and syllables that would enable him to say it. (Winchester, 2001. p. 1)

Trusler's goal was in fact to help people choose words that were more eloquent than their day-to-day language, rather than to find a precise word to describe a particular concept—and it required the use of an initial word-lookup.

There was an even earlier, similar book, not mentioned by Winchester, by Abbé Gabriel Girard, published in 1762, called *A New Guide To Eloquence: Being A Treatise Of The Proper Distinctions To Be Observed Between Words Reckoned Synonymous*.

Trusler's book was followed in 1794 by "British Synonymy; or, an Attempt at Regulating the Choice of Words in Familiar Conversation," by Hester Lynch Piozzi, a friend of the lexicographer Samuel Johnson. It was another attempt at helping people choose the right word for the occasion, rather than the right word for the idea. There were many other "synonymies" (Winchester, 2001, p.1) published before Roget's First Edition, and even some after that. For example, Robert Soule's *Dictionary of English Synonyms & Synonymous Expressions Designed as a Guide to Apt and Varied Diction* appeared in 1871. But all were developed with the same end in mind, and none was designed or expected to facilitate the expression of ideas, as was the explicit design goal of Roget's Thesaurus.

Several of the early synonymies, because they were concerned with word usage, included meta-information about the context in which each word would most appropriately be used. This was necessary because not all synonyms are equal. Roget was not concerned about usage (that was left up to the reader) but with classifying ideas. Consequently many of the "synonyms" in his thesaurus are, apparently, no such thing.

Synonymy

An understanding of the concept of synonymy and closely related concepts (discussed further, below) is important to this study because it is one of the main relations in Roget's Thesaurus. It is also important in order to understand why Roget's Thesaurus is more than just a synonym dictionary.

English differs from other languages in its degree or number of synonymous words—it has many more (Williams, 1975). English, at its heart, is a Germanic, or more specifically, Anglo-Saxon language, but has accumulated words from many sources; mainly Latin, Greek, Danish, and French. This has been accomplished indirectly by assimilation, but mostly through conquest and occupation. Consequently the English language has usually two and often three words to describe common concepts. The Anglo-Saxon word is usually the common term, while the French or Latinate term(s) tend to be used in more formal situations. For example, *kin, relatives, consanguinean*; *wise, prudent, sagacious*; *share, allot, apportion*. That is not to say that the default, most frequently used, or common words are all Anglo-Saxon. "Plain" comes from Latin and "simple" comes from French.

The lay meaning, or at least, expectation of synonymy is that two words are interchangeable in all contexts. But words may be called synonyms if they share only one sense. After analyzing synonymy in RIT, Old (1996, p 1.) observed that:

> Two words which share a sense are commonly called synonyms. They are equivalent in the context of that sense, and can be used interchangeably in a sentence which utilizes that sense. Two words are rarely synonyms for each other for all of their senses.

An example of two common synonyms is "over" and "above." No native English speaker would question whether they are synonyms. Yet "over the hill" (as in "I live there") is not the same as "above the hill." *Over* and *above* occur together in a Roget's sense or *Synset* ("a set of synonyms," [Miller, G., Beckwith, Fellbaum, Gross, Miller, K., & Tengi, 1993]) that includes the word *overhead,* but not in a Synset that includes the spatial sense of *past* and *beyond*. Old went on to define "word equivalents" which are, in all contexts, interchangeable. Word equivalents always occur together in any of their respective Synsets, and never occur independently. In other words, they may describe several senses, and always the same senses, so are a type of "perfect synonym." Old noted that word equivalents (and also synonyms necessarily, because word equivalents are also synonyms) may be classified by type. For example, *abbreviation* {amidst, amid, midst, ‘mid; F.B.I., Federal Bureau of Investigation}; and *spelling variants* {airplane, aeroplane; Führer, Fuehrer; Odin, Woden}. Some word equivalents could be called “regular” synonyms (words which are not tied by any obvious mechanism such as form). For example, {absurdly, ridiculously; accountant, bookkeeper}.

The case in which Roget’s classificatory system produced Synsets that involved apparently no synonyms at all, is “lists.” Lists of such things as occupations, parts (such as parts of sailing ships), and other such objects are common in the thesaurus. Entries in lists may not satisfy a purist’s definition of being synonyms but they do satisfy Roget’s goal of classifying words according to their ideas, or concepts. This is consistent with the philosophy of the English philologist John Tooke. Roget was an admirer of Tooke, whose book on language, *Epea Pteroenta* (Winged Words)*,* Roget quotes several times in the Introduction to the First Edition. Tooke (1786; as cited by Roget, 1852, p. i) believed that we cannot “thoroughly understand the nature of the SIGNS, unless we first properly consider and arrange the THINGS SIGNIFIED.” There is no doubt that Roget interpreted the “THINGS SIGNIFIED” as ideas or concepts, rather than actual objects of the world.

Roget’s classificatory plan (based on the relationship between words and concepts) was similar to that of Ferdinand de Saussure, the founder of structuralism. Saussure believed that words do not unite a thing and its name, but an *image acoustique* (*significant,* or signifier) and a concept (*signifié,* or thing signified) (Saussure, 1916, pp. 98-99). Saussure also believed that the relation between words in a language is an “associative relation” (sometimes referred to, by others, as a paradigmatic relation), and that the meaning of an individual word is determined by its relation to other words in the way that the role of a chess piece on a chessboard is determined by its relation to the other chess pieces. In this scheme, Roget's word entries are an associative relation, rather than "synonymy."

In RIT, the lists are mainly arranged under Paragraphs, with each element in the list being an entry in a separate Synset. Lists may also be a series of Synsets where the entries in the Synsets are synonyms, while the Synsets themselves are ordered and bound by the "topic" of the list. Lists, like synonyms, also have a range of types. For example, (all from Categories in the Roman numeral level Class: Religion) lists of parts ("parts of the Mass"); lists of activities ("rites"); lists of objects, or things ("Holy days"); lists of classifications (from Category 61: Classification; "botanical and zoological classifications"); and lists of sequences (from Category 107: Irrelation; "Geological Ages"). About 10% of the RIT entries (approximately 19,000) are in lists. Note that elements in lists are used in the same contexts, even although they are rarely interchangeable in those contexts; and that they still share semantic content—that of the notion (or topic) described by the list's label.

Roget also accounts for "paronymous" words, which are "different parts-of-speech from the same root, which exactly correspond in point of meaning[1]" (Roget, 1853, xix). *Para-*

[1] Roget credits the definition to "*A Selection of English Synonyms,* edited by Archbishop Whately." A paronym is also called a "conjugate," which is defined in Webster's (1976) as "having the same derivation and therefore usually some likeness in meaning." On the other hand Webster's defines "paronymous" as "2 a : formed from a word in another language b : having a form similar to that of a cognate foreign word." The common concepts here are "form" and "derivation," which imply a common root. From this perspective these definitions are consistent with the definition that Roget chose.

means “near” and *–onyma* means “name” (or, the special name of a thing). “Confronting” and “in front of” are examples from RIT. The English adjective “stark,” which has the meaning “rigid,” and the archaic meaning “strong;” and the related (through a common root) German adjective “stark” (not found in RIT) which has the meaning “strong, powerful;” are also paronyms. Again, this is not a synonymy relation at work.

Winchester (2002) categorizes Roget’s Thesaurus amongst the other “vocabulary expanders”— synonymies for the verbally insecure—so in fact mistakes Roget’s classification by ideas for being an error, or shortcoming:

> …the central shortcoming in Roget’s Thesaurus, as I see it, stems not from the book’s troublesome structure but from something quite different—from Peter Mark Roget’s Panglossian regard for the intellectual merit of his likely readership. (p. 3)

Roget never intended his work to be a book of synonyms. He realized that all words, not just synonyms, are known by the company they keep. Furthermore, he realized that opposites, or antonyms, were also closely related and should be categorized nearby to each other in his scheme. He had some difficulty convincing his publisher that this difficult-to-typeset layout was a sensible thing to do, but eventually got his way; and all early editions of the thesaurus were organized into (where relevant) synonym-antonym opposed categories, until his system was corrupted by later editors who felt the costs outweighed the benefits.

Antonymy

It is not obvious that antonyms have more in common than other, arbitrary pairs of words, but under examination it becomes obvious that opposites are poles on a common dimension. For example, hot and cold share the dimension (or attribute) of temperature. They also exist on a continuum, or range, as do many other antonyms—temperature is graded. Words between hot and cold on this continuum are {cold, cool, tepid, luke-warm, warm, hot}, and of course, can be modified by qualifiers such as very (hot) and quite (cold). So hot and cold have much more in common than, say, hot and green.

Roget's organization of Categories took account of antonyms and classified them under super-Categories, or *hypernyms*. A hypernym is a broader or more general term that in turn describes or represents a notion shared by, or implicit in, both of the lower-level Categories. In relation to the hypernym, the thesaurus Categories are *hyponyms*—representing specializations of the notion described by the hypernym.

Many antonyms consist of a positive, or base sense, and a negated form of the same word. For example, "logical" and "illogical." These are most often formed using prefixes that are of Latin and Greek origin (*in-, im-, mis-, dis-*, and so on; but also *–less*, as in *brainless*). "Not" may be used, but (formal) logic tells us that the negation of something is "everything else," not the opposite of it. For example "not hot" does not mean cold, although context usually constrains it from meaning "everything else." Negation in English is a complex topic and will not be fully addressed here, but note that the Latin and Greek prefixes used to generate opposed concepts usually do not mean "not." {*a-, in-,* *im-, il-, non-*} in fact may; but *an-* (without), *anti-* (against), *de-* (separate), *dis-* (apart), *dys-* (difficulty), *re-* (backward), *un-* (front, before[2]) each use a different mechanism to differentiate the opposed form from the basic form of the word.

It is clear that, like synonymy, antonymy is much more complex than it first appears. It is also pervasive. From real-world concepts like “up” and “down” to abstractions like “superior” and “inferior,” much of language semantics is composed of duals and symmetries. Science and mathematics use the concepts of duality and symmetry from the operations of addition and subtraction, to super-symmetry theory (Ito, 1993, Index[3]). Even balance (as in dynamic equilibrium and homeostasis) involves a duality. We may be predisposed to classify meaning into dualities by our senses and perceptions, which utilize similarity and difference, contrast and equivalence, to recognize and order the world. Also, the need to integrate orthogonal dimensions such as “what” and “where” requires that we often pay attention to two or more things at a time.

George Miller, while developing WordNet, his model of the mental lexicon, recognized that antonyms are closely related[4].

> The strongest psycholinguistic indication that two words are antonyms is that each is given on a word association test as the most common response to the other. (Miller et al., 1993, p. 24)

[2] Un- is derived from the Indo-European root, ANT-, which meant front (as, perhaps, in “confront”) or forehead, though it has come, through usage, to mean “not.” Also un- is Germanic, not Latinate or Greek. It is also worth noting that il-, im-, ir- (as in, “irregular”) are all derived from the related Latin prefix, in-.

[3] The Encyclopedic Dictionary of Mathematics has 39 major entries for *Dual-* and 36 for *Symmetry-*).

[4] He included antonymy as a relation in the WordNet structure, and it is the only relation that is applied to all four main parts-of-speech (nouns, verbs, adverbs, adjectives).

So, although one might expect words in opposition to be poorly associated with each other, they are in fact very strongly associated. Miller et al. also cite an unpublished study by Fellbaum and Chaffin (1990), who found that:

> subjects were most successful in completing analogies that involved an opposition relation. Moreover, analogies based on opposition relations took the least time to complete. (C. Fellbaum, personal communication, June 17, 2002)

Analogies and metaphors, like antonymy, are relations between things (such as objects, processes, or situations) that have some features or objects in common, and some features or objects that are different (Old & Priss, 2001). This type of "mapping" may be fundamental to the organization of human cognition. Metaphor and analogy are discussed further, below.

Miller also recognized that antonyms were often graded, and organized all of the WordNet adjectives as a type of opposed category relation. He observed that words like "large" and "small" are antonyms, but while "big" and "little" are also antonyms, "large" and "little," are not so, in quite the same way—despite the fact that "large" and "big" are synonyms, and "small" and "little" are synonyms. He argues that this is because (with the exception of "a handful of frequently used adjectives (most of which are Anglo-Saxon),") antonyms are commonly formed by a morphological rule—that of adding a negative prefix "that changes the polarity of the meaning [of the word]." So "antonymy is not a relation between meanings" (Miller et al., 1993, p. 28) but a relation between words.

Miller et al. (1993, p. 28) also observed that, at least in adjectives, there is usually a strong antonymic relationship between a primary pair of words, or "direct antonyms," so that their opposed Synsets should be regarded as "clusters of adjectives associated by semantic similarity to a focal adjective that relates the cluster to a contrasting cluster at the opposite pole of the attribute." Miller et al. called the relationship between these synonyms of the primary pair "indirect antonymy." For more on antonymy, especially as regards "markedness" and gradation, see Miller et al. (1993).

Roget recognized that antonyms were not simple opposites or negations, in the way that he recognized that synonymy was, in the American idiom, "a can of worms:"

> The investigation of the distinctions to be drawn between words apparently synonymous forms a separate branch of inquiry, which I have not presumed here to enter upon; for the subject has already occupied the attention of much abler critics than myself, and its complete exhaustion would require the devotion of a whole life…. The purpose of this work, it must be borne in mind, is, not to explain the signification of words, but simply to classify and arrange them according to the sense in which they are now used, and which I presume to be already known to the reader. I enter into no inquiry into the changes of meaning they may have undergone in the course of time. I am content to accept them at the value of their present currency, and have no concern with their etymologies, or with the history of their transformations; far less do I venture to thread the mazes of the vast labyrinth into which I should be led by any attempt at a general discrimination of synonyms. The difficulties I have had to contend with have already been sufficiently great, without this addition to my labors. (Roget, 1852, pp. xvi-xvii)

He made a similar observation of antonyms (although he never actually uses the word "antonym" in any of his writing—he uses the terms "correlative" and "opposed" instead) while discussing the vagaries of the senses of words, as described in dictionaries:

> It may even happen that the very same word has two significations quite opposite to one another. This is the case with the verb *to cleave,* which means *to adhere tenaciously,* and also *to separate by a blow. To propugn* sometimes expresses *to attack;* at other times *to defend. To let* is to hinder, as well as *to permit. To ravel* means both *to entangle* and *to disentangle. Shameful* and *shameless* are nearly synonymous. *Priceless* may either mean *invaluable* or of *no value. Nervous* is used sometimes for *strong,* at other times for *weak.* (Ibid, p. xvii)

Although his examples are often archaic terms, the principle still holds, further endorsing the view that antonyms are semantically closely related to each other.

Roget also observed what Miller et al. (1993) later reported: that “two ideas which are completely opposed to each other, admit of an intermediate or neutral idea” (Roget, 1852, p. xiii); that is, that antonyms are often graded. Furthermore, he observed that many words or ideas have multiple opposed words or ideas. He gives many examples, of which “give,” versus “take” or “receive,” is illustrative. An example from his Categories is Category 677: Use, versus Category 678: Disuse (opposed Categories in the thesaurus structure); immediately followed by another potential antonym of *use*, Category 679: Misuse. He also observes that, “contrast in form does not always bear the same contrast in meaning” as in, for example “malefactor” versus “benefactor;” and difference versus indifference. Finally he makes the observation that “there seldom exists an exact opposition between two words that may at first sight appear to be the counterparts of one another.[5]“ “Untruth” is not the exact antonym of “truth” because it has the connotation of lying, which draws in other concepts such as morality and deceit. Except in the context of “telling the truth,” where the alternative concept is implied, “truth” is a neutral term.

It appears that the common problem linking the issues discussed above is one of context and usage. “Malefactor” and “untruth” are used to represent concepts that have negative social connotations and through the common context of their use those connotations have become a part of each word’s semantics. They are no longer simply the other side of the

[5] This is the same conclusion drawn in earlier discussions about synonyms. Perhaps, rather than synonymy and antonymy as relations, associative and correlative would be better terms for relations of this type.

semantic coins of “benefactor” and “truth.” Usage is what determines the meaning, or senses, of any word, and the context in which it is used is what disambiguates one sense of the word from the other senses of the word. It would be surprising if it were any different for antonyms.

That is not to say that antonymy, or opposed concepts, or negation do not exist—just that they exist in a much more complex, organic way than simple classification or formalization can represent. Roget, although overwhelmed, was not deterred; 75% (780) of the Categories in the 1852 edition were arranged in opposition. A further 35 Categories form triples of the Misuse versus Use-Disuse type—Categories not selected to participate in antonymous pairs but that are closely related to one, or both, of an antonymous pair. Other examples are Convexity-Flatness-Concavity, and Teacher-Learner-Learning. These triple-forming Categories are not made explicit in Roget’s original structure, although they do all occur as consecutive Categories in the text.

Of the remaining 200 or so Categories, about half could be found partners by simply negating the terms used to label the existing Categories—although the new Categories would be small, and sparsely-populated. The other half is comprised of semantic, physical, or numerical “fixed points.”

A fixed point, as used here, is a concept that does not lend itself to being contrasted with anything else—although it may be ordered (set in order with other concepts) or subdivided into parts. Examples are Category 75: Class; Category 86: List; Category 29:

Mean (average); Category 84: Number; Category 105: Infinity; Category 233: Limit; and numerous categories labeled for concrete nouns such as World, Mankind, Stock Market, and Music (but not Melody, which is contrasted with Discord).

Categorical notions are all represented by nouns in Roget's hierarchy. If adjectives or verbs were admitted then vaguely antonymic terms could be imagined for the solo, un-paired, fixed-point Categories. Examples generated from the Categories listed in the previous paragraph might be *classless, unlisted, innumerate, finite, unworldly, inhuman*, and so on. But these, as Categories, would not be of the broad notions type to which Roget managed to assign, on average, 95 semantically related words.

As mentioned earlier, duality, symmetry, and many other forms of opposed, contrasting, or polarized "things" exist in the natural, mathematical, and scientific (not necessarily mutually exclusive) worlds. That is not an idiosyncrasy of the English language. Zen Buddhism and the dual philosophical concepts of "yin" and "yang" (positive-negative; female-male)[6] suggest that the phenomenon of duality is not an artifact of Indo-European-derived culture, Western thought, or Roget's beliefs. In the Chinese database used for this study, the most common word element (singly, or as a part of compound words, or in phrases) is bu4 ("no," "not")[7]. Corpus-based frequency statistics also show that bu4 is pervasive. After de5 (of; at; on), yi4 (one), and shi4 (to be; is; "yes"), bu4 is

[6] Strictly speaking yin1 means "shady side" and yang2 means "sunny side," but their senses have been extended to represent a wide range of "antonymous" philosophical ideas. "Yin: The female principle; feminine, negative, hidden, dark, softness, earth, internal, small as opposed to *yang*: positive, masculine, male, obvious, [sun-] light, hardness, etc" (Muller, 1997, [Radical 170]). "Yin and Yang, the two opposing principles in nature, the former feminine and negative, the latter masculine and positive" (Lanbridge's Concise Chinese-English Dictionary, 1985, p. 1422).

[7] Pinyin is only the pronunciation. The character or glyph may have several forms. Bu4 has several meanings, but the main meaning is "no," or "not."

fourth in frequency. “Bu4 zu2” means insufficient, while “zu2” means sufficient; “bu4 zhun3” means forbid, while “zhun3” means permit. Bu4 acts for Mandarin Chinese as in-, ill-, or un-, do for English (D. Li, & W. Wang, personal communication, June 2002).

Not all negated words in Mandarin Chinese are formed by using bu4. Two other equivalents of un- are fei1 and wu2; mei2 is the negative prefix for verbs; and there are other words that translate as “not,” such as fu2, mi3, po3, wei4 (and wu2, mentioned earlier). For anti-, the prefix used to indicate opposition, fan3, dui4 or de3 kang4, are used (D. Li, & W. Wang, personal communication, June 2002). And, as in English, the most common Chinese antonyms (*hot, cold*; *many, few*; and so on) exist as distinct words, with no negation or affixes.

Mandarin is in a language family unrelated to Indo-European languages. Yet the same mechanisms of antonymy, identified and discussed by Roget in when he built his thesaurus, are applicable to Mandarin as well.

Roget was well justified in organizing his thesaurus taking account of the polarities found between antonymous concepts—despite the fact that his publisher opposed it and subsequent editors simply eliminated all evidence of it—, with one exception. The recently published 150th Anniversary Edition of the British branch of Roget’s Thesaurus has restored the “opposed categories” organization.

Universal Conceptual Structures

Roget believed that his classification system, because it was based on the organization of concepts rather than words, could be used to align different languages for comparison or learning.

> …the principles of its construction are universally applicable to all languages, whether living or dead. On the same plan of classification there might be formed a French, a German, a Latin, or a Greek Thesaurus, possessing, in their respective spheres, the same advantages as those of the English model. (Roget, 1852, p. xxii)

According to the Mawsom (1922, p.35) such later works, capable of being used for comparison with English, did appear. For example, *Dictionnaire Idéologique* by T. Robertson (Paris, 1859); *Deutscher Sprachschatz* by D. Sanders (Hamburg, 1878); *Deutscher Wortschatz, oder der passende Ausdruck* by A. Schelling (Stuttgart, 1892), and more recently the German thesaurus, *Der Deutsche Wortschatz nach Sachgruppen*, by Franz Dornseiff (1970; first published in 1933). But no attempts to implement Roget's dream have yet been successful. No matter how well concepts (or words) are aligned, there is always some slippage of the semantics (Old, 1995). This is because the connotation of a word or idea—the penumbra of associations and implications, as opposed to the denotation (the thing itself)—varies from language to language, even from culture to culture. It is also because idiomatic and metaphoric usage, pop culture or political correctness, can hijack the meaning of a word, or blend it with others, and so bring new connotations to existing concepts producing misalignment between the semantics of languages.

Even two English editions of Roget's can be aligned only to about eighty percent. This is in part due to editorial judgments on how the conceptual structure should be organized. Even words and concepts from so-called "closed sets," such as prepositions (discussed further, below), vary greatly between closely related languages where the corresponding words are cognates (have the same ancestral/etymological root). Here again, context and usage are paramount[8] in determining meaning.

And a final note on Winchester's criticism of the organization of Roget's Thesaurus. Winchester, among all his other errors in evaluating Roget's intentions and work, also demonstrates his misunderstanding of the Thesaurus in his criticism of the size of the index:

> Today the index to the British edition is twenty pages longer than the thesaurus itself. The index to Roget's International Thesaurus, in America, though set in a typeface two points smaller than that of the main body of the book, still occupies half the number of pages the thesaurus does; it would be far longer were it printed in the same size. (Winchester, 2001, p. 3)

What Winchester did not observe, or chose to ignore, about the Word Index was that it is arranged as a list of head words, with sub-lists by part-of-speech, and sub-listed senses; each sense with a differentiating word, and a Category and Paragraph index number. It takes up more space in part because of the formatting and page-space overhead required of organizing these differentiators in the manner of a standard alphabetic dictionary; and in part because the complexity of language semantics is such that providing alternative access points to the classified concepts takes more space than simply grouping them. Winchester's criticism does not detract from Roget's organizational scheme, but endorses

[8] There is a related, strong movement to generate multilingual versions of WordNet. EuroWordNet includes Dutch, Italian, Spanish, German, French, Czech. and Estonian, linked to an "Inter-Lingual-Index."

it as a succinct, efficient and intuitive method of organizing concepts (and words) of the English language.

It should be noted that Winchester's criticism might have been even harsher if he had only known that the index is incomplete. Many words are omitted, including archaic and obsolete words; foreign terms; Latin quotations; and many other such entries that the average thesaurus-user would never be expected to use as entry points when searching for a concept, and so are not included by the editors.

Models of, and Research on, the Structure of Roget's Thesaurus

Research on Roget's Thesaurus over the last thirty years has included the modeling of connectivity within Roget's (S.Y. Sedelow, 1991, 1993; S.Y. Sedelow & W. A. Sedelow, Jr., 1969, 1986a, 1986b, 1992, 1994a, 1994b; W. A. Sedelow, Jr., 1985, 1993; W. A. Sedelow, Jr. & S.Y. Sedelow, 1979a, 1979b, 1983; L.J. Old, 1991a, 1991b, 1993, 1996b, 1999, 2000, 2002, 2003). As part of this line of research a mathematical model of abstract thesauri, of which Roget's Thesaurus is one instantiation, was developed by Robert Bryan (1973; 1974). The model includes the definition of chains derived from T-graphs. T-graphs are similar to cross-tables in that a relation is defined on two sets and represented by a table of columns and rows, where a cross or a dot represents a relationship between an element on a row and an element in a column. The dots (or crosses) are referred to as *entries*. The elements in this model are word-strings—which may be single words, compound words, or phrases—and senses (sense definitions, or synonym sets), and a relation between them.

Bryan defined a series of chains linking the entries, either along the rows (word associations), along the columns (senses associations), or both. The most restrictive, the Type-10 chain, is a double chain, which requires at least two words to share a sense or two senses to share a word, in order to participate in the chain. This ensures links are not arbitrary, as happens when two senses are linked by homographs (identical spellings but with disjoint meanings) such as *lead* (the metal) and *lead* (to command).

If links are allowed freely between senses with shared words or words with shared senses, an almost fully connected node-arc (vertex-link) graph, or network, of semantically linked words and senses, is derivable from RIT. The more restrictive Type-10 chains partition the thesaurus into (disjoint) equivalence relations, as they have the properties of being transitive, reflexive, and symmetric. John Talburt and Donna Mooney (1990a, 1990b) processed the electronic version of *Roget's International Thesaurus*, 3rd Edition (RIT) to derive all of the nodes of all possible Type-10 chains. The result was 5,960 disjoint parts (partitions, or equivalence relations) referred to as *components*. Talburt and Mooney assigned unique numbers to each component and counted the number of links in each. As each word must have at least two senses in common and each sense must have at least two words in common, the minimum link is actually a set of four entries (two words and two senses). Talburt and Mooney named these links of four entries "quartets."

Components derived from RIT ranged in size from the largest component, TMC-69 (Talburt-Mooney component, index number 69), of more than 22,242 linked entries; to more than three thousand components consisting of only four entries (single-quartet, single-node chains). The existence of TMC-69 despite the tight constraints imposed by Type-10 links illustrates the massive interconnectivity within English semantics and between English words. This network ties together such related words as *streak* and *stripe,* but also such disparate words as *grain* and *fashion*.

Jacuzzi (1991) observed that quartets forming chains could link face to face (forming a structure like the pattern of a large crossword puzzle), and (with Old) by "corner-joins" (see Diagram 1). In other words, given three senses, *s1, s2*, and *s3*, with *s1* having only two words (*w1, w2*; both synonyms) and *s3* having two words (*w2, w3*) and *s3* having three words (*w1, w2, w3*), *s1* could join to *s2* by two words (*w1* and *w2*) and *s2* to *s3* by two words (*w2, w3*).

By reprocessing the RIT data and further constraining quartets to only "face joins," Jacuzzi partitioned the data into 10,341 smaller components[9]—almost twice as many as Talburt and Mooney. These ranged from the largest component, VJC-184 (Vic Jacuzzi component, index number 184), comprised of 1,490 entries, to 6,574 single-node, single-quartet chains. The largest TMC-69 component was broken down by this process into 2,507 smaller components, of which VJC-184 was the largest.

While TMC-69 was a loosely inter-connected network of word and sense associations, VJC-184 is a very densely bound network—a core of the core connectivities of the largest Talburt-Mooney component. This central, core "semantic network" has been the subject of research by Old (2000).

[9] Strictly speaking, these are not partitions. It was later observed that, in order to split a corner-join, the entry at that node must appear in both of the resulting "components,"

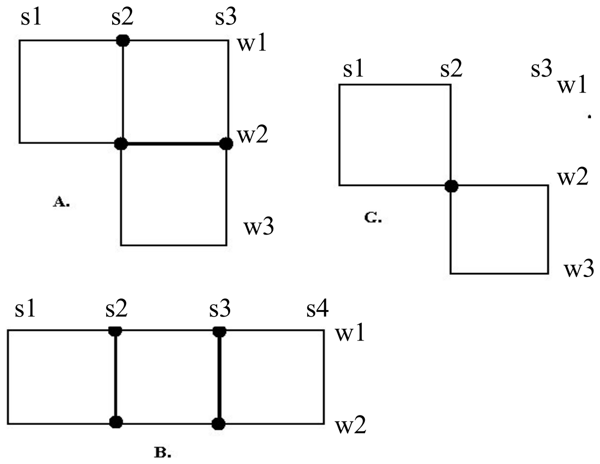


Diagram 1. **Face joins (A, B), and corner-join (C), of Type-10 chains**

The terms most likely to provide connectivity, and therefore constitute the cores of components, are polysemous (having many senses), common words. For example, the word *over* has twenty-two senses in RIT. It is an obvious synonym of *above*, which shares seven of *over's* senses. But *above* cannot be substituted for the word *beyond* (a synonym for *over* in one of *over*'s other senses) in the following sentence: "we live just beyond the next hill." So even *over* and *above,* which are indeed typical synonyms, have senses that are not held in common. One consequence of this is that the senses and words which make up the networks of tightly-knit components in Roget's are related in ways that cause transitions from concept to closely-related concept, that slowly veer, twist, blend, and travel away from whatever concept is chosen as a starting point. In this example a chain might follow the following path: *above* => *over* => *past* => *beyond* =>

exceeding => *extra* => *leftover* => *remaining.* Previous analysis has shown that such chains can also rapidly turn to antonymic senses.[10]

The words least likely to provide semantic connectivity are highly specialized technical or scientific terms (typically concrete nouns), long quotations, idiomatic phrases, and foreign terms (although even some of these can be found linking concept to concept). Archaic terms and senses may still participate in chains through their association with more modern terms and senses. There were many disconnected words (insufficiently connected to other words, or existing in lists of parts, or specialized topics) that remained after the processing of the Type-10 components. Many of these words can still be incorporated as direct synonyms into existing Quartet entries without danger of muddying the homograph partitions. But even so, 20,000 words cannot be included because they are isolated words, or monolexic (single-entry) Synsets. These are called here "singletons."

Inter-Connectivity and Connectedness in Roget's Thesaurus

The majority of the work on the structure of Roget's Thesaurus, by far, has been done by the couple Walter A. Sedelow Jr. and Sally Yeates Sedelow. Their work covers more than three decades and 140 papers. It extends from early research funded by the American defense department, on translating Soviet military strategy, to the computational analysis of Shakespeare, to a proposal for an interlingual communication support system; and follows from their early work on *Science and the Language of History* (W. A. Sedelow, Jr., 1957), *The Narrative Method of Paradise Lost* (S. Y. Sedelow, 1961), and *The History of Science as Discourse* (W. A. Sedelow, Jr. & S. Y. Sedelow, 1979a). The

[10] Cross-references, which are shadowy, incomplete, but explicit reflections of the implicit connectivity identified by chains, also show this same behavior.

common thread here is the analysis of common threads—the inter-connectedness, interdependencies, and hidden patterns within complex (mostly textual) entities.

Most of their work has aimed at the application of computers to natural language, computational discourse analysis, and models of whole-language semantic space (as opposed to, for example, topological models of individual word senses). They demonstrated the feasibility (and desirability) of using computational techniques "when [one is] studying the associative structures which carry the meaning-load for entire languages" (W. A. Sedelow, Jr., 1991, p. 88).

The insights and principles of connectivity and connectedness within complex systems, which they developed, they applied to research on the intellectual dependencies within colleges and universities. They recommended a graph-theoretic (network) approach to representing relationships, interconnectivities and interdependencies among reading assignments, courses, faculty, departments, majors, etc., from the administrative level down to the semantic space of the textbooks used.[11] A prototype computer system was developed which used token-passing Petri nets to model the dynamic interconnections, but they were ahead of their time and knew that data-flow machines were a better fit. Recent developments in grid computing make full-scale modeling of this type of problem feasible.

The Sedelows foresaw the possibility of:

[11] This research is published in *Getting at Disciplinary Interdependence: A Report on Research*, L. J. Old (Ed.), Dr. Walter A. Sedelow Jr., Dr. Sally Yeates Sedelow, Co-Principal investigators. University of Arkansas Press (450 pages), 1990.

> …utilizing a single scalable science and technology for all of these researches, at each and every level of complexity. Thus, there is one general purpose methodology…from the senses of a given word, up through the level of the associative semantics of whole vocabularies, and on through the level of what creates the sense of cohesiveness within a verbal text, and through [to] the level of the grouping of such texts. (W. A. Sedelow Jr., 1991, p. 89)

The model of abstract thesauri (part of which was discussed earlier in relation to semantic chains) developed by Robert Bryan, a mathematician and former student of the Sedelows, came out of their research. Bryan's formal model also defined graphs, stars, and molecules (in analogy to the chemical entities). It brings order to the organic, loose collections of words and senses in the thesaurus structure and, according to Sally Sedelow, both empirical and theoretical work are "empowered by the formal model's facilitating access to the richly complex associative structure of the Thesaurus" (S. Y. Sedelow, 1991, p.108).

As noted previously, their grand model takes account of dynamics within the inter-connectivity of complex systems. They were supporters, friends, and associates of Ross Quillian during his excommunication in the times of McCarthyism. Working within the paradigm of associational psychology, Quillian (1967) proposed a spreading-activation theory of human semantic processing. Collins and Loftus (1975) note that this was a forerunner of a number "of global theories of semantic processing based on network representations, in particular, those of Anderson and Bower (1973), Norman and Rumelhart (1975), and Schank (1972)" (Collins & Loftus, 1975, p. 424; quoted in S. Y. Sedelow, 1991, p.108).

The Sedelow's work on Roget's Thesaurus is too broad and covers too many years to be summarized in one section, but suffice it to say that they concluded that RIT "is a remarkably accurate representation of English word sense relationships" and "might be accurately regarded as the skeleton for English-speaking society's collective associative memory" (ibid).

Small Worlds

This topic, small-worlds, is given a separate section as it describes a model that can account for much of the implicit structure of Roget's Thesaurus.

The small-world model (Travers & Milgram, 1969) derives from the observation that people find, when first introduced, that they know people in common, There are many other variations on this theme, such as "went to the same school," "come from the same town," and so on, but Stanley Milgram set out to quantify how separated, or not, people really are from each other in terms of connections through other people. His experiment, where he had people pass letters to friends and acquaintances, recording the paths taken by the letters, confirmed our common assumption: that it really is a small world.

A mathematical model developed from Milgram's experiment has been found to be applicable to diverse natural phenomena (Watts 1999; Watts & Strogatz, 1998). The essence of the model is that in some large networks, such as social networks, the connectivity is such that no point, or node, in the network is ever far from another. The global human social network is only as wide (in terms of the average number of nodes needed to connect any two people) as 6 nodes—or six degrees of separation. This was corroborated recently by Watts and colleagues (Dodds, Muhamad, & Watts, 2003) in an experiment that used email instead of letters.

Small-worlds may be characterized by particular measures. Word association data has about (on average) 3.0 degrees of separation. Old (2000) showed that Roget's Thesaurus satisfies the criteria of being a small-world network, and Young (1993) showed that the neural network of the brain also fits the criteria. Other work (Steyvers & Tenenbaum, 2001; Motter, de Moura, Lai, & Dasgupta, 2002) has found that Roget's Thesaurus (1911 edition) has about 3 degrees of separation. WordNet has a higher degree, but this may be due to the fact that it has been organized into a classification structure that separates verbs from nouns from adjectives, and separates more general words from more specific words.

A small-world network is not a homogeneous network—it is "lumpy," with sparse areas and highly connected clusters. Kleinberg (1999) showed that the World Wide Web is also a small-world. Because URLs are directed (links go in only one direction) Kleinberg classified the highly connected nodes (Web sites) into those that linked to many Web pages and those that were linked to by many Web pages. The Google search engine also . Old (2000) theorized that the high-density clusters in Roget's Thesaurus were primitive notions, elaborated by thousands of years of human experience and language development. Steyvers and Tenenbaum (2001) theorized that the Roget's network clusters were instead the results of the seeding of concepts in early child development, as children learned speech.

The distribution in the connectivity is caused by growth and preferential attachment. *Preferential attachment* means that the probability of a new node being connected to an existing node is proportional to the number of links that this node already has (Albert & Barabasi, 2002, p. 74)—in other words, the rich get richer. Steyvers and Tenenbaum (2001, p. 5) note that the growth process "is a kind of semantic differentiation, in which new concepts correspond to more specific variations on existing concepts and highly complex concepts (those with many connections) are more likely to be differentiated than simpler ones." Conversely, a *complex concept,* with many connections, may be seen to act as genus, attracting and classifying new, like-minded (although differentiated in some way) concepts. Those that are *simple concepts* are already differentiated and unambiguous by their isolation. This last comment is discussed further in Chapter 5, Section: *Part-of-speech Patterns*, under *Genus and Differentiae*.

The small-world model suggests the (common-sense, perhaps) probability that the underlying meaning of words form a vast interconnected semantic network. The words developed to express these meanings, if they formed a complete coverage (and Roget's entries do, to the extent that the list is kept current), would also form such a network. We know how the Categories arose—by Roget forming clusters of like meaning words, and categorizing them by general notion. But if the actual organization of words is a small-world, how then do the Categories remain separated as words are added? Roget's son, and the second editor, John[12] knew this was a problem:

> Any attempt at a philosophical arrangement under categories of the words of our language must reveal the fact that it is impossible to separate and circumscribe the several groups by absolute boundary lines. Many words, originally employed to express simple conceptions, are found to be capable, with perhaps a very slight modification of meaning, of being applied in many varied associations. Connecting links, thus formed, induce an approach between the categories; and a danger arises that the outlines of the classification may, by their means, become confused and eventually merged (Roget, J. L, 1879, p. ix).

The alternative is for all related senses (Synsets of words) to be repeated, separately under their relevant Categories. But that also has drawbacks.

> Were we, on the other hand to attempt to include, in each category of the Thesaurus, every word and phrase which could by any possibility be appropriately used in relation to the leading idea for which that category was designed, we would impair, if not destroy, the whole use and value of the book. For in the endeavour to enrich our treasury of expression, we might easily be led imperceptibly onward by the natural association of one word with another, and to add word after word, until a group would successively be absorbed under some single heading, and the fundamental divisions of the system be effaced. The small cluster of nearly synonymous words, which had formed the nucleus of the category, would be lost … and it would become difficult to recognize those which were peculiarly adapted to express the leading idea (Roget, J. L, 1879, p. x).

So either the categories become so interconnected that they are indistinguishable, or they become so big their core ideas cannot be discriminated. This reflection of the small-world phenomenon became more of a concern for Roget Jr., as he added more and more words. The only solution he foresaw was to use cross-references (this was in contradiction to his father's advice, which had been to repeat related Synsets under every category). So the cross-references now also participate in the small-world network and "may … be looked

[12] Though overshadowed in history by his father's work, J. L. Roget demonstrated an intelligence and word sense similar to his father. Having learned from his father, he produced the edition that was chosen as the basis of the American branch of Roget's Thesaurus.

upon as indicating in some degree the natural points of connection between the categories" (ibid, p. xi). They solve the big problem, that "as would be in any classification of language, a large proportion of expressions… lie on the ill-defined border between one category and another" (ibid, p. xi)

The thesaurus structure reflects the characteristics of the small-worlds model introduced by Milgram. A comparison between RIT and the Oxford English Dictionary shows that the same information is stored in both texts (Old, 1991b; 1993). The main difference is that the words in RIT found labeling classifying concepts (Classes and Categories), are found at the level of sense definitions in the OED. In-as-much as the OED also associates words in sense definitions, it is most likely also a small-world network. With both these texts in electronic form, and word association data available, using small-worlds as a common model opens the possibility of real comparison and interchange between lexicons, and automatic data-mining of free text.

Metaphor and Analogy

Perhaps the amount of connectivity in Roget's is surprising to the reader. The main reason is that polysemy is so common among English words. Also, as mentioned above, English often has Latin, Greek, French, and other foreign terms in addition to the Germanic (Anglo-Saxon) words to describe the same sense. This probably exaggerates the amount of synonymy in the language, but how do words become polysemous in the first place? Probably by old words being put to new uses. There is a whole spectrum of mechanisms by which this is done, all of which relate to or boil down to, metaphor and analogy[13].

By taking the context in which a word is usually used, or the connotation(s) of the word, or the features or attributes of the word's meaning, some facet of its sense is used in a new situation creating, in effect, a new word. This new word may label a new concept, or an old concept now made explicit for the first time. If others in addition to the original neologist use the word in that same sense, it enters the language—just as foreign terms or technical terms enter the language to label or describe new concepts.

The longer a word is around (the older it is) the more likely it is to be used in new ways, especially if it is used for some feature that permeates our daily lives or is important to our world model. Not only the original sense (what ever that may be) can be extended, but also the new senses can be extended; sometimes becoming so removed from the original sense that the synonyms of the original word will no longer serve as synonyms for the word in its new senses. For example *over* is known to derive from something like "UPER-" (hypothesized Indo-European (IE) root). It is seen in German as *über*, Latin as *super*, and Greek as *hyper*. It is no coincidence that it sounds like *upper* and shares the morpheme "ove" with *ab-ove* (IE root "UPO-"). So intuition alone should tell us that it might have something to do "up-ness."

[13] Mechanisms for denoting the concepts may involve word formation devices including affixing of derivational morphemes on an existing word (e.g., "jump"-"er"), combining two or more words to form a compound word (e.g., "street"-"corner"), and blending of existing words (e.g., "brunch" from "breakfast" and "lunch"). Nouns can be extended to become "denominal verbs" and verbs can be nominalized.

But, to extend the example used earlier, if I walk daily to my friend's house by following the path that goes over the hill, that is, I walk over the hill, then I might say my friend's house is (by following the path) *over* the hill—or *beyond, past, on the other side of*, the hill. However I will never say that my friend's house is <u>above</u> the hill, because there is no "up-ness" involved in the position of my friend's house in relation to the hill. On the other hand if I pile corn up in a bushel basket but put a greater amount than a bushel in measure, I can now say the measure has gone *beyond, past, over* <u>and</u> *above* the amount of one bushel because the *extra* amount is situated at an upward position with respect to the full measure. So *above* and *over* share something in their meaning, semantics, or essence, yet differ in some way that excludes *above* from meaning (being a synonym of the phrase) *on the other side of.*

Above comes from Teutonic *abufan* (*a-be-ufan*). The prefix *a-* means *at* (so *abufan* > at-by-up), and it is that which separates the semantics of *above* from the semantics of *over,* in the example above. That is not to say that changes to the form of the word are necessary to separate the semantics of its descendants. Not one of the following words, along their (very rough) meanings, is a homograph—they all derive from single IE roots: *appropriate* means both "take" and "correct;" *fast* means both "diet" and "quick;" *hide* means both "skin" and "conceal;" and *bank* has many meanings which appear unrelated on the surface.

Claudia Brugman and George Lakoff (Brugman & Lakoff 1988; Brugman, 1981) observed these differences and overlaps in meaning between senses and called them “transformations” forming “metaphorical links” between senses. By analyzing the transformations of senses of *over* they were able to build a graph which they called a radial category. They called it “radial” because they viewed it as like a cartwheel with the “central,” basic, (or possibly original), over/above sense at the center, with the other senses extending transformation-on-transformation from it. They showed that the reason these transformations worked was that there were always some features in common between adjacent senses, and that a transformation involved “carrying” some of those features across, adding some new features unique to the new sense, and leaving some features behind. According to Brugman and Lakoff the common feature that made it an *over* sense was that of having some relation between some landmark object and some reference object, or “trajector”. For example, “the light is over the table” has the light as the reference object (trajector), the table as the “landmark” object, and *over-ness* as the integrating, orientating relation between them. Such sets of features Brugman and Lakoff represented diagrammatically and named them “image schemas.”

Lakoff, who has written several books on metaphor and meaning, believes that the essence of metaphor is “the understanding and experiencing of one thing in terms of another,” and argues “that metaphor is pervasive in everyday life, not just in language but in thought and action” (Lakoff & Johnson, 1980, p. 3). Douglas Hofstadter prefers the term “analogy” (which will not be discriminated by technical definition, in this dissertation) and states “the grasping of one situation in terms of another is so common the we tend to forget that what is going on is, in fact, analogy” (Hofstadter, 1995); and “it’s the very blue that fills the whole sky of cognition—analogy is *everything*, or very nearly so, in my view" (Hofstadter, 1999, p. 1). More recently, Theodore Brown has written that "… metaphorical reasoning is at the very core of what scientists do when they design experiments, make discoveries, formulate theories and models, and describe their results to others – in short, when they do science and communicate about it" (Brown, 2003, p. 14).

Understanding and experiencing one thing or situation in terms of another relates back to antonymy, as discussed earlier. In antonymy there is usually a reference point (or object, or word, or sense) to which the contrasting sense is compared (*logical—illogical*). The shared dimension (*temperature*, for *hot-cold*; or *logic* for the previous example) provides the bridge, or mapping. Roget quotes the great philosopher, Hume, as noting:

> an universal observation which we may form upon language, that where two related parts of a whole bear any proportion to each other, in numbers, rank, or consideration, there are always correlative terms invented which answer to both the parts, and express their mutual relation. If they bear no proportion to each other, the term is only invented for the less, and marks its distinction from the whole. (Hume, *Essay on the Populousness of Ancient Nations,* quoted in Roget, 1852, p. xiii*)*

Examples of the "correlative" terms are "man and woman," "stranger and citizen;" while carpenter, and tailor have "no proportion to" (are not in approximately equal numbers to) those who are not carpenters or not tailors. In other words, when classifying things we invent labels to differentiate only those things that are of approximately equal value. Perhaps this follows from the need to have some commonality, or axis, to provide a bridge for mapping between the two concepts.

The principle of commonality, or shared features, obviously also applies to synonyms as well. As stated earlier, synonyms rarely are perfect mirrors of each other, but without some overlap or commonality they would not be interchangeable in any sense or situation.

Summary

Roget's Thesaurus is more than a synonym dictionary. Roget intended it as a classification of ideas and developed a hierarchical classification structure to classify words that expressed or represented similar ideas, together. This resulted at the lowest level of the hierarchy in not only sets of synonyms, but also lists of words related to each other by relations such as meronymy (part-whole), and orderings such as taxonomic terms.

Roget also recognized deeper relationships between ideas and arranged words describing opposed concepts under adjacent Categories—Categories that shared a common hypernym; this was accomplished despite resistance from his editors. Later researchers capitalized on his insights and developed models and formalisms that connected Roget's work to areas such as natural language understanding for computers, automatic translation between languages, models of the mental lexicon, and information processing.

A final note on terminology; the terms used here (from general to specific) *notion, concept, idea,* and *sense,* have not been defined. *Concept* is used mostly, here, as it is the term used in cognitive science; but when talking about (the more general) Categories, *notion* is used; when talking about Roget's goals, *idea* is used; and when the meaning of specific words or Synsets is discussed, *sense* is used. The synonyms for these words from RIT include, {*apprehension, assumption, attitude, belief, concept, conception, conclusion, consideration, estimate, estimation, fancy, feeling, hint, idea, impression, inkling, intimation, judgment, meaning, mental impression, notion, observation, opinion, percept, perception, personal judgment, position, presumption, reaction, sentiment, suggestion, suspicion, theory, thinking, thought, view, way of thinking}*. Goldstone and Kersten, in a comprehensive paper on concepts and categorization, define a concept simply as a "mentally processed idea, or notion" (Goldstone & Kersten, 2003, p. 600). "The concept of **dog** is whatever psychological state signifies thoughts of dogs" (ibid). The same may be said here of *notion, idea*, and *sense*, as they are discriminated only by context as described above.

Chapter 3 Research Questions and Methods of Analysis

Research Questions

The "entry point" (Alford, 1998) for this research is:

- What are the patterns of connectivity within Roget's Thesaurus as they relate to the conceptual structure?

This entry point has been refined and reformulated through iterations between theoretical and empirical "tracks of analysis" (Alford, 1998). The initial breakdown of the entry point is:

- What is the explicit structure of meaning, as reflected by the synonymy and polysemy relations in Roget's Thesaurus?
- What is the implicit structure of meaning, as reflected by the synonymy and polysemy relations in Roget's Thesaurus?
- What models may account for these patterns?

Structure, as used here, relates to connectivity. Connectivity is based on relations, where a relation is defined as a subset of all possible connections between (the Cartesian product of) two sets. For example, the "word-sense" relation is between two disjoint sets, Words and Senses. The elements in the two sets in a relation may be identical. For example, the *synonymy relation* is a relation between Words, a "word-word" relation; so that word *a* (drawn from the set A, of all words) is related to word *b* (drawn from the set B, of all words); where word *a* could be identical to word *b*. An obvious example of a set is the “synset,” discussed frequently throughout this dissertation. But there are also sets of cross-references, sets of Categories, sets of parts-of-speech (such as the set of verbs), and sets of senses. Word fields and semantic neighborhoods are also sets.

The explicit structure, such as the hierarchy (Classes and Categories), body (senses and Synsets), Word Index and cross-references, of Roget’s Thesaurus is described in Chapter 4 using graphs, and illustrations, along with examples from the text. The explicit structure is also compared to the explicit structure of some selected, similar data sources.

The implicit, or tacit, *inner structure* is explored in Chapter 5 using the methods described in Section: *Overview of Methods of Data Analysis*, later in this chapter; including descriptive statistics. In addition to tacit patterns formed by the more-explicit relations of polysemy and synonymy, hidden patterns in the grammatical categories (parts-of-speech, such as noun and verb), cross-reference structures, and word-sense structures are also explored in Chapter 5.

Limitations

Roget’s Thesaurus, like any lexicon, consists of lemmas. It is not text in service. For example, only the lemma **run** (vb), and not {*runs, running, ran*}, is represented. This may reduce the generality of results gained from studying Roget’s, for example in comparison to data from text corpora such as the Brown Corpus (a collection of texts built in the 1960s as representative of American written language). Consequently, knowledge of, and data, related to word morphology would need to be used for some applications of results from this research[14].

There are errors in the original text that forms the basis of the main thesaurus database. For example there are cross-references that were not updated when Category numbers were updated between editions. There were typographical errors in the machine-readable version, although these have largely been corrected in the database version. There were errors in attribute values generated for the machine-readable version and these have been carried over into the database version. For example 70 of the 71,384 Synsets (WordNet terminology for "sets of synonyms," or senses) have incorrect synonymy counts (the number of synonyms in a sense). Such errors are still being discovered, and corrected through hand editing. Sampling, however, suggests that attribute errors are less than one per thousand entries for all attributes.

This study emphasizes an information processing, computational-lexicographic approach. Only where relevant have current research and knowledge from the fields of linguistics and cognitive science been included.

The grammatico-syntactic structure of language is ignored in this study, except as it relates to compound-word forms.

[14] Wordlists relating base-words to their conjugated or pluralized forms can be used to expand lemmata, and are readily available as downloadable files from the Internet.

Except as correlated with word association data, the generality of conclusions taken from associational data in this study may be applicable only to information processing and lexicon development, not to human cognition.

Except as correlated with corpus word frequency data, the generality of conclusions taken from this study is applicable only to information processing and lexicon development, not to theories of natural language usage.

The age of the data used here is relevant in that the latest version of Roget's Thesaurus used was published in 1962. However, as will be shown, the semantic structure is little affected by those words that are added to the language most frequently, such as those from science and technology.

The focus of this research is on concept patterns in the semantics of Roget's Thesaurus. In order to constrain the breadth of the study, and its implications, the depth of analysis of, and comparison with, the compared data sets has been limited. However the potential application of new insights into meaning resulting from this study should outweigh these limitations.

The 1962 edition of Roget's International Thesaurus is a snapshot in time, both of Roget's Thesaurus, Western culture, and the English language. Consequently, results and conclusions drawn from studying this edition are placed in context by comparing them with results drawn from comparable data sets.

- to control for the types of changes that occur in Roget's Thesaurus over time, major results are compared with results from a database version of the 1911 edition of Roget's Thesaurus (edited by Sylvester Mawson)
- to control for structural or organizational patterns resulting from philosophical and design decisions made by Roget in the construction of Roget's Thesaurus the structure and organization of a relational database version of WordNet is compared
- to attempt to account for the evolution of broad, over-all, or global semantic patterns a database of Indo-European roots is used
- to attempt to account for local semantic patterns, results are compared with a database of word association norms
- to control for Indo-European-centric concept patterns Roget's Thesaurus is compared with a database of Mandarin Chinese.

This has required databases to be constructed and studies to be conducted on these control sources similar to those conducted on Roget's International Thesaurus.

Summary of Sources of Information

Roget's International Thesaurus (RIT), 3rd Edition (1962) in database form, is the central data source used for this research. The semantics and word associations of Roget's Thesaurus have been "culturally" validated for 150 years. Partly for this reason RIT was chosen as the basis for the "lexicon in the background"—"a general purpose reference work of semantically-related words” (S. Y. Sedelow, 1974, p. 2)—for a study of translational associations, funded by the Office of Naval Research, in support of natural language processing efforts to translate Russian Military Strategy in the 1960’s and early1970’s. RIT was converted to machine-readable form at that time, and converted to database form by this author in the early 1990’s.

The machine-readable form consisted of entries (one word may have several entries), each with 22 attributes. Attributes, coded as integers, consist of features associated with each entry in the actual text. Examples of these attributes are the entry’s part-of-speech, its font type (bold or italic, and so on), polysemy (total number of senses in RIT for the word which the entry represents), and the RIT Category in which the entry is found (its location in the text).

The database version, in addition, includes lookup tables for the codes (what the integer codes mean in human-readable text); the structure of the full Roget hierarchy; labels for the classes and categories; cross-references, in a separate table; the antonymy relation between categories; and several tables to facilitate processing, such as a separate Entry index. Following research and processing by Talburt and Mooney (1990a, 1990b) and Jacuzzi (1991), indexes to derived partitions, or components, were also added to the database (see Section: *Models of, and Research on, the Structure of Roget’s Thesaurus*, for details).

RIT is still under copyright. Dr. Walter Sedelow Jr. and Dr. Sally Yeates Sedelow, Professors Emeriti, University of Arkansas, were granted rights to it for research purposes. They have allowed the thesaurus to be used for this dissertation research, and are consultants and informal committee members for this dissertation.

In addition, the 1911 edition of Roget's Thesaurus, available as an electronic book from Project Gutenberg (Project Gutenberg, 2002), has been converted to database form by this author, and is compared here with the 1962 RIT version. The 1911 edition is virtually identical to Roget's first (1852) edition, and retains the synonyms-antonyms opposed category organization of the original Roget's hierarchy. Differences and similarities between the 1911 version and the 1962 version highlight changes and stabilities in technology, culture, and word usage; and editing philosophies.

Other sources of data used for this study are:

- Word association data from the University of South Florida (Nelson, McEvoy, and Schreiber, 1998). This source has been converted to database form by this author. The original data consists of about 5,000 "cue" words, or prompts, and about 10,000 "target" words, or responses. Almost all of the cue words are also targets, and can be found as entries in RIT.
- Indo-European Roots and English Base Words Database. This is used to analyze semantic patterns in RIT and test theories of concept development (developed by this author from various sources).

- Chinese-English Association Dictionary. This is used as a non-Indo-European language source for comparison with patterns in RIT and for testing some theories developed during this research.
- WordNet lexical database converted to relational form by Dr. Uta Priss, School of Library and Information Science, Indiana University and Randee Tengi, Princeton University Cognitive Laboratory. WordNet is an alternative method of structuring synonym sets, based on psycholinguistic theories of human lexical memory developed by Dr. George Miller at Princeton University. Connectivity patterns in RIT are compared to those of WordNet.
- Word lists such as frequency data, homographs (developed by this author, and discussed in Old, 1991a), and common "base" words, from various sources.
- A dozen British and American editions of Roget's Thesaurus published between 1852 and 2002.

For a more detailed description of the data sources and acknowledgements of their sources see *Appendix A: Sources of Inform*ation.

Overview of Methods of Data Analysis

In order to identify the patterns of connectivity within Roget's Thesaurus the elements that constitute the thesaurus are first identified and a terminology is defined. Next, the explicit structure of the thesaurus is identified and the elements and structure quantified with descriptive statistics. Finally, the implicit or hidden inner structure of the thesaurus is identified.

The primary method of analysis for the implicit structure of RIT is visualization. The thesaurus is detailed and complex—a vast semantic topology. Visualization reduces cognitive load, and a well-designed graphic, according to Tversky (1999), can "compensate for limitations in human memory and information processing." Card, Mackinlay, and Shneiderman view information visualization as an extension of cognitive working memory, defining it as "the use of computer- supported, interactive, visual representations of abstract data to amplify cognition," (Card et al., 1999, p. 7). The results of such interactions with the thesaurus data are used in this dissertation to illustrate the inner structure. The interactive visualizations are developed as "adjustable mappings from data to visual form to human perceiver" (Card et al., 1999, p. 17). That is, following Card et al's prescription, by the transformation of raw data to database relations; "visual mappings" of database relations to visual structures; and transformations that produce various views on the visual structures, controlled by adjusting parameters such as position, scaling, and restricting the view to certain data ranges.

In order to view localized data sets (fewer than 100 variables), where a relevance metric between words (Old, 1996) can be applied or where a correlation matrix can be derived, multi-dimensional scaling (MDS) has been used.

Where a topological rather than topographical model is appropriate (where distances between words or senses are not so important as relatedness, adjacency or broad patterns of connectivity) network or graph visualizations are used. For example tree structures are used for viewing hierarchies and formal concept analysis (Wille, 1989), or FCA, is used to derive structural relationships of *semantic neighborhoods.* FCA is especially useful where both graph duals are relevant. That is, where it is most useful to represent relationships between senses, relationships between words, and relationships between words and senses concurrently.

A *semantic neighborhood*, like a word field (Wortfeld) or semantic field (Bedeutungsfeld)[15], refers to a set of semantically related words. For this research a semantic neighborhood differs in that both words and senses are included, and is defined here as the senses of a given word, along with the synonyms for each sense. A *restricted neighborhood* is defined here as a neighborhood which includes only those synonyms that occur in more than one sense of the word. A *sense neighborhood* can be defined symmetrically for senses. In either case "sense" may be replaced with any concept at any level in the RIT hierarchy, and "word" may be replaced by a set of words. The set of words can be from different Synsets. So, for example, the restricted neighborhood of the

[15] According to Elsen (2001), Barrett (1982) was the first to use the term "semantic field;" The first to use "Bedeutungsfeld' was Ipsen (1924); and the first to us "Wortfeld" was Trier (1931).

set {over, above} will include just those words which are synonyms of both words, and just those senses which are shared by both words. *Neighborhoods* are further defined, formally, in Priss, 1996 (pp. 38-39).

Information cartography (Old, 2002) will be used for concurrent visualization of multiple feature dimensions and global patterns within Roget's Thesaurus; and comparisons of Roget's with relevant other lexical databases such as WordNet, Indo-European roots, or Chinese language data. Information cartography is a dynamic, interactive process, and conclusions (relevant configurations) or illustrative examples have been exported as information maps for inclusion in this dissertation.

In addition to the visualization methods of analysis, the data has been browsed via three sets of interfaces—a Structured Query language (SQL) -based database interface; a GUI forms-based database interface; and a web-based interface that includes text and graphical options for interaction and querying, where queries may be made against the on-line version of the thesaurus databases.

Most of the detailed analysis for this research was done using Structured Query language (SQL) to generate the semantic neighborhoods, descriptive statistics, data for visual displays, coordinates for information maps, and contexts (matrices) for FCA lattices.

For further explanations and illustrative examples of the methods mentioned here see *Appendix B: Methods of Data Analysis*.

Chapter 4: Research Findings: The Explicit Structure of Roget's Thesaurus

The explicit structure of Roget's Thesaurus consists of three main parts. Following the front matter is the top level of the *hierarchy* represented by the tabular *Synopsis of Categories.* This is followed by the *body*, or *Sense Index* of the book, which continues the hierarchy down to the lowest level. The Sense Index lists the 1,000 or so Categories (also called headwords, or lemmas, by some researchers) representing the notions found at the most detailed level of the "Synopsis." Each Category contains the Entries—instances of words ordered by part-of-speech and grouped by sense, or *synset* (Miller et al., 1993). Cross-references to similar concepts elsewhere in the book are found adjacent to some Entries. At the back of the book is the *Word Index*, listing the words in alphabetic order, along with their senses ordered by part-of-speech. The senses are represented in the Word Index as references to locations in the Sense Index.

On any particular page the relations of synonymy and antonymy can be seen. Although not all words collocate as synonyms (some occur in lists), most do; and although not all adjacent Categories present as antonyms, about eighty-percent do; moreover, although not all Categories contain cross-references at the Synset level, likewise, about eighty-percent do.

This chapter describes the three parts of the book and the organization of their component elements.

Detailed Structure of Roget's Thesaurus

Roget's Thesaurus was developed as a classification of word meanings. The hierarchical structure is a classification tree with the highest-level Classes based on the Aristotelian Categories. The lowest level grouping contains the Entries—sets of words representing similar ideas, and often referred to as the sets of synonyms, or "synsets."

Sets of synonyms represent a "sense" in RIT and the most representative, or frequently used, words for that sense are in bold-type. The terms "sense" and "Synset" are discriminated here, in that *sense* means the concept, or idea, represented by the location in the classification tree where the set of synonyms is found; whereas *Synset* means the set of synonyms corresponding to a sense, along with their attributes such as parenthetical information, cross-references, bold-typeface, location in the classification tree, and so on. *Location in the classification tree* is identified using index numbers (discussed further, below). Entries in a Synset share a common part-of-speech, as well as a common idea or sense.

Synsets are grouped into Paragraphs. Entries in a Paragraph share a broader notion—one that overlaps with the ideas or senses represented by the individual Entries. For example, the following Paragraph contains three Synsets:

{filament, shred; thready, threadlike; ropy, stringy, wiry}[16].

[16] Synonyms are separated by commas, Synsets by semi-colons, and paragraphs by periods. Consequently, some thesaurus literature refers to the sets of synonyms, or Synsets, as "semi-colon groups."

Paragraphs are arranged by part-of-speech, so Entries in a Paragraph also share a part-of-speech in common. Paragraphs usually do not have an identifying label, or word that captures the notion shared by the Synsets contained in the Paragraph. Exceptions are lists or collections of words (usually nouns) on a topic such as *ships* or *countries*. For Synsets containing close synonyms, the bold-type Entries serve the function of labels. The Paragraphs containing such Synsets could be viewed as being labeled by the set of bold-type Entries within them.

Paragraphs are enumerated, ordered by part-of-speech: Noun > Verb > Adjective > Adverb > Preposition > Phrase; and occasionally including Pronoun > Interjection > Conjunction > Interrogative. Collections of Paragraphs are grouped under a broader, labeled notion called a Category. Adjacent Paragraphs that share part-of-speech but not a common idea still share the broader notion represented by the Category label.

In RIT the hierarchy has a six-level structure: a hierarchy, or tree. The top-level Classes are at "level 1;" the Synsets are at "level 6;" the Paragraphs are at "level 5;" and the Categories are at "level 4." The two remaining levels, 2 and 3, are sub-classes indexed by Roman numerals (Roman-level subclass) and letters (Letter-level subclass), respectively. Some branches of the RIT tree are "short"—there are missing letter-indexed (Letter-level) subclasses. In order to normalize the tree structure for the RIT database so that all Categories, Paragraphs, and Synsets are on the same level, the RIT editors duplicated the Roman-level subclasses to form Letter-level classes in the cases where they were "missing" (this is discussed in more detail in Chapter 5).

All Categories are uniquely identified by an integer number. So, for example, the hierarchy (or tree branch; or "path") for the first sense of the word *ordering* in RIT has the following structure (showing the Level Number, Index, and Label for each level):

1. **1** Abstract Relations
2. **IV** Order
3. **A** Order in General
4. **60**: Arrangement
5. **1**: ***Nn*** Reduction to Order
6. **5**: {arrangement, ordering, "confusion unconfused" [Young]}

As noted earlier, a word may have more than one sense. An Entry in the thesaurus represents just one sense of that word. A word may have several Entries in a single Category if it has several closely related senses, or, as is sometimes the case, several parts-of-speech. For example, the word "cut" can act as a verb ("to cut") and a as noun ("a cut"), both of which appear in the Roget Category "Disjunction." The collected senses of a word, therefore, may be represented as a sub-tree of the thesaurus hierarchy. Figure 2 shows the "sense tree" for twelve of the twenty-two Roget senses of the word *over*, down to the fourth, or Category, level. There are two senses under Category 36: Superiority. Category 40: Addition, which contains one Entry for (one sense of) *over*, is highlighted.

Broad Structure of Roget's Thesaurus

As stated earlier, the text of Roget's is divided into three sections: the multilevel Hierarchy, which is a tree structure; the "body" or "numbered section," which is ordered by Category Number, called here the "Sense Index"; and the alphabetically-ordered "Word Index" at the back of the book. Figure 3 shows these components, schematically.

The database version of RIT contains tables only for the Hierarchy and the Sense Index of the book. The Word Index is generated, as needed, from the Sense Index.

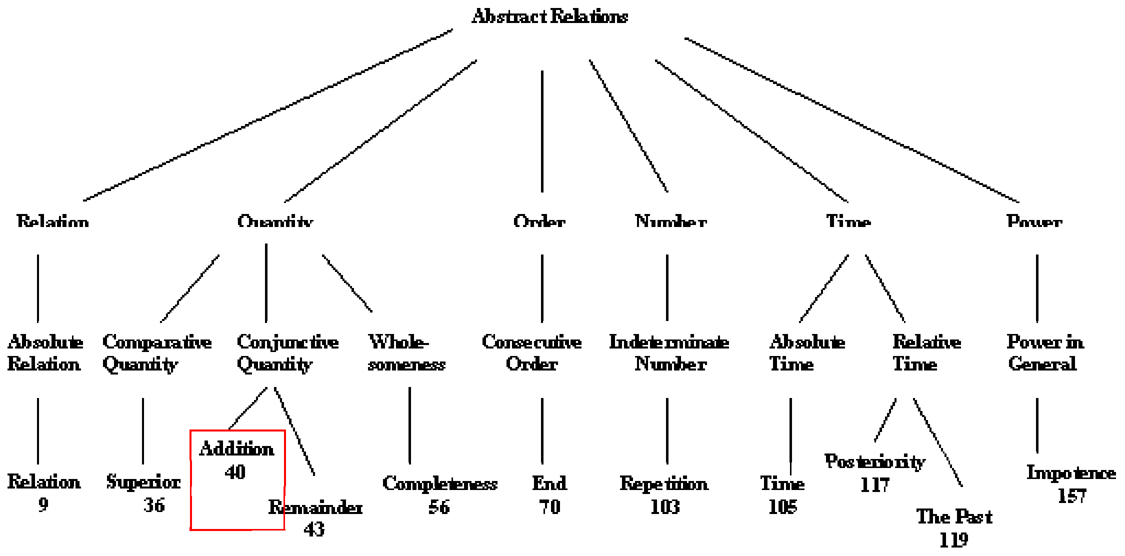


Figure 2. **Partial Sense Tree showing the conceptual hierarchy for senses of "over" (12 of 22 senses)**

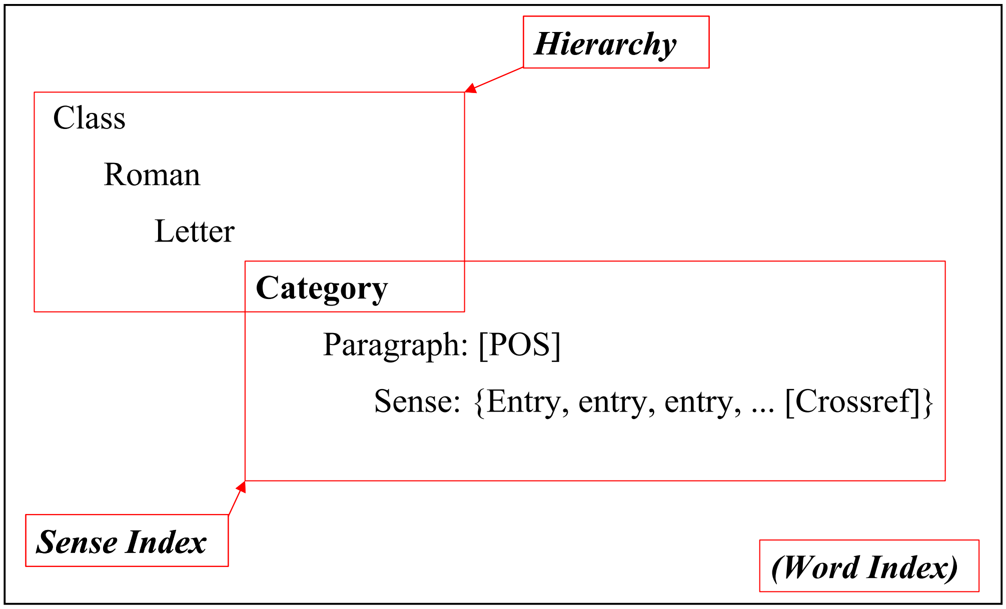
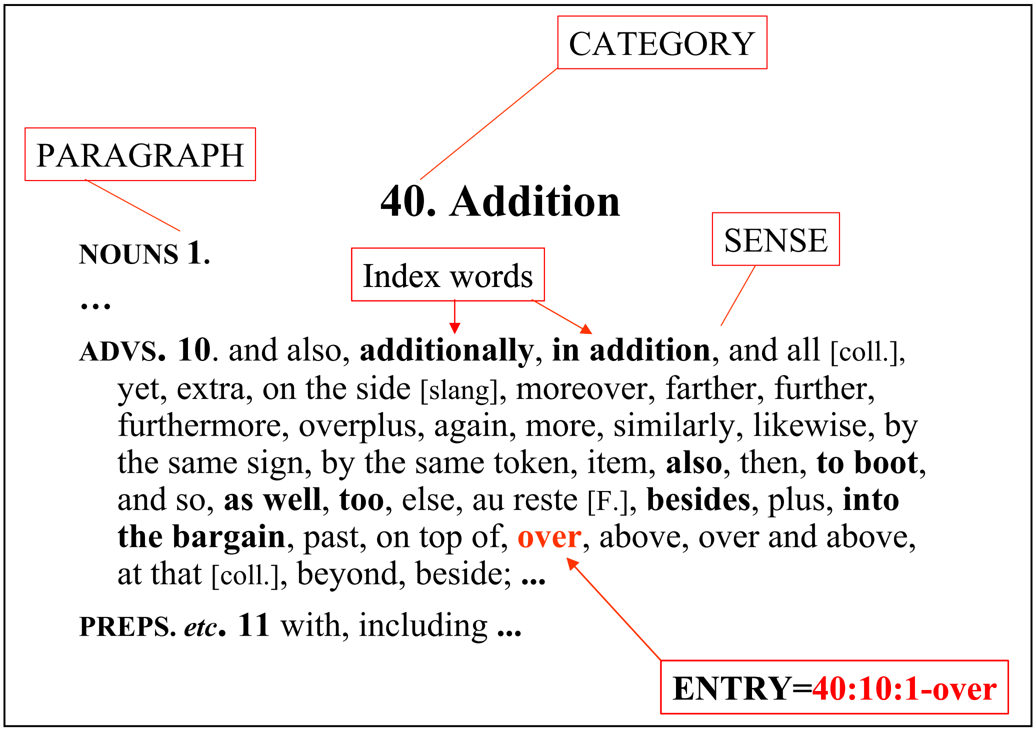


Figure 3. **Overall structure of the text of Roget's Thesaurus**

Because Category numbers are unique, they provide a bridge joining the Synopsis of Categories to the Sense Index and from the Word Index to the Sense Index.

Figure 4 illustrates the elements of a sense (or Synset) in the Sense Index, with an example of one sense, or Entry, of the word "over" (highlighted in red) taken from Category 40: Addition; Paragraph 10: Adverbs; Synset 1 (indexed as 40: 10: 1). Bold-type synonyms from the Sense Index, such as "additionally," are generally chosen as index words to act as discriminators in the Word Index. Synset 40:10:1 is also one sense of "and also," one sense of "additionally;" and so on, for all of the other Entries in 40:10:1.



40. Addition

NOUNS **1.**

…

ADVS. **10**. and also, **additionally**, **in addition**, and all [coll.], yet, extra, on the side [slang], moreover, farther, further, furthermore, overplus, again, more, similarly, likewise, by the same sign, by the same token, item, **also**, then, **to boot**, and so, **as well**, **too**, else, au reste [F.], **besides**, plus, **into the bargain**, past, on top of, **over**, above, over and above, at that [coll.], beyond, beside; **...**

PREPS. *etc*. **11** with, including **...**

***Figure 4*. The Sense Index of Roget's Thesaurus.**

To show how the same information is restructured in that section of the text the equivalent Word Index entry for “over” is illustrated in Figure 5. Here the index word from the Sense Index, “additionally,” identifies the particular sense. Category labels are not used as differentiae in the Word Index, as they are too general. Paragraph labels, however, may be used—especially if the entry is in a list.

The Synset index number for Entry “40:10:1—over” is highlighted in red. Only the Category and Paragraph index numbers are used in the Word Index of the text—possibly because of the inordinate effort that would be required to update Word Index Synset numbers during revisions to the Sense Index. This is not a concern for the automatically generated Word Index of the database version of RIT.

Occasionally, if the entry is a member of a class of words such as in a list, meta-information similar to that found in a dictionary may be used. For example, the word “blouse” appears in the Sense Index list, **230:54** (Category 230: Clothing; Paragraph 54: Waists, Shirts). That sense is entered in the Word Index under *blouse* as “**types of ~ 230.54**.”

Under some circumstances the senses of a word can be found in a phrase. Indexing these correctly requires heuristics not available to the database version of RIT. Under the Word Index entry *home* there are phrases such as **“**be at -**,” “**at - with**,” “**bring – to” (“be at home,” “at home with,” “bring home to”)**.** In the automatically-generated Word Index, *be at home* is instead listed after the entry *be at cross purposes with* and before the entries *be* *at home in*, and *be at home with*; while *bring home to* is listed after bring. That is, none is listed under home**.**

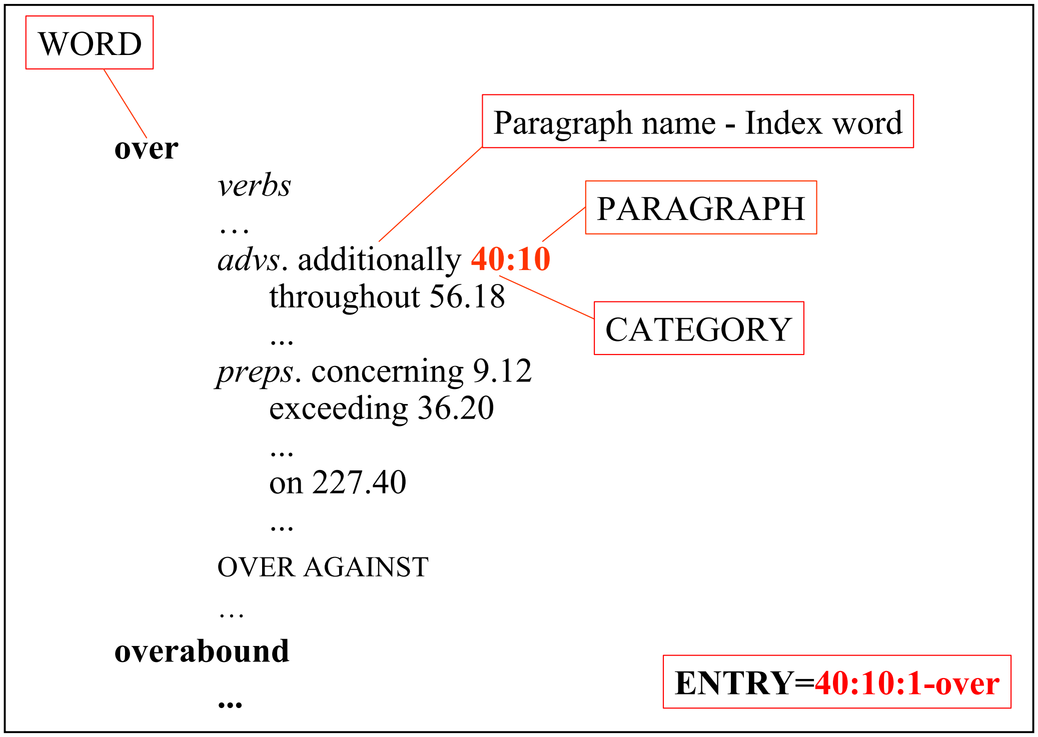
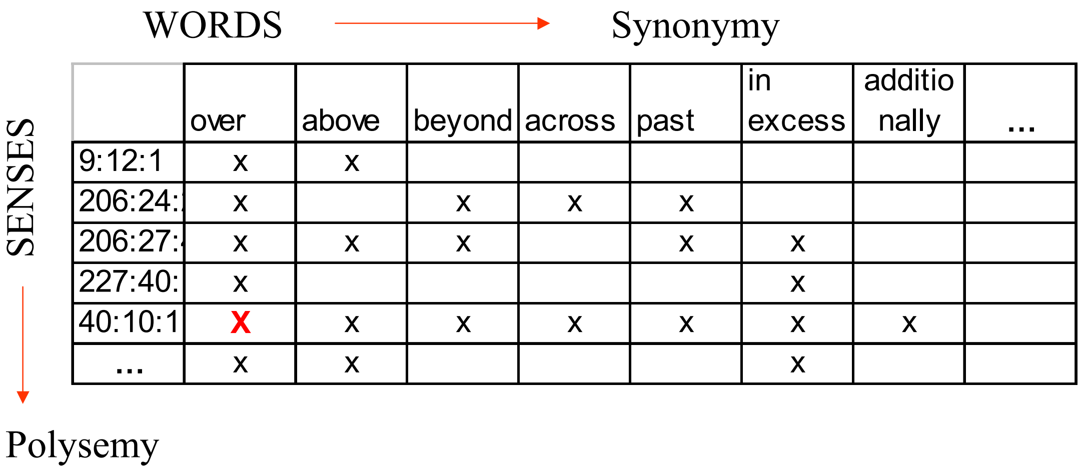


***Figure 5*. The Word Index of Roget's Thesaurus**

The relationship between the Sense Index and the Word Index is symmetric. The number of senses of a word is termed the word's "polysemy." The symmetric attribute, the number of words (synonyms) that describe a sense, is termed the sense's "synonymy." Together they constitute the Word-Sense relation. This can be seen in Figure 6, which represents some of the senses and synonyms of the word "over." The Entry, "over-40:10:1," which occurs above in both the Sense Index and the Word Index (Figures 4 and 5), is represented by a red "X." This is a succinct representation of the relation between words and senses, but it lacks the detailed attributes and context found in the text.



	over	above	beyond	across	past	in excess	additionally	...
9:12:1	x	x						
206:24:	x		x	x	x			
206:27:	x	x	x		x	x		
227:40:	x					x		
40:10:1	X	x	x	x	x	x	x	
...	x	x				x		

Figure 6. **The Word-Sense relation. (Synonymy is the number of words in a sense. Polysemy is the number of senses a word has.)**

Cross-References

Cross-references, as they appear in the text of Roget's Thesaurus, are similar to Entries. That is, they exist in Synsets, separated by commas, as do regular Entries. They differ in that they are index numbers, not members of the set of Words, and represent a relationship between their own Synset and semantically related, but remote, senses.

Cross-references are an explicit shadow of the implicit structure of Roget's Thesaurus. They are analogous to the links between Synsets implied by words shared between Synsets. So "Cross-reference," like synonymy, is a relation. But Cross-reference differs in several ways. A cross-reference is directed—it has an origin and it has a destination. In other words it is not symmetric (although there may be cross-references running in both directions). A further idiosyncrasy is that a cross-reference points not to a single sense, but to a set of senses—always at the Paragraph level or at the Category level.

An example from Category 1: Existence, Paragraph 2: Reality, found in the third Synset, references Category 515: Truth; Paragraph 5: Genuineness. Using index notation as found in the text, the cross-reference appears as follows: **515.5**. The full cross-reference as stored in the database is **1:2:3—515:5.** The referencing, or source, Synset contains {authenticity, 515.5}. The first Synset of the referenced, or destination Paragraph, **515:5**, contains {genuineness, authenticity…realness, reality…}.

The character string, "reality," occurs in both locations—as the Paragraph label in the source location, and as an Entry in the destination Synset—and the Synsets share a string in common: "authenticity". This is a regular, or *normal* cross-reference, but cross-references do not necessarily indicate shared strings, or words in common between the source and destination locations. The main purpose of a cross-reference is to indicate shared meaning, not shared words.

A Synset may contain several cross-references, and these may point to a sequence, or range of senses. For example, a cross-reference found in Category 299: Arrival, points to three Paragraphs:

(Source)
Category 299: Arrival; Paragraph 4: Welcome; Synset: 1 {welcome, greeting}
(Destination)
Category 923: Hospitality, welcome; Paragraph 2: Welcome
Category 923: Hospitality, welcome; Paragraph 3: Greetings
Category 923: Hospitality, welcome; Paragraph 4: Greeting

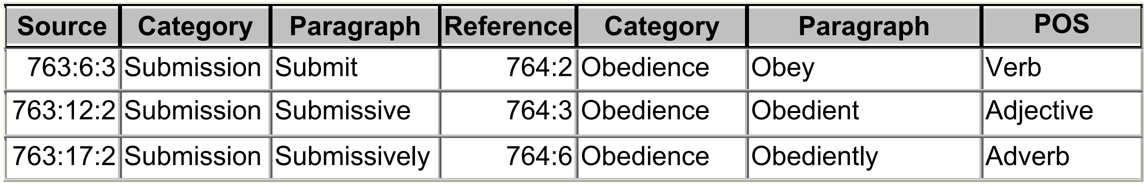
Using the index notation, this is represented as: **299:4:1—923:2-4**. In the RIT database this is represented by three separate rows in the Cross-references table:

299:4:1—923:2
299:4:1—923:3
299:4:1—923:4

This type of cross-reference (belonging to a set of two or more sequential cross-references) will be referred to as a *range* cross-reference.

A cross-reference to a whole Category, a Category-only cross-reference, would appear in the text as “entry_1, entry_2 … entry_n, **923**” (note the missing Paragraph index), and would be stored in the database as **299:4:1—923:0**. These will be referred to as *category* cross-references.

When a set of cross-references from one Category refers to a remote Category at multiple, but non-consecutive locations, there is clearly a strong semantic relationship between the two Categories. In the following example the cross-references serve as links between similar concepts listed under the verb, adjective, and adverb parts-of-speech in both Categories. This type (belonging to a set of two or more concurrent cross-references) will be referred to as *multiple* cross-references. Note that no character strings, or words, are shared between the source and destination.



Source	Category	Paragraph	Reference	Category	Paragraph	POS
763:6:3	Submission	Submit	764:2	Obedience	Obey	Verb
763:12:2	Submission	Submissive	764:3	Obedience	Obedient	Adjective
763:17:2	Submission	Submissively	764:6	Obedience	Obediently	Adverb

.

A final cross-reference type, although rare (total 57), is an internal reference, where the source and destination Categories are the same, but the Paragraphs are different. This is termed here a *self-reference*. This type occurred frequently in the original and older editions of Roget's Thesaurus (total 1,946). The goal was to link a concept in one part-of-speech section to the same (or similar) concept in another part-of-speech section; or to link the source Synset to a more-specific set of words relating to the same concept. An example of the first case is found in Category 459: Animal Sounds, in Synset **459:1.2** (nouns) containing {[a] **call,** [a] **cry**; [a] howl, … [an] ululation}, linking to **459:2** (verbs) containing {[to] call [to] cry; [to] **howl**, … [to] ululate}. An example of the second case is found within Category 123: Oldness, in Synset **123:4:3** containing "archaeology," linking to **123:22**, a Paragraph listing branches of archaeology such as "paleoanthropology" and "Egyptology."

There are approximately 3,619 cross-references in RIT. Of these, 2,986 point to Paragraphs and 576 point to whole Categories. Of the Paragraph-referencing cross-references there are 606 of the *range* (sequence of paragraphs) reference type; 1,304 of the *multiple* (concurrent paragraphs) reference type; 1076 of the *normal* (single-paragraph) reference type; and 57 of the *self-reference* type.

There are no cross-references linking truly anonymous concepts. Antonymous notions are classified in adjacent Categories so the Editors may have considered such references to be redundant. There are, however, a few (twelve) cross-references between adjacent Categories linking such concepts as *bequest* and *inheritance.*

As stated earlier, cross-references form a kind of shadow, or skeletal, network structure of the thesaurus as a whole. This implicit structure is discussed further in Chapter 5. The terms, *category-, range-, multiple-, normal-* and *self-referencing* cross-references will be used in that discussion also.

Chapter 5: Research Findings: The Implicit Structure of Roget's Thesaurus

This section describes and illustrates the results of analysis of Roget's International Thesaurus, 3rd Edition (RIT) using the methods outlined under Section: *Overview of Methods of Data Analysis* and expanded on under *Appendix B Methods of Data Analysis*. The analysis focuses on patterns found among the RIT cross-references; patterns found in the semantics of the parts-of-speech of Entries; local views such as semantic neighborhoods; patterns emerging from global views of RIT such as word-overlap (with implied semantic overlap) between Categories; and concludes with core connectivity patterns such as Type-10 Chain components and *conceptual* and *semantic switching centers* among senses and words.

The structures of the main comparison data sources, the 1911 edition of Roget's, WordNet, and word association data are also discussed here, in context, in order to provide insights into the reasons for the differences enumerated in the Section: *Descriptive Statistics*. As RIT and the 1911 edition differ slightly in their structure, this discussion begins with a brief explanation of those differences.

The original and 1911 editions of Roget's Thesaurus had six Classes at the highest level in the Hierarchy. RIT has eight Classes. The Classes that were added were "Physics" and "Sensation."

1.Abstract Relations
2.Space
3.Physics - ***added***
4.Matter

5.Sensation - ***added***
6.Intellect
7.Volition
8.Affections

These additional two Classes represent essentially a reorganization of the lower levels of the older editions, especially of the associated Categories, rather than the addition of new concepts.

The original Hierarchy was a variable-height tree whereas RIT has a mostly balanced 6-level tree. For example, the 1911 edition has nine levels for the Synset 441:1:2, found in Category 441: Vision.

CLASS III: Words Relating to MATTER
SECTION III. ORGANIC MATTER
2. SENSATION
2. Special Sensation
(6) Light
(iii) PERCEPTION OF LIGHT
#441. Vision
1. ***Nn*** **vision**
2. {glimpse, glint, peep}

It can be seen from this example from whence the additional RIT top-level Class "Sensation" had its origins. This and similar rearrangements may reflect differences in philosophy, or perception, between Dr. Roget and the modern RIT editors, of how higher level concepts should be classified. The concepts represented by the Categories under these Classes may be more "primitive" (general) than their associated Categories deeper in the original Roget hierarchy, and so deserving of a separate, higher-level classification. Or the RIT editors could have made the changes simply to construct a more balanced, regular six-level tree.

Examples of the 100 RIT Categories affected by this editing include 139: Change, 140: Permanence, 152: Cause, 153: Effect, 324: Physics, 325: Atomics, 327:Heat, and 332: Cold. These and the other affected Categories are found in 16 Roman-level subclasses. Examples of the Roman-level subclasses are 1:VII: Change, 1:IX: Causation, 3:II: Heat, 3:III: Light, 5:I: Sensation in General, 5:III: Taste, and 5:IV: Smell. In the context of other Categories and Classes in RIT, these are indeed primitive.

Descriptive Statistics

The electronic version of the 1911 edition of Roget's Thesaurus has 1044 Categories, numbered 1 through 1,000. The additional 44 Categories were added (inserted) since the publication of the original Roget's Thesaurus, which had 1,000 Categories. RIT has 1043 Categories numbered 1 through 1040. The three extra Categories are also insertions, such as 306a: Food.

The reason that the more modern 1962 version (RIT) has fewer Categories than the 1911 edition is that when the 1962 version was revised by the RIT editors they split some Categories, present in the 1911 version, into several new Categories in RIT. The editors also aggregated some Categories in the older version into single Categories in the later version and, on the whole, the number of aggregations was greater than the number of splits. Many of the Categories were renamed during this process. An example of the renaming process is where Category 298: Food of the 1911 edition of Roget Thesaurus was split into First Edition RIT Categories 306: Eating, 307: Nutrition, and 329: Cooking.

Then in the Third Edition of RIT, “Food” was inserted back in as Category 306a: Food. All the while, many new words were also added to the Categories.

For perspective on the organization of Roget’s Thesaurus in general, a comparison can be made with WordNet. Like Roget’s, WordNet classifies words into sets of synonyms. But unlike Roget’s, WordNet does not have a single-tree structure. George Miller believed that the mental lexicon, which WordNet was to emulate, had a different organization from that reflected by Roget’s classification tree (the hierarchy). Roget’s has a single label (for the most part) for each node in the hierarchy, and every label is a noun. Miller also constructed WordNet in this way, for nouns, but with a slight difference. Instead of single words labeling the nodes, he reorganized the synsets that contain nouns, to form the tree. So every node is a synset—the higher synsets representing more general notions, and the lowest synsets representing specific concepts. He constructed a separate tree, in the same way, for verb synsets. So, in the following example from the WordNet verb tree, the two lower synsets are both nodes representing specific concepts connected to the upper synset-node, which represents the more general notion implicit in the two lower nodes.

{act, behave, do}
    {act, act as, play}
    {act, dissemble, pretend}

For adjectives Miller chose an opposed arrangement similar to Roget’s opposed synonym-antonym arrangement of Categories. For example, *big* and *small* are opposed. *Big* and *little* are then connected by the fact that *small* and *little* are synonyms. In other words, he used the strongest antonyms as the backbone for the concept then organized the related words around it.

Furthermore, Miller defined many "semantic relations" between entries, and between synsets—relations that were missing from Roget's Thesaurus[17]. An example is meronymy, the Part-Whole relation. RIT is almost barren of semantic relations when compared to WordNet, but it has the most important ones, synonymy and the Word-Sense relation. In controlled vocabulary (special topic thesauri) terminology these are known as Related Term (RT), and Broader-Term/Narrower-Term (BT-NT), respectively.

The average number of words in a RIT Category is about 190. The average number of words in a 1911 Roget's Thesaurus Category is about 85. The Categories that increased the most between versions were (using the RIT Category labels) 413: Animals (1539 words added), 306:a Food (1058 words added), and 410: Plants (958 words added). All of the Categories that increased by at least 450 words (23 Categories) were concrete topics of this nature, or related to technology such as aeronautics, automation, tools and machinery, electricity, and chemicals. These Categories represent the largest Categories in either of the two thesauri compared, so these increases in word numbers represent the continuing important of the notions rather than their recent popularity. The idea that new words are, in a sense, attracted to existing popular topics—the rich Categories get richer—a characteristic of small-world networks, as was discussed in Chapter 2.

[17] The comparison is relevant, as Miller used Roget's Thesaurus as the source data for the construction of his model of the mental lexicon. In this sense WordNet could be seen as a version of Roget's Thesaurus.

When the change in number of words is viewed as a percentage increase the ordering of the Categories changes and the technology Categories rise to the top of the list. The Categories representing astronautics (97% increase), electronics (95% increase), and physics (92% increase) are first, second and third on the list. Also high on the list are the Categories representing atomics, botany, radiation and radioactivity, radar and radiolocators, radio, aircraft, and rockets and flying missiles. These were apparently the topics that underwent the most change between 1911 and the early 1960's. The top-twenty Categories, by percentage increase, all have a percentage increase greater than 84%. There is one RIT Category with a 100% increase—a totally new Category—741: United Nations.

RIT has about 200,000 Entries. This compares to WordNet's 238,000 and the 1911 Edition's 130,000 Entries. RIT has about 113,00 individual words. This compares to WordNet's 174,000 words and the 1911 Edition's 54,000 words. In RIT's defense, the WordNet lexicographers have included a lot more "things" (concrete nouns), such as the complete list of 18,368 names (including Latin names) from the Plant and Animal Kingdoms. Such entries rarely have synonyms and contribute very little to the semantic connectivity within the lexicons, although they do of course make for much better lexicons.

Words with only one sense (a polysemy of one) are called monosemous. About 80% of WordNet words are monosemous. About 75% of RIT words are monosemous (more than 78,000 words). At the other extreme are the highly polysemous words, the majority of which are among the most common words in everyday English usage: words such as "cut," "over," "head," and "line." But not all common words are polysemous. The exceptions are words such as articles, conjunctions, and personal pronouns ("the," "and," "of," "he," "she", and so on), which have at most two senses, yet are the most frequent words of all in daily usage. These words, along with the prepositions ("over," "beyond" and so on) are sometimes referred to as closed sets, because, unlike words naming plants or technology, for example, no new words are being added to the set. A summary of the differences between the 1911 Edition, RIT, and WordNet is given in Table 1a.

	1911 Edition	**RIT**	**WordNet**
Classes	6	8	25 "unique beginners"
Levels (min-max)	6-9	4-6	No fixed bounds[18]
Entries	130,000	200,000	238,000
Words	54,000	113,000	174,000
- per Category	85	190	N/A
- monosemous	65%	75%	80%
Average polysemy	2.4	1.8	1.3

Table 1a. **Comparison of Roget's International Thesaurus (1962) with Roget's Thesaurus (1911) and WordNet 1.7**

The ratio of words to Entries for each lexicon reflects the proportion of semantically rich words to concrete single-sense, or monosemous words. If monosemous words are excluded from the calculation the average polysemy in RIT almost doubles to 3.4 senses—a difference of 1.6 senses. The complementary ratio, the average synonymy of a Synset in RIT (the average number of Entries found in a Synset) is 2.8 words. If monolexic (single-entry) senses are excluded the average synonymy is increases to 3.8

[18] According to Christiane Fellbaum, Research Psychologist and chief lexicographer at WordNet, "the deepest noun tree is 15 and the deepest verb tree is 5 [levels]" (C. Fellbaum, personal communication, June 2002).

words—a difference of only 1.0 word, on average. Figure 7 illustrates the data on average polysemy and synonymy in RIT.

One might expect the monosemous words to all occur as singletons in monolexic Synsets such as lists, so that when removing the *mono*-'s from the calculations one could expect more-or-less similar results between words and Synsets. Most of the monosemous words are indeed in monolexic lists, but many more serve a vital function as implicit differentiae in *multilexic* Synsets. That is, they characterize the Synset giving it a unique identity. When a dictionary definition is constructed it has two components: a genus and one or more differentiae (*genus proximus et differentiae specificae*). A simple example is: "a cactus is a plant (genus) which has thorns, and lives in the desert (two differentiae)." The differentiae could, alone, define the sense but the genus gives it context. The genus, in addition, serves to classify the topic word, while the attributes differentiate it from other members of that class. Between the two, genus and differentiae, ambiguity is avoided.

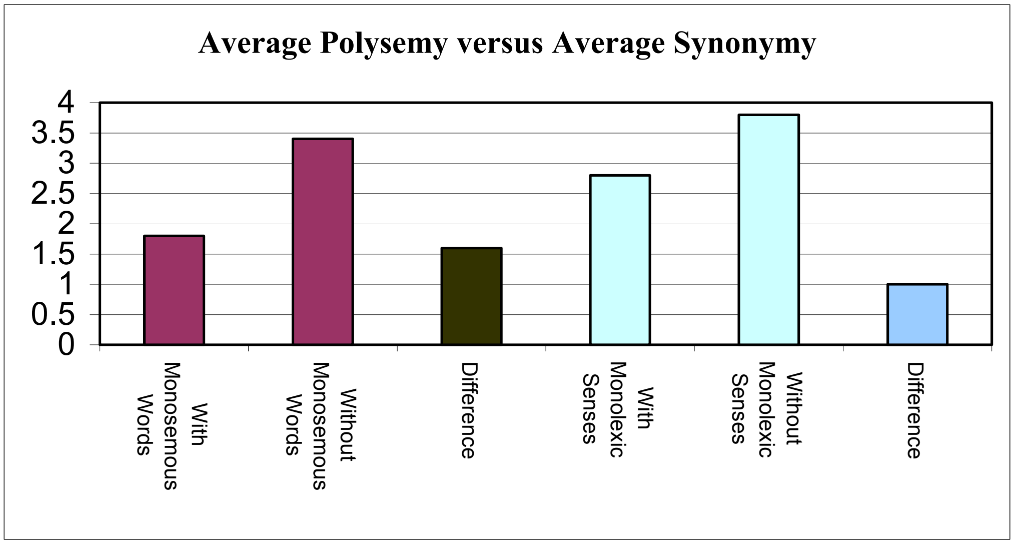


Figure 7. Effect on average polysemy of removing monosemous entries compared to the effect on average synonymy of removing monolexic entries

A polysemous word, such as *plant*, in a Synset can also act as a genus, giving the sense context. Identifying the genus and differentiae in a thesaurus Synset opens the door to understanding why *above* and *over* act as synonyms in one context but not in another—they can both substitute for "on top of" in one sense, but only "over" can substitute for "on the other side of" in another sense. This is discussed further under Section: *Visualization of Implicit Patterns* and Section: *Part-of-Speech Patterns.*

	Categories	**Paragraphs**	**Synsets**	**Entries**
Total	1,043	14,254	71,398	199,242
Avg. Per Category		14	68	191
Avg. Per Paragraph			5	14
Avg. Per Synset				3

Table 1b. **Frequencies for elements of the RIT Sense Index**

The numbers associated with the explicit structure of the RIT Sense Index are shown in Table 1b. On average a Category has about 191 entries. These 191 entries are divided among about 14 Paragraphs (that are, in turn, sub-classified by part-of-speech). Paragraphs average five sets of synonyms (Synsets, or senses) each; and each Synset contains about three Entries. In other words it takes, on average, about three words to define a sense.

Parts-of-speech are distributed through the Entries (or, conversely, it could be said that the Entries are distributed across the parts-of-speech) as in Table 2. The high number of nominal entries, as discussed earlier, is accounted for by the many lists of *things* and *parts*, most of which are concrete nouns. About half of the noun entries (48,000) are monosemous. Entries, not actual words, are counted here as most words have more than one part-of-speech. For example, “after” has five parts-of-speech and 13 senses, each identified as an Entry (see Table 3).

Part-of-Speech	POS Count
Noun	108418
Verb	40462
Adjective	39257
Adverb	9323
Preposition	503
Pronoun	12
Interjection	867
Phrase	437
Conjunction	131
Interrogative	13

***Table 2.* Frequency of Part-of-Speech, by RIT Entry**

It is possible, however, to count the number of *unique character-strings per part-of-speech*. This way, for example, the four prepositional senses of the word *after* are counted once under Prepositions, and the five adverbial senses of *after* are counted once under Adverbs. There are about 121,000 unique string-POS instances. The distribution is about the same as for Entries, but because of polysemy, much reduced in numbers for the “regular” parts-of-speech (Noun, Verb, Adjective and Adverb). Noun and adverb totals are reduced by about 30% and verb and adjective totals are reduced by about 50%. In other words, the various senses of a polysemous word tend to be of the same part-of-speech, especially when the word is an Adjective or verb.

Part-of-Speech	“after” count
Noun	1
Adjective	2
Adverb	5
Preposition	4
Conjunction	1

***Table 3.* Frequency of Part-of-Speech for the Word “after”**

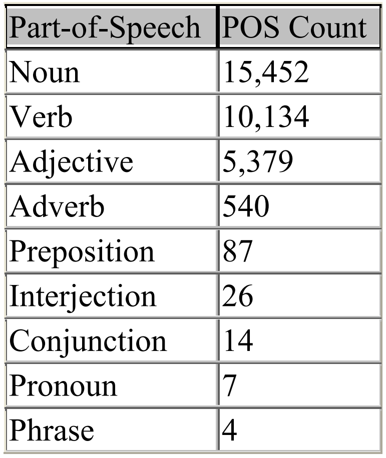
Words belonging to closed-set, low-polysemy parts-of-speech, are relatively unchanged by this calculation.

The longer a word is around (aside from those in closed sets), the more senses it tends to accumulate. We have already seen that the largest recent growth in Roget's Thesaurus, if not in the English language, is in concrete nouns—names of things—especially in the area of science and technology. These are recent additions and of low, or single, polysemy. A comparison with the distribution of Indo-European roots part-of-speech information is of interest at this point. The distribution of parts-of-speech in the Indo-European Roots Database is given in Table 4.

Part-of-Speech	IE Root Count
Verb	620
Noun	348
Adjective	131
Preposition	26
Pronoun	16
Verb or Adjective	3
Verb or Noun	3

***Table 4.* Frequencies of Part-of-Speech for Indo-European roots**

It is clear from Table 4 that verbs predominate among the Indo-European roots. This is quite different from the distribution of POS in RIT. The *roots* database includes 5,500 English base-words (*run*, but not *ran* or *running*; *apple*, but not *apples)* matched to the Indo-European roots from which they were derived. These base-words were compared with their corresponding RIT Entries and a part-of-speech distribution was calculated (see Table 5).



Part-of-Speech	POS Count
Noun	15,452
Verb	10,134
Adjective	5,379
Adverb	540
Preposition	87
Interjection	26
Conjunction	14
Pronoun	7
Phrase	4

Table 5. **Frequencies of Part-of-Speech for RIT Entries/base-words derived from Indo-European roots**

This ordering is the same as that derived from RIT rather than that derived from the roots database. One possible explanation is that the Indo-European roots that are nouns are more productive than those that are verbs. That is, they produce more words, and those words have more senses. An alternative conclusion is that many RIT nouns have as their origins Indo-European roots which were verbs. A selection from the database confirms the second explanation.

The words derived from an IE root often belong to a range of parts-of-speech, and even when the descendant belong to single parts-of-speech, their cognates (siblings) can belong to different parts-of-speech. By counting the descendants of each root by part-of-speech, then ordering the list of roots by the numbers of words of each part-of-speech that they produced, the roots can be viewed as being primarily noun-producers, or verb producers, and so on. All of the top twenty RIT-Entry-producing Indo European roots listed by this method were, themselves, verbs. Fourteen of those twenty roots produced, primarily, large numbers of Entries that were nouns—the other six roots produced verbs.

Some roots made it into the *top twenty* by producing both verbs and nouns. KAP- {grasp, hold}, at 6th and 7th positions, produced 111 verbs and 108 nouns. The KAP- verbs include {*accept, anticipate, captivate, capture, chase…*}; the nouns include {*acceptance, acceptability, anticipation, cable, capsule, captive, case, chassis…*}.

Summary

RIT is comparable to, but differs qualitatively and quantitatively from the 1911 Edition and from WordNet. The Thesaurus has expanded considerably in the number of words it contains since the original version, but the expansion has not been across the semantic board. Most additions have been nouns. The broadest notions in both WordNet and the two versions of Roget's, in terms of numbers of words, are common, everyday topics such as animals and plants; and these continue to be the largest sources of new words. The greatest rate of increase in new words, however, has been concentrated in the areas of science and technology.

The distribution of words by part-of-speech in RIT shows nouns to be by far the largest group. Curiously, though, the most frequent etymological source of RIT words, including the nouns, is Indo-European roots that are verbs. This information becomes more relevant in later discussions on hidden, or implicit, patterns in the Thesaurus.

Cross-Reference Patterns

As described in the discussion of the RIT cross-references in Chapter 4: *The Explicit Structure of Roget's Thesaurus*, there are five types of cross-references, termed here *category-, range-, multiple-, normal-* and *self-referencing* cross-references. These references, or links, are directed—they go in only one direction—from the source (referencing) to the destination (referenced) location.

There is also an implied relation back from the referenced location to the source. This is equivalent to the concept of Internet *back-links* (also called *reverse links* or *backward links*); and in citation analysis called a "citation"—as opposed to a "reference" (Small, 1978, p. 339), which corresponds to a regular RIT cross-reference or Internet hyperlink (a URL on a Web page).

It is easy, looking at a published document, to see what other papers that document references, but impossible to see what citations it has. Likewise, by looking at a Web page alone one cannot know what pages link to it. Search engines do, however, provide back-links on request; these show which Web pages link to a particular page (provided they are indexed by the search engine). The Google (Brin & Page, 1998) search engine uses the number, or count, of back-links that a Web page has as part of its measure of importance of the page on the Internet (Google, 2003, [PageRank Explained]). Using a database of thesaurus cross-references it is possible to identify the equivalent "cross-reference back-links."

A thesaurus cross-reference, such as 1.2.3 -> 515.5, from Category 1 to Category 515, could imply that there is a reciprocal relationship from Category 515 back to Category 1. However, as stated earlier, cross-references are directed arcs or links. While the count or number of links to a Category or concept may be significant in studying the importance of it, the semantics of the source and destination are not equivalent, so that cross-reference back-links are not a semantic relation. In support of this, the thesaurus editors supply return, or reciprocal cross-references in about only one third of the cases.

In citation analysis the relationship is also directed. Information about the importance of a document can be gained from citations, but the document owes nothing to the citing documents. On the other hand, the referencing document owes much to the cited document. What is owed is usually contained in the semantics of the text of the citing document Text related to the citation is called the “citation context” (Small 1978), and in Internet link terminology “anchor text.” The Google search engine takes advantage of this also, in classifying Web pages; the anchor text in hyperlinks on referencing pages is used for indexing the Web pages.

For all three situations, cross-reference, hyperlink/back-link, or citation/reference, the arc between locations has different implications depending on which direction the arc is followed. Also contributing to the asymmetry for cross-references may be the fact that they are specific-to-general; the source is always a specified Synset, and the destination is always at least a Paragraph (several Synsets), and often a whole Category. That is because cross-references are meant to lead the thesaurus user to a broader notion, not just another sense of the word adjacent to the cross-reference source—such information could be achieved simply by looking in the Word Index, at the back of the thesaurus.

Cross-references, taken by part-of-speech of the source Synset where the cross-reference is found in the RIT text, follow approximately the same distribution as words by part-of-speech. Within the sub-types (C-Category, Range, and so on), however, there are some noteworthy exceptions. All of the whole-Category links, except three, are from noun Synsets. All of the reciprocated, C-Category links are from Noun Synsets. This is probably because the hierarchy, or classification tree, is a noun tree. The Categories are all labeled by nouns, hence any references to them will be from Synsets where the label occurs as an Entry in the Synset, and the Synset will therefore most likely be a set of nouns.

Types of Cross-references; Reciprocating, Chains, Graphs and Cycles

Using the database version of RIT it is possible to calculate and classify reciprocal (symmetric, or reciprocating) cross-references between pairs of Categories or concepts (pairs for which cross-references go in both directions) —a tacit relation. This allows the five types of explicit thesaurus cross-references to be further sub-classified into those which are directed, and those for which there is also a reciprocal, and symmetric cross-reference of the same type. Table 6 lists the types, along with their frequency in the thesaurus. “C-Category” means “whole Category, Category reciprocated cross-reference.” “Xref” means “normal, regular, cross-reference.” “SelfRef” means “Self-referencing; from one part of a Category to another part of the same Category.” Self-referencing cross-references already reciprocate (are symmetric), so no new type is defined.

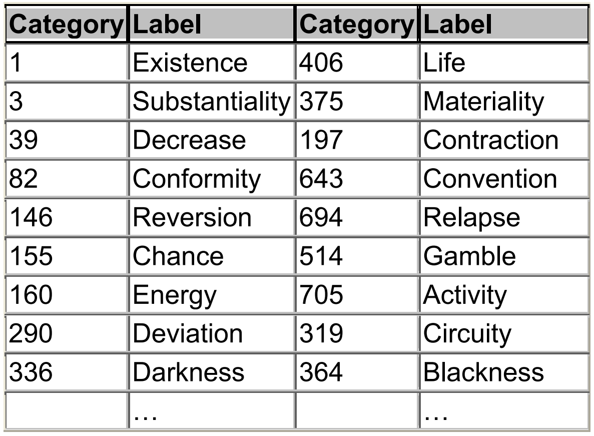
For brevity, from here on, types of cross-references will generally be referred by their abbreviated names, and cross-references may be called links. So a whole-Category-to-whole-Category reciprocated cross-reference will be referred to as a “C-Category link.” In addition, to reduce complexity, where possible only Category numbers and labels will be used for illustrative examples.

Cross-reference Type Code	Cross-reference Type	Abbreviated Meaning	Count
0	)—-(	C-Category	194
1	—-(	Category	382
2	}—-{	R-Range	29
3	—-{	Range	577
4	<===>	M-Multiple	380
5	===>	Multiple	924
6	<—->	X-Xref	216
7	—->	Xref	860
8	>—-<	SelfRef	57

Table 6. **Classification of cross-reference types in RIT**

Like the chains defined by Robert Bryan in his model of abstract thesauri, the cross-references are also graded (from strong to weak), and form chains. Chains of cross-references form networks within the thesaurus—the syndetic structure. The reciprocal cross-references provide evidence of semantically strong links between locations in the thesaurus, and form sub-networks within the larger networks. The notions or concepts found at each end of the links are practically the same, but are differentiated by context (both in daily usage and in the thesaurus).

The strongest links, by type, are the reciprocated whole-Category, C-Category cross-references, or links. An example is the link between 687: Therapy and 686: Healing Arts (**686.1.2-687.0** + **687.1.2-686.0**). These two Categories are found adjacent to each other in the thesaurus. Some examples of C-Category links that are found in quite separate parts of the thesaurus are listed in Table 7.



Category	Label	Category	Label
1	Existence	406	Life
3	Substantiality	375	Materiality
39	Decrease	197	Contraction
82	Conformity	643	Convention
146	Reversion	694	Relapse
155	Chance	514	Gamble
160	Energy	705	Activity
290	Deviation	319	Circuity
336	Darkness	364	Blackness
	…		…

Table 7. **C-Category links between Categories**

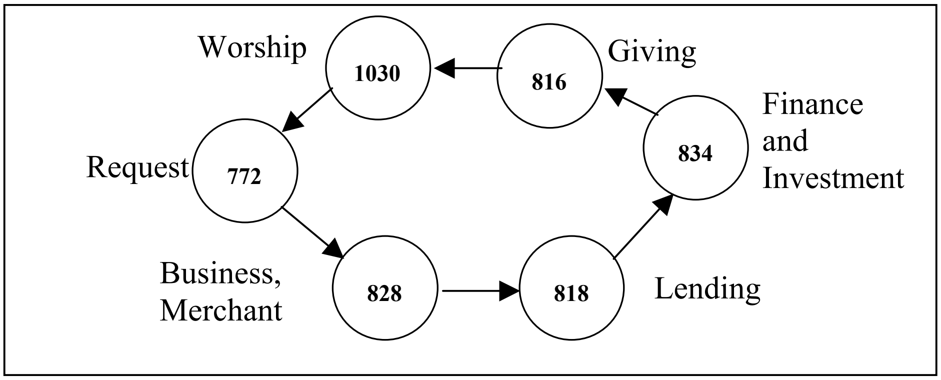
Links in the cross-reference network may form *cycles*, *chains* or *graphs*. Technically, chains, cycles, stars, and any other structures formed by linking locations (nodes), are all graphs, but they will be discriminated here by their form. The following graphs illustrate an example of each:

Chain (C-Category type):

949: Ill-humor)—(**867**: Discontent)—(**539**: Disappointment)—(**519**: Disillusionment

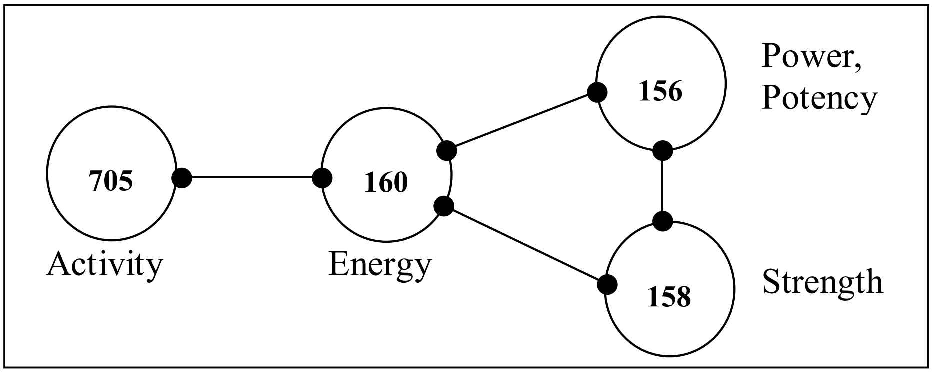
Such chains are not transitive and there is no C-Category link joining the non-adjacent Categories. In long chains the meaning transforms, or diverges away from the meaning of the beginning node. Chains may split or merge forming more complex structures.

Cycle (*normal* Xref type—but only the Categories are labeled here):



These are semantically weak links between Categories, but not meaningless. They represent narrow, specific ideas that may not be expected in a Category from the Category's label, or to be shared between particular Categories. Category 816:Giving, between 1030: Worship and 834: Finance and Investment, contains Entries such as *donation, contribution, tithing, subsidize, offering and oblation*, which clearly entails both finance and (at least one facet of) religious institutions.

Graph (C-Category type):



This graph has the attributes of an equivalence relation and behaves like a partition on Categories with C-Category links. This is analogous to the Type-10 chains on words and senses, identified by Bryan.

The examples given are relatively simple graphs with homogeneous link-types. Cross-reference graphs that are much more elaborate are more frequent, especially when link-types are unconstrained. For example, taking the following chain formed by C-Category links as a basis—*as a skeletal structure*:

562: Learning)—(**474**: Knowledge)—(**466**: Intelligence,Wisdom)—(**533**: Imagination

Two other non-C-Category cross-references run between Categories 474 and 562; a second link runs between 474 and 466; and both 474 and 466 link *externally* to Category 731: Skill in elaborate ways. Focusing on cross-references involving Categories 474 and 466, there are 17 unreciprocated cross-references to other Categories alone. Between the varying numbers of links-per-Category, and the varying levels of semantic strength (depending on link-type), such a sub-network resembles part of a small-world network—which it probably is.

The Categories related to the skeletal chain that were not identified here involve notions of *intellect, sanity, ignorance, discrimination, judgment, memory, foresight, information, learning, teacher, student, poetry, expedience*, and *dullness*. This broad, interwoven semantic field is implicit in the thesaurus; it is not found under one classification structure. It crosses class boundaries and can be identified only through automated methods.

A cross-reference is not only between a Synset and a remote Paragraph or Category. It is also between the hypernyms, or upper level nodes of the hierarchy above that Synset, and the hypernyms above the referenced Paragraph or Category. Most cross-references do not cross Class boundaries. That is, they usually reference Categories within the same Class. Those that do cross boundaries reflect strong relationships between the Classes.

Implications of Cross-references among Upper Levels of the Hierarchy

There are 33 C-Category links that cross Class boundaries. There are no C-Category cross-references between Classes 3: Physics and 5: Sensation (the Classes created and added to RIT by the American editors), and any other Class. That is, they have no connection to other Classes through C-Category links. The Class-crossing C-Category links are represented by the links in the graph in Figure 8a, labeled by the number of links that occur. The relationship is strongest between the Intellect and the Affections Classes.

An example of the nature of the relationships between Classes shown in Figure 8a can be seen analyzed in Figure 8b—which is part of one of the example C-Category chains given earlier.

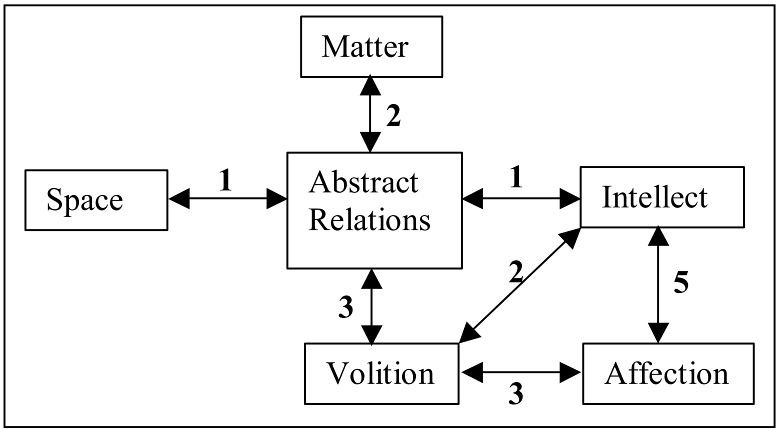


Figure 8a. **Counts of C-Category links crossing Class boundaries**

The example in Figure 8bsuggests that, despite the fact that their semantics and words overlap (as evidenced by the strong C-Category link between the two Categories), a qualitative division exists between these almost-equivalent concepts found categorized under the Affections and Intellect Classes. A second example, more elaborated, is given in Figure 8c to support this observation.

Class level:
7: Affections C6: Intellect
Category Level:
867: Discontent)—(539: Disappointment

Figure 8b. **C-Category link at the Class level**

C-Category links between the Classes of Affections and Intellect, such as (for a further example) [921: Unsociability)—(611: Uncommunicativeness], suggest that whether a concept has social-emotional connotations, or is purely intellectual (at least to the observer) affects the semantics of practically identical concepts, and consequently, the way in which the concepts are classified. In this way Hope can be seen as the emotional equivalent of Expectation, an emotionally neutral intellectual notion; and Care the intellectual equivalent of Caution. Likewise science (an intellectual pursuit) does "541: Prediction," but when non-scientists make claims about the future they are said to {*foretell, augur, divine, prophesy, forecast*…}—and it is called "1032: Occultism."

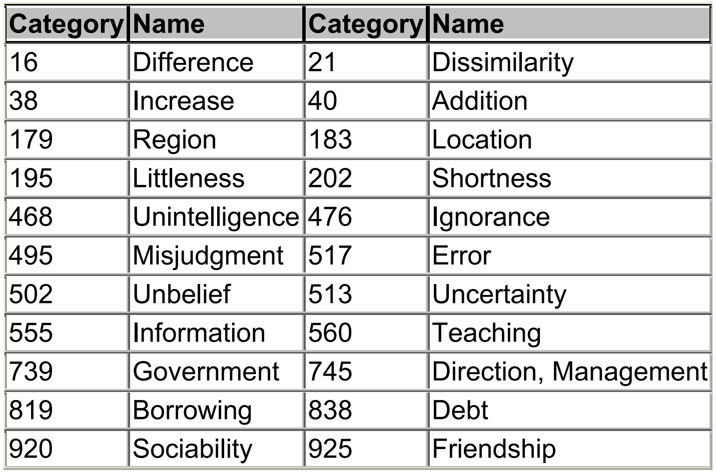
Similar analysis can be made of the second level, or Roman Classes, of the hierarchy; and the third-level, or Letter Classes. Examples of strong C-Category links that cross only Letter Class boundaries (both derive from the same top level Class and Roman Classes) are given in Table 8.

At this low level of the hierarchy Categories are related only by the fact that they share the very broad notion of their Roman Classes—they represent dimensions of the Roman Class notion. For example, Categories 16: Difference and 21: Dissimilarity share only Roman Class II: Relation. They are discriminated by their Letter Classes, A: Absolute Relation and B: Partial Relation.

Class level:		
7: Affections		**6: Intellect**
Roman Class Level:		
I. Personal Affections		**II. States of Mind**
Letter Class Level:		
D. Contemplative Emotions		**D. Anticipation**
Category Level:		
886: Hope	)—(	**537: Expectation**

***Figure 8c.* C-Category link shown at all Class levels**

The relationship unearthed by C-Category links shows that distant Categories can bear a close, possibly redundant, semantic relationship. This is not a criticism of Roget's hierarchy (although the hierarchy may deserve criticism) as semantics is multi-faceted and multi-dimensional and it should be expected that not all facets of meaning shared between two notions could be represented by a single relation, or even a single structure. The words classified under a Category in one Class (or facet) will be different from the words classified under a Category in a different Class (or facet), even though the notions which the Categories represent may seem the same. Category 537: Expectation contains 147 Entries and Category 886: Hope contains 154 Entries—but they share only 10 words.



Category	Name	Category	Name
16	Difference	21	Dissimilarity
38	Increase	40	Addition
179	Region	183	Location
195	Littleness	202	Shortness
468	Unintelligence	476	Ignorance
495	Misjudgment	517	Error
502	Unbelief	513	Uncertainty
555	Information	560	Teaching
739	Government	745	Direction, Management
819	Borrowing	838	Debt
920	Sociability	925	Friendship

***Table 8.* C-Category links that cross Letter Class boundaries only**

Set Implication

The mathematical concept of "set implication" suggests that subsets imply the superset. In general, a word in RIT that has a set of senses that is a subset of senses of a second word, by set implication, implies the second word. In Table 9a, the words on the right have fewer senses than the words on the left, and those senses are a subset of the senses where the words on the left are found in RIT. So the words on the right imply or infer the words on the left.

SuperSet	**SubSet**
3-D	stereoscopic
abandoned	deserted
about	circa
allow	deem
allowance	stipend
bloody	gory
blunt	take the edge off
blush	turn red
brief	short and sweet
calm	tranquil
caustic	escharotic
…	

Table 9a. **Set implication between RIT words. Subset implies Superset (right to left)**

The words on the right are more rare (have fewer senses) and are more specific, and the words on the left are more polysemous (have more senses) and are more general. Implications of this sort can form chains: *poodle* and *terrier* both imply *dog*; d*og* and *cat* in turn imply *animal*; animal implies *living thing,* and so on. In this way Synsets, being subsets, can be seen to imply Paragraphs, which in turn imply Categories; and so on up the hierarchy. A cross-reference carries with it these implications. Set implication associated with cross-reference is illustrated schematically in Diagram 2.

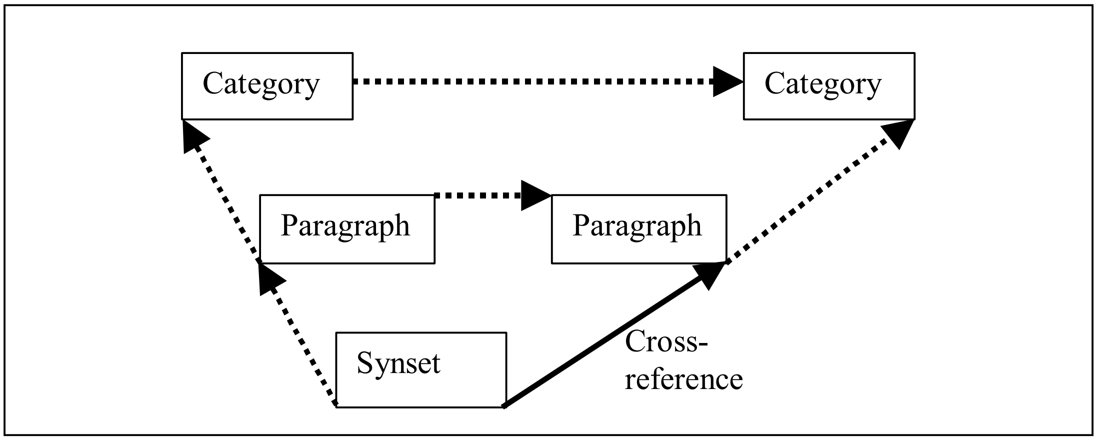


***Diagram 2*. Implications (dotted lines) in cross-reference (solid line)**

The source Synset of a cross-reference is always a smaller set of concepts than the destination of the cross-reference, but it does not always contain a string (a word) that is contained in the destination set. Of the whole-Category links, those that contain an identical string in both the source and destination provide semantic evidence of set implication in cross-references—the source Category concept implies the destination Category concept (there is an inference from the source to the destination). Table 9b illustrates this (the source Category concepts are on the left).

Source	Source Name	Destination	Destination Name
27	Disagreement	793	Disaccord
30	Equality	14	Identity
34	Greatness	194	Size
38	Increase	196	Growth
82	Conformity	643	Convention
119	Past	123	Oldness
140	Permanence	112	Perpetuity
143	Continuance	110	Durability
168	Reproduction	22	Imitation
179	Region	183	Location
197	Contraction	39	Decrease
262	Furrow	395	Channel

Source	Source Name	Destination	Destination Name
489	Measurement	29	Degree
491	Discrimination	894	Fastidiousness
513	Uncertainty	502	Unbelief
529	Inattention	532	Neglect
538	Inexpectation	918	Wonder
539	Disappointment	867	Discontent
553	Manifestation …	512	Certainty …

Table 9b. **Set implication in cross-references**

The source Category notion implies the destination Category notion, but not vice versa. This is a similar pattern to that found in the description in Chapter 2 of metaphor and analogy. "The ship ploughed the sea," is a metaphor drawn between a plough ploughing the land, and the behavior of a ship on the sea. The direction is from the source of the metaphor to the situation at hand, transferring the semantics of one situation to the other through the similarity in the behavior—in this case the mapping of the relation between the *actor* and the *agent*. It is unlikely that the converse, "the plough sailed the land," would be accepted or understood as a metaphor, because there is no semantic element in "sailing" that is shared between the two situations. "The driver lost control and the car ploughed through the crowd" is also directional—from plough to car. It is possible to say, "drive the plough," but it would not generally be considered metaphoric.

Most Categories participating as sources and destinations of cross-references share the anchor term. Examples of those that do not are Category 137: Regular Recurrence, anchored on *holy days* referencing a Paragraph in Category1038: Religious Rites, that contains a list of holy days; Category 123: Oldness, anchored on *ancient manuscripts* referencing a Paragraph in Category 600: Writing, that contains a list of important ancient manuscripts; and Category 161: Violence, anchored on *windstorm* and referencing a Paragraph in Category 402:Wind, that contains such terms as *sand spout, dust-devil, cyclone* and *hurricane.* Although there is no shared anchor term and no inference between the Categories, there is still method in these cross-references.

Fan-in and Fan-out; Semantic Hubs and Authorities

Except for the cycle, the examples given earlier have all been semantically strong cross-references. The majority of cross-references, however, are of the weaker types—from a Synset, to one or two paragraphs, unreciprocated by a link of the same type. Source Categories have as many as 27 outgoing links, of all or any types. Destination Categories have as many as 33 links directed to them. Using electrical circuit terminology these cross-reference counts are referred to as *fan-in* and *fan-out*[19] (*fan*, because connections with many links look like a fan when drawn on a circuit board diagram).

Categories with high fan-in or fan-out are analogous to the hubs and authorities (Kleinberg, 1999) identified in studies of the distribution and density of hyperlinks to and from Internet Web pages. Those Categories with a high fan-in are like semantic authorities, referred to by other Categories; those Categories with a high fan-out are like the hubs, referring to assorted Categories across the thesaurus for *semantic-authority*. Like Web pages, not all Categories have links, and some are never referenced; and Categories may participate in both sets.

[19] Also known (discrete mathematics, and computer science) by the terms: in-degree and out-degree.

Tables 10a and 10b show the top twenty Categories by fan-in and fan-out, or cross-reference count. The top couple of Categories by cross-reference count are intellectual in nature, but on the whole the Categories represent negative emotional notions such as *sadness, falseness, deception, uncertainty, displeasure* (existing in both sets), and other notions with negative connotations such as *disease* and *weakness*. Previous research (Old, 2000) on semantic densities (areas highly interconnected by Type-10 chains) within RIT hypothesized that those concepts have to do with the fight-flight response, and survival. These most highly connected Categories (by cross-reference) may reflect the same phenomenon.

The common themes for authority Categories in RIT are (considering all Categories, and totaling cross-references at the Roman Class level—totals included):

- 6:I: *Intellectual faculties and processes*, 349 links;
- 8:I: *Personal affections*, 348 links;
- 6:III: C*ommunication of ideas*, 281 links;
- 7:I: *Volition in general*, 231 links;
- 2:IV: M*otion,* 213 links.

The same Roman Classes appear for the hub Categories, except that V*olition in general* disappears; and the order of I*ntellectual faculties and processes*, and C*ommunication of ideas*, is reversed. 6:I: *Intellectual faculties and processes* includes Categories 513: Uncertainty, 472: Insanity, Mania, and 469: Foolishness among its top authorities; and 6:III: C*ommunication of ideas* includes 614: Falseness and 616: Deception.

Cat#	Label	Fan In Count	Cat#	Label	Fan In Count
466	Intelligence, Wisdom	33	864	Displeasure	19
474	Knowledge	30	159	Weakness	18
870	Sadness	26	469	Foolishness	18
512	Certainty	26	855	Excitement	17
542	Foreboding	24	532	Neglect	17
614	Falseness	21	227	Covering	16
646	Motivation, Inducement	21	112	Perpetuity	16
616	Deception	21	336	Darkness	15
513	Uncertainty	21	907	Vanity	15
472	Insanity, Mania	21	967	Disapprobation	15

***Table 10a.* The top 20 (of 821) destination Categories by fan-in. Authority-like nodes.**

Cat#	Label	Fan Out Count	Cat#	Label	Fan Out Count
572	Art	27	418	Sex	19
562	Learning	25	537	Expectation	18
1002	Lawsuit	23	697	Protection	18
973	Improbity	22	876	Amusement	18
614	Falseness	21	270	Transference	18
684	Disease	20	635	Choice	18
514	Gamble	19	870	Sadness	17
688	Psychology, Psychotherapy	19	616	Deception	17
541	Prediction	19	864	Displeasure	17
680	Uncleanness	19	540	Foresight	16

***Table 10b*. The top 20 (of 782) source Categories by fan-out. Hub-like nodes.**

Almost all of the semantically strong cross-references and most (75%) of the unreciprocated cross-references between cross-reference hubs and authorities occur within the Roman level Classes. In other words, a strongly connected hub and authority pair will usually occur within a single Roman Class. This suggests strong coherence within Roman Classes. For example, there are reciprocal, R-Range links between

Categories 466: Intelligence, Wisdom (Paragraph 5: wisdom) and 474: Knowledge (Paragraph 5: scholarship); both from Roman Class 6:I: *Intellectual faculties and processes*. The shared Entry is “erudition.” This link can be summarized as:

6:I-466 Intelligence, Wisdom {wisdom < erudition > scholarship} Knowledge 474-6:I

Table 10 lists examples of the links running from hubs to authorities that cross Roman Class boundaries; it illustrates the type of coherence that exists between the otherwise disjoint Roman Classes.

RC1	Cat1	Category Name1	ParaName1 <	**entry**	> ParaName2	Category Name2	Cat2	RC2
2.IV	270	Transference	carrier	letter carrier	postman	Messenger	559	6.III
2.IV	284	Propulsion	shooter	Nimrod	hunter	Pursuit	653	7.I
2.IV	308	Ejection	get rid of	throw away	discard	Disuse	666	7.I
2.IV	323	Agitation	agitated	excited	excited	Excitement	855	8.I
6.I	469	Foolishness	absurd	ludicrous	humorous	Humorousness	878	8.I
6.I	495	Misjudgment	misjudge	misconstrue	misinterpret	Misinterpretation	551	6.III
6.I	513	Uncertainty	bewildering	enigmatical	enigmatic	Unintelligibility	547	6.III
6.I	513	Uncertainty	hang in doubt	falter	hesitate	Irresolution	625	7.I
6.III	544	Latency	latency	dormancy	inertness	Quiescence	267	2.IV
6.III	544	Latency	latent	dormant	inert	Quiescence	267	2.IV
6.III	547	Unintelligibility	enigmatic	puzzling	bewildering	Uncertainty	513	6.I
7.I	624	Obstinacy	obstinacy	opinionatedness	dogmatism	Certainty	512	6.I
7.I	629	Avoidance	dodge	shrink	pull back	Reaction	283	2.IV
7.I	633	Eagerness	overzealous	fanatical	fanatic	Insanity, Mania	472	6.I
7.I	648	Allurement	lure	ensnare	trap	Deception	616	6.III
7.I	655	Way	passageway	inlet	place for entering	Ingress, Entrance	301	2.IV
8.I	860	Impatience	impatient	impetuous	impulsive	Impulsiveness	628	7.I
8.I	860	Impatience	impatient	anxious	eager	Eagerness	633	7.I
8.I	876	Amusement	sportsman	racer	speeder	Velocity	268	2.IV
8.I	881	Dullness	triteness	clichee	platitude	Maxim	516	6.I
8.I	902	Ostentation	pompously	bombastically	grandiloquently	Grandiloquence	599	6.III

***Table 10c*. Links running from hubs to authorities crossing Roman Class boundaries.**

The semantic connections between the distant clusters are clearly reasonable and could bring into question the reasonableness of the locations chosen for the Categories and Classes that participate in the clusters, in the classification system. However the majority of semantically strong links exist within the Roman Classes; this sample is representative of only about 15% of the total cross-references between the core hubs and authorities; the other 85% are internal to their Roman Classes. This sample probably illustrates John Lewis Roget's assertion that:

> Many words, originally employed to express simple conceptions, are found to be capable, with perhaps a very slight modification of meaning, of being applied in many varied associations. Connecting links, thus formed, induce an approach between the categories; and a danger arises that the outlines of the classification may, by their means, become confused and eventually merged (Roget, J. L, 1879, p. ix).

Furthermore, the relations in this sample often represent implications, cause and effect, or general-to-specific instances, rather than equivalence. For example, *impatience* is an internal state, while *impulsiveness* is observable, and it could be said that the first leads to the second. Also, these are further examples of the *multi-facetedness* of semantics discussed earlier—that similar notions are not identical notions. The context (such as emotional or intellectual context) often demands a whole different vocabulary, and justifies the apparent redundancy of some Categories in different parts of the classification hierarchy.

Cross-reference Coupling and Co-referencing

In citation analysis, any sharing of references is known as 'bibliographic coupling.' For the complementary comparison, that between citations of two documents, any sharing of documents is called 'co-citation.' (Egghe & Rousseau, 1990, cited in Wouters, 1998; Garfield, 2001, p.3). Using this in an analogy with cross-references, if Category A1 and A2 contain cross-references to Category B, we could call this "cross-reference coupling" and say that A1 and A2 are *coupled in* B. If A2 also contains a cross-reference to Category C, we could call this "co-referencing" and say that B and C are *co-referenced by* A2. Finally, if a cross-reference linking A2 to B is reciprocated by a cross-reference from B to A2 we could call this a *reciprocal* cross-reference.

Coupling

By selecting cross-references that are coupled, that is, they both contain cross-references to a common third Category, we can view some of the *implicit* semantics contained in cross-references. Two illustrative examples follow:

Example 1:

Source A
Category 765 Disobedience, Paragraph 5: Rebel
Source B
Category 738: Lawlessness, Paragraph 3: Anarchist

are *coupled in* (both contain cross-references to)

Destination
147: Revolution, Paragraph 3: Revolutionist

Rebel and Anarchist are Paragraph labels. They are also members of the Synsets containing the two cross-reference Sources. The concepts they represent have implications for the Paragraph they reference—they are *facets* of the cross-referenced concept. Cross-references are directed so there is no backward implication to the Sources. Consequently no transitive implication exists between the Sources even though the two Sources share some semantics with the referenced, Destination concept.

Example 2:

Source A	Source Concept A	Source B	Source Concept B	Destination	Coupled In
256:14:8	Concavity: indent	566:18:11	Indication: mark	261:4	Notch

An *indent* (source concept A) is a feature, while a *mark* (source concept B) is a symbol (in this context). A *notch* (the destination concept) may be either a feature, *or* used as a symbol. In other words, the source concepts are quite different facets of a *notch*, as opposed to being equivalent concepts.

Further examples are listed in Table 11. These have been filtered as there are 1,574 coupling cross-references in RIT. The words following the colons in the sources are Paragraph labels of Synsets from which the cross-references originate. In these examples the Paragraph labels of the referenced locations, or concepts, have been omitted as there are often several Paragraphs referenced (that is, in order to simplify the representation).

Although the coupled cross-references do not have implications for each other, together they certainly have implications for the concept they are coupled in. Even at the more-abstract Category-only level it can be seen (see Table 12. The Paragraph labels are omitted in these following examples).

Source A	Source Concept A	Source B	Source Concept B	Destination	Coupled In
343:33:27	Radio: Receiver Units	344:17:8	Television: Receiver Units	342:18	Electronics
431:6:6	Sourness: Sour	432:6:5	Pungency: Pungent	428:6	Unsavoriness
491:2:2	Discrimination: Discernment	540:1:9	Foresight: Foresight	466:5	Intelligence, Wisdom
2:8:6	Nonexistence: Unreal	533:19:5	Imagination: Imaginary	518:9	Illusion
1032:10:1	Occultism: Divination	1033:1:15	Sorcery: Sorcery	541:15	Prediction
520:8:7	Assent: Assent	775:10:2	Permission: Permit	773:3	Consent
903:9:1	Pride: Vain	908:9:2	Boasting: Boastful	907:10	Vanity

Table 11. Cross-reference coupling.

Source A	Source Label A	Source B	Source Label B	Destination	Coupled In
1002	Lawsuit	945	Forgiveness	1005	Acquittal
1002	Lawsuit	1006	Condemnation	1007	Penalty
337	Shade	613	Concealment	227	Covering
612	Secrecy	921	Unsociability	611	Uncommunicativeness
510	Probability	481	Reasoning	500	Belief
640	Custom, Habit	82	Conformity	643	Convention
140	Permanence	1	Existence	110	Durability
255	Convexity	357	Elasticity	196	Expansion, Growth
684	Disease	690	Impairment	204	Narrowness, Thinness

Table 12. Cross-reference coupling at the Category level.

Such implications as illustrated in Tables 11 and 12 are implicit in the Thesaurus. They cannot be identified from the text alone. Diagram 3 illustrates this schematically. The arrows with dotted lines represent implications.

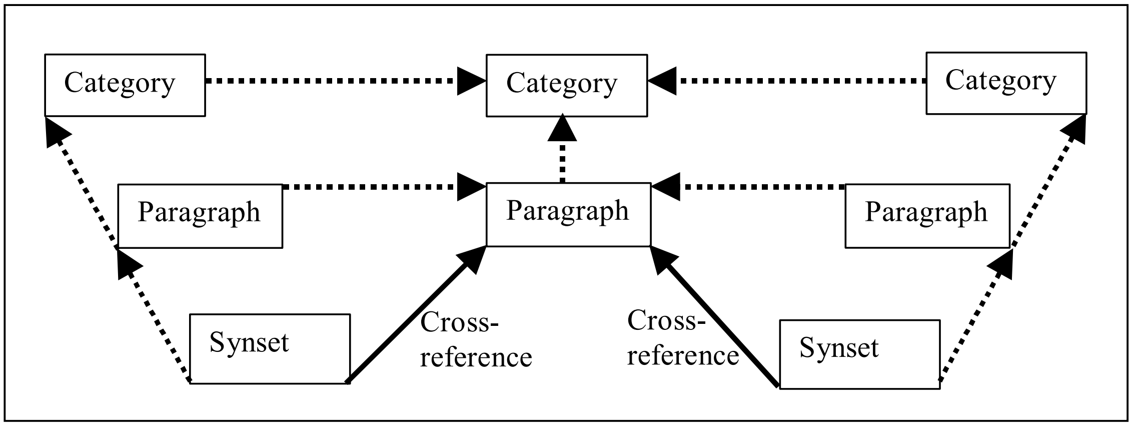
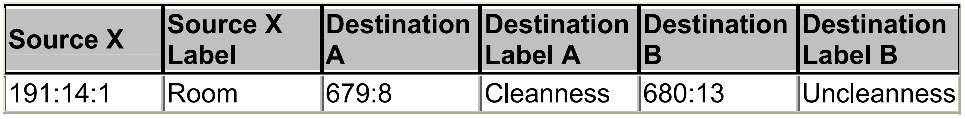


***Diagram 3*. Implications (dotted lines) in cross-reference coupling (solid lines).**

Co-reference

The complementary cross-reference type to *co-reference coupling*, *co-referencing cross-reference*, is simple. Any Synset with two or more cross-references is a candidate for the source, creating a co-reference relationship between its two or more cross-references. For example, Synset **191:14:1** found in Category 191: Room, contains five cross-references: {kitchen 329.3, storeroom 658.6, **lavatory 679.8, water closet 680.13**, smoking room 433.14}

Selecting the two highlighted cross-references, it can be stated that, "Paragraphs 679.8 and 680.13 are *co-referenced by* 191.14.1."



Source X	Source X Label	Destination A	Destination Label A	Destination B	Destination Label B
191:14:1	Room	679:8	Cleanness	680:13	Uncleanness

Cleanness and *uncleanness* do have something in common, in that they are antonyms. Other co-references show less in common. The Entries in 191:14:1 (*kitchen, storeroom* etc.) are list-like, rather than a set of synonyms. The only thing they have in common is that they are rooms. The cross-references take them out of the *room* domain to unrelated domains. Other examples (briefly) are: Presence and Hearing are co-referenced by Drama; Knowledge and Activity are co-referenced by Learning; and Mankind and Structure are co-referenced by Zoology. The only thing semantically in common between co-referenced concepts is the source concept by which they are co-referenced. There is no inference between them.

There are only nine co-reference source Synsets in RIT, most of them lists; and 191:14:1 is the largest. The semantic overlap between the Destination Paragraphs co-referenced in the source Synsets is weak, and co-reference is altogether a weak relation in RIT. Diagram 4 illustrates the implications schematically.

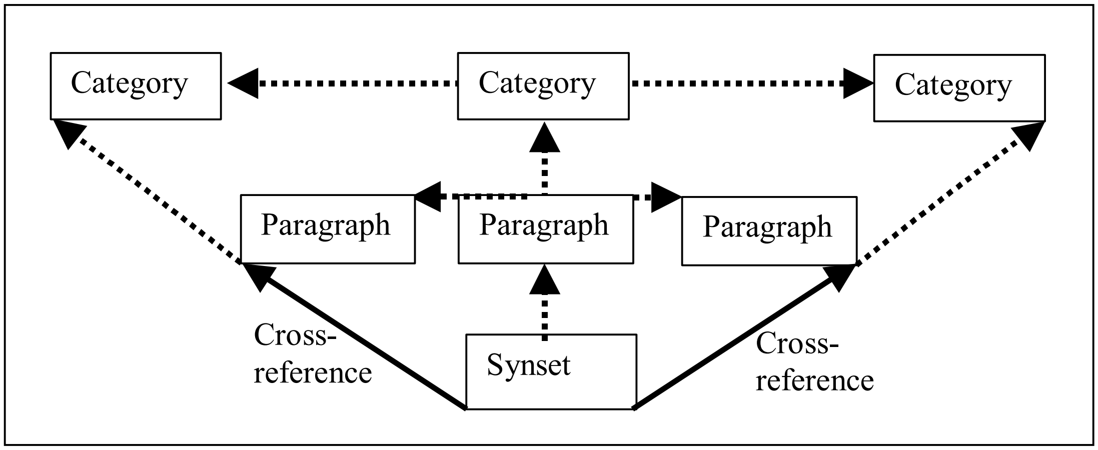


***Diagram 4*. Implications (dotted lines) in co-referencing (solid lines)**

Summary

Cross-references form an elaborate network of links throughout the thesaurus. Latent semantic information can be extracted from the cross-references by classifying them (analogously to Bryan's chain-classification of word-sense links), then selecting relationships among the different types of cross-references; by calculating the density of cross-references at specific levels of the hierarchy; and by studying the semantics shared by disparate locations in the thesaurus linked by cross-references. The links range from semantically strong C-Category reciprocated cross-references between whole Categories in completely different Classes, to weak self-referencing links that reference locations within their own source Categories.

Sub-structures formed by sets of cross-references are analogous to network structures found by other researchers focusing on Entries and Synsets, such as among Talburt and Mooney's (1990), and Jacuzzi's (1991) components; and to other connected data.

Citations, hyperlinks and cross-references, unlike other forms of RIT connectivity, are all directed links. Cross-reference link densities are similar to those found among the hyperlinks of Web pages, suggesting the same hub-like and authority-like connectivity; and to citations, allowing the transfer of citation analysis techniques in the search for latent, or implicit semantic information.

Although it has not been tested to date, cross-references form what is probably a small-world network. There is strong coherence within Classes at the top, Roman, and Letter levels—the majority of the cross-references, by source and destination, fall within the bounds of the same class. There is also a significant minority of cross-references crossing boundaries. Strogatz (2001) points out that a small-world network is somewhere between networks of random connections (with isolated fragments, or components) and regular networks (up to fully connected). The latter may be highly clustered, but with long paths required to across the clusters. These are analogous to Classes and Categories. By adding random links [“the slightest bit of rewiring” (Strogatz, 2001, p.273)] to models of networks of this type, they soon transform into a small-world network. The added paths act like short-circuits cross-linking clusters and parts of clusters, facilitating short paths across and between them. These random links are analogous to the cross-references that cross Class boundaries.

Frequency counts of cross-references also follow power curve distributions—another characteristic of small-world networks.

Part-of-speech Patterns

As stated in the Section: *Descriptive Statistics*, many words have multiple parts-of-speech (POS). This provides a type of link between nouns, verbs, adjectives, adverbs and prepositions through the words they share in common. Nouns are the largest group (about 70,000 words) while prepositions are the smallest group (about 400 words). Exploring all of the relationships between POS would be a book in itself, so the following discussion, while acknowledging other parts-of-speech, focuses mainly on the relatively small, closed set of prepositions. The semantic overlap between POS is explored initially, then, using the sense definitions of Indo-European roots of prepositions and words listed as prepositions in Roget's Thesaurus this Section discusses the semantic origins of English prepositions. Finally, a model of prepositional semantics that accounts for the POS overlap between prepositions and other POS is proposed.

Overlap among Parts-of-speech

Prepositions, as a class of words, have been referred to as a closed set. The "set" is the set of words that are eligible to be called prepositions. It is closed probably as a consequence of the fact that the words defined as (or classed as) prepositions describe a limited set of concepts (for example spatial and temporal relations) that do not change—unless our consensual reality changes.

Prepositions are not, on the other hand, a stable set. The semantics of individual prepositions is mutable across time, and among related languages. Non-standard or idiomatic use of prepositions can become the standard, while the "correct" or traditional usage goes out of fashion. Or not… An educated Scot uses the word *outwith* (archaic to some[20]) where the average English speaker would instead use *outside of,* or *except*. *Outwith* is a perfectly good preposition and unambiguous to its users.

While an English speaker standing before a house might say that the rear garden is *beyond, to the back of* or *behind* the house, but never *after* the house; a Dutch speaker would say it is "*achter* het huis." *Achter* means "after." It has the same Indo-European language root as *after,* and has the same basic semantics[21] in both languages. Even though Dutch and English are about as close as any two languages can be without being dialects, this preposition has evolved to be used in different ways.

Words that are prepositions do not have a clear semantics even within the same language. Where a teacher speaking American English, referring to a poorly written essay, might tell a student to "do it *over,*" a British teacher would only ever say, "do it *again*."

Even prepositions commonly considered synonyms may vary or disagree in the senses they describe. *Above* can be a synonym of *over* in the sense of "higher up," but not in the sense of "across"—one may live *across* the road or *over* the road, but not *above* the road (and still mean the same thing).

[20] From Middle English according to Webster's 3rd Edition, 1965. Though it is not in Roget's Thesaurus, *outwith the law* (illegal) is.

[21] It is still acceptable English to say, "Take the first turn right *after* the set of lights;" or "After you :)"—but we are more likely to use it in its analogous temporal form: "… after 10 o'clock;" or "… after I get up."

Furthermore, there is considerable overlap between the set of words that are called prepositions and words from other word classes (parts-of-speech). Crystal (1989, p. 92) points out that word classes

> … are not as nearly homogeneous as the theory implies. Each class has a core of words that behave identically, from a grammatical point of view. But at the "edges" of a class are the more irregular words, some of which may behave like words from other classes.

This Section offers no solutions to this apparent confusion, but attempts to illustrate it as a natural, and even desirable, feature of prepositions—and of language generally.

The prepositions used here are drawn from the 411 entries found in RIT, which groups words of similar meaning together by part-of-speech (Synsets are grouped in Paragraphs and Paragraphs are grouped by POS). WordNet (Miller et al., 1993) does the same, and contains a richer set of relations, but does not contain prepositions. The comparisons made here between prepositions and other parts-of-speech are limited to nouns, verbs, adjectives and adverbs.

Prepositions are a small set compared to other parts-of-speech. While prepositions are a closed set, nouns are ever-increasing as science and technology advance and new words are needed to describe new concepts. Other parts-of-speech are being added to as well (for example, to be "ENRONed"), although not as rapidly. Table 13 shows the word-count-by-part-of-speech for words in RIT.

The other data sources used in this study have different numbers, but the distribution is about the same. In Table 13 words are counted only once per part-of-speech. The word *line*, for example, is found as a noun Entry in 31 different thesaurus senses but is counted here only once as a noun. The preposition a*fter* is just one of the 411 prepositions counted here. It is also counted once under each of the other parts-of-speech as its 13 senses, or entries, are spread across all five parts-of-speech.[22] The difference between entries and words is that an entry represents one sense-instance of a word, while word is a particular string of characters. So *after*, with 13 senses, is represented in RIT by 13 entries. Four of those senses are prepositional, so the word *after* has four prepositional entries.

Part-of-Speech	POS Count	Percent
Noun	69017	57.4%
Adjective	23171	19.3%
Verb	21368	17.8%
Adverb	6346	5.3%
Preposition	411	0.3%

Table 13. **Part-of-speech count of words in Roget's Thesaurus**

Approximately half of the prepositions found in Roget's Thesaurus have more than one sense and so are polysemous. Many of those words are classified elsewhere in the thesaurus under different parts-of-speech. In Figure 9 the percentage of overlap among parts-of-speech has been illustrated graphically using pie charts. In this case entries were chosen, rather than words.

Entries found classified under only one part-of-speech are ignored here, as they do not contribute to the analysis of overlap between parts-of-speech, and also because the more than 105,000 unique entries in this category (of the total 200,000 thesaurus entries) would make

[22] *After* is found as a synonym of *afternoon* and *evening* in one nominal thesaurus sense.

the overlapping entries for smaller parts-of-speech, invisible. So *betwixt*, for example, which occurs only as a preposition, is ignored. *After*, which occurs in all five parts-of-speech, is included in the calculations for all five pie charts.

The arrows serve as a rough indicator of the main allegiance owed by a part-of-speech to another part-of-speech. For example, verbs and nouns share a high percentage of words (77% and 87% respectively), indicated by a thick, double-headed arrow; 47% of prepositions are also adverbs (indicated by a thick arrow) and 32% are also adjectives (indicated by a narrower double-lined arrow); and 57% of adverbs are also adjectives (indicated by a thick arrow). The relative proportions shown here are not normalized numbers for each part-of-speech (for example there are many more nouns and verbs than prepositions), but a clear indication, at least, is present in the illustration.

Note that among the different parts-of-speech only adverbs (that is, 8% of adverbs that occur in other parts-of-speech) are also found as prepositions in any significant numbers. Those same entries constitute the 48% of entries represented on the Prepositions pie chart labeled "4 48%" (in white).

In real numbers, 287 words classified as prepositions in Roget's Thesaurus are also found in senses other than those classed as prepositional. For example, 33 of these words also occur as nouns[23]. Table 14 shows the actual overlap in terms of word-counts (including conjunctions).

[23] An *over* (Nn) is a cricket term for a period of play—but that sense is not included in this American edition of Roget's Thesaurus. Examples of verbs are "to *further* a cause" and "to *near* a conclusion."

These overlaps are formed with 198 of the 411 prepositions. There are a further 213 prepositions that do not overlap with any other part-of-speech.

Part-of-speech	Overlap
Adverb	137
Adjective	87
Noun	33
Conjunction	*18*
Verb	12

Table 14. **Number of words shared between prepositions and other parts-of-speech.**

Part-of-speech Overlap for the Preposition Over

In Figure 10 the overlap between prepositions that occur as synonyms of *over* in various senses with various parts-of-speech can be seen represented as a *Concept Lattice* (Wille, 1982). This forms a kind of topology of *over,* its senses, and the word that are found accompanying it in those senses—its synonyms. The lattice includes only *shared* synonyms of *over*—those words that occur with *over* in more than one sense. As with Figure 9, the words that have been omitted occur in only one part-of-speech and do not contribute to the connectivity or overlap between parts-of-speech, or senses. They would however differentiate or discriminate senses which otherwise contain identical sets of words. This is discussed further under the Section, *Genus and Differentiae*, below.

A concept lattice is generated automatically from a relation between two sets, *objects* and *attributes*. In this example the *objects* are words from Roget's Thesaurus while their *attributes* are the senses of the words. A polysemous word can occur in more than one sense (as several entries) and a sense can contain more than one word—hence the graph structure formed is a lattice, not a tree. The nodes/circles are called *formal concepts* and are labeled above by the index numbers of the senses and below by words found in those senses. Index numbers are of the form:

Category#:Paragraph#:Sense.

Although a formal concept is defined as the set of all of its attributes (words) and all of its objects (senses), for economy of representation words and sense index numbers are used as labels only once. Words label the lowest formal concept in which they occur and index numbers label the highest formal concept in which they occur. Thus a lattice is a partial ordering, where formal concepts higher in the lattice structure are labeled by senses that contain more synonyms, and formal concepts lower in the lattice are labeled by senses that contain fewer synonyms.

Symmetrically, formal concepts lower in the lattice are labeled by words that have more senses, and formal concepts higher in the lattice are labeled by words that have fewer senses. No information is lost through this method of labeling only once per word and once per sense—the complete sets of senses and words can be read from the lattice as illustrated in the following examples. *Formal concepts* will be referred to, from here out, simply as *concepts.*

Senses are read off the concept lattice top down. To the top and right of the centre of the lattice can be seen sense 227.40.1, a prepositional sense from Category 227: *Covering.* This sense of *over* contains the following set of entries that share more than one sense with *over*: {*on top of, on, upon, above, over, o'er*}. These entries can be found on the lattice by following the lines (or links) down from the *Covering* concept, as follows: the concept below and to the left is labeled with *on top of*; the concept below and to the middle is labeled with *o'er*; following the link down to the right there is a concept labeled with *upon* and *on;* and finally, the concept below and linked to both the lower-left and middle concepts (labeled with *on top* of and *o'er*), is labeled with *above.* Together these labels make up the set of shared entries, or synonyms, of *over* found in Roget's Thesaurus Category 227: Covering, Paragraph 40, Sense 1.

The four senses labeling the bottom node contain no other entries (besides *over*) that are found in more than one sense of *over.* The top node is unlabeled as there is no sense that contains all of the words.

To find the senses of a particular word the lattice is read from the bottom up. So for example the word *over*, which is found in all senses, labels the lowest concept—all of the senses of *over* can be found by tracing the lines up (and conversely, all of the senses can be seen to contain the word *over* by tracing the lines down from them).

Above has six senses shared with *over*, {36.13.1; 206.24.2; 206.27.4; 227.40.1; 661:27:1; 40:10:1}, three of which are adverbial, one of which is prepositional, and two of which are adjectival. These can be identified and read off the lattice by tracing the lines up from the concept that is labeled with *above.*

The *scope* of the concept labeled by *above*, reading the lattice upwards, is the set of six senses of *above*; while the scope of the same concept, reading downwards, is the set of words that are contained as synonyms in the two senses that label that concept (*over* and *above*). In Formal Concept Analysis (Wille, 1989) the set of objects (the set of words) is called the *extent* of a concept; and the set of attributes (the set of senses), the *intent* of the concept.

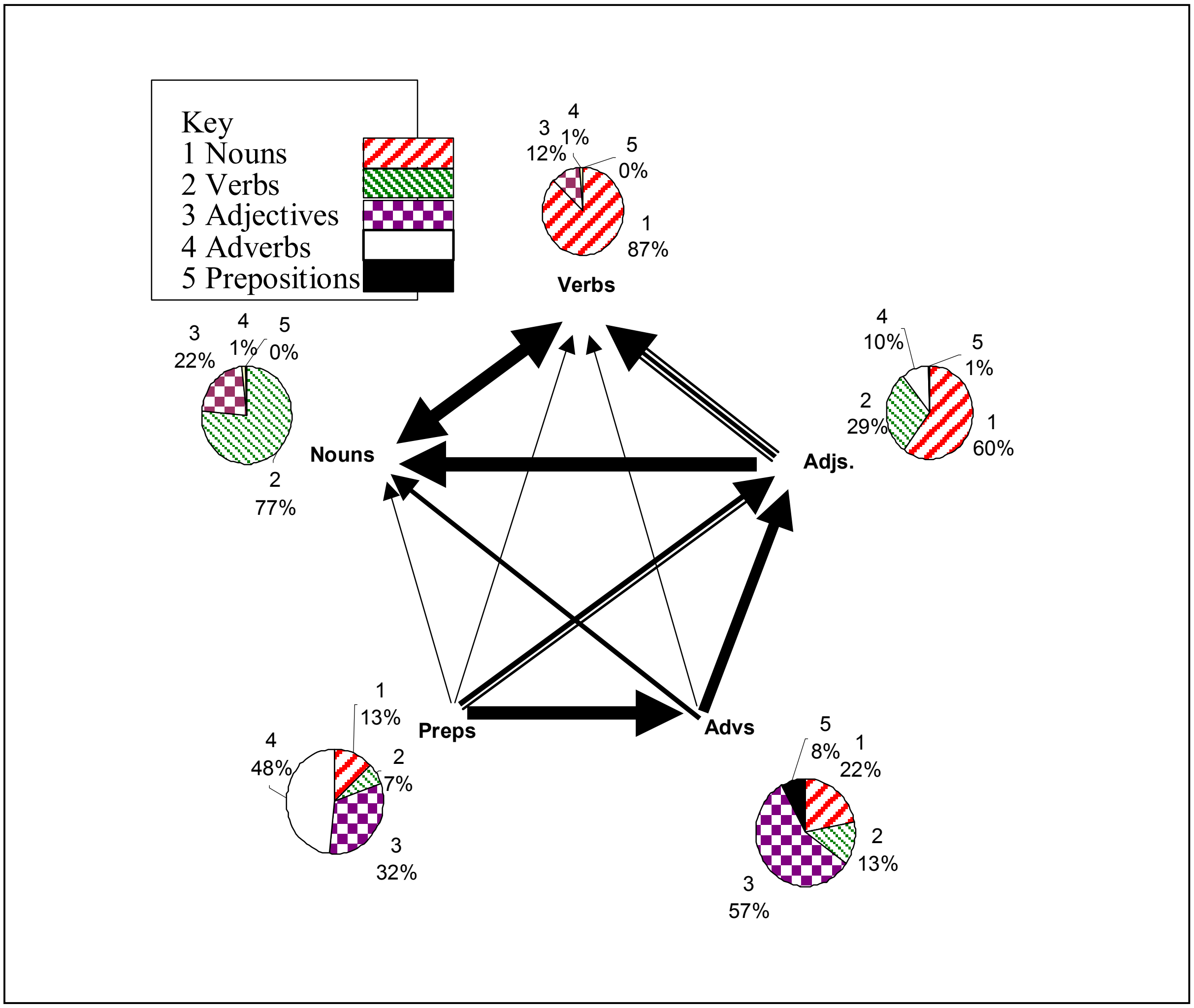


Figure 9. **Percentage overlap between parts of speech in Roget's Thesaurus**

It is not necessary to navigate the lattice expertly or understand the underlying mathematical formalism. Simply comparing adjacent concepts should convince the reader that this automatically-derived graphic has presented the senses of *over* in a coherent way—a way which supports Brugman and Lakoff's (1988) assertion that senses of a word are related and that there are gradual transitions, or *transformations*, as one navigates from closely to more distantly related senses. Similar lattices can be derived for any word in Roget's Thesaurus that has senses crossing part-of-speech boundaries.

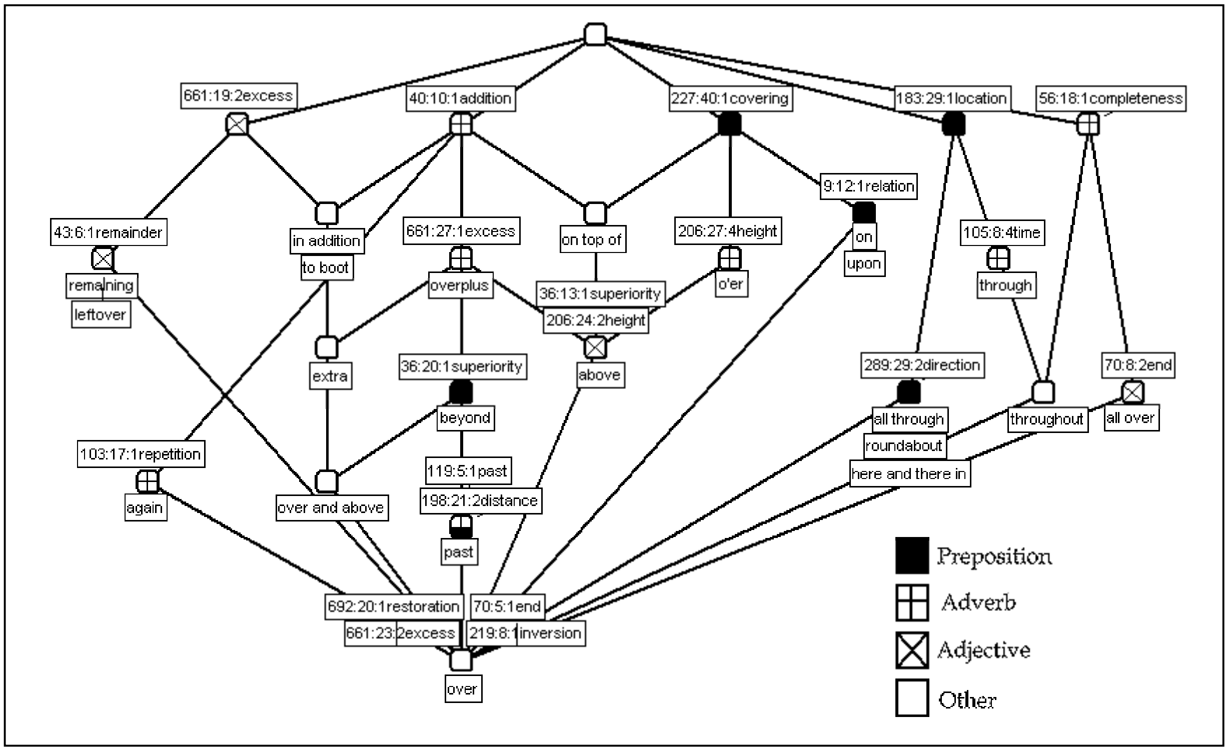


Figure 10.* Lattice showing the topology of relationships between entries, senses and parts of speech for the word *over

Figure 11 shows the concept lattice of *above*—also restricted to synonyms that occur in more than one sense. Six of the seven senses are shared with *over* (c.f. Figure 10). The seventh sense differentiates *above* from *over* in this lattice.

The automatically constructed lattices show that many closely related adjectives, adverbs and prepositions can be selected by focusing on a single word, and illustrate the overlap and blending among parts-of-speech, and among some words. These words are examples of the type described by Crystal (1987) as being at the "edges" of the word classes. They are the glue that ties the senses together, and incidentally, some of the most common (polysemous and high-frequency-usage) words in the thesaurus.

Genus and Differentiae

In contrast to Brugman and Lakoff's "radial category" of senses, there is no central sense evident in the lattice. None-the-less, the sense with index 40.10.1 from Category 40: Addition*,* an adverbial sense, shares words with many of the other senses. In the thesaurus it has 37 entries. Of the 37, 24 are words that have more than one part-of-speech, and 31 are polysemous. Of those with more than one part-of-speech, 14 double as adjectives, 12 double as prepositions, 4 as nouns[24], and 3 as verbs. Of the remaining "idiosyncratic" words (single-instance words, omitted from this lattice), *additionally, moreover,* and *furthermore* occur in the thesaurus only in this sense—they characterize it, differentiating it from other senses. They are the stripes that separate this tiger from other big cats—they distinguish this sense from other senses.

Synset 40.10.1, along with its idiosyncratic words, and relationships to other senses via those shared words, hints at what is at the core of prepositional semantics, it illustrates the concept of *genus and differentiae* used to construct sense-definitions in dictionaries. A

[24] In uses such as: the *more* the merrier; a blast from the *past*; a movie *extra;* a real *plus.*

simplified dictionary example would be: “A cup is a type of container (genus) that has a handle (differentia number one) and is used for drinking liquids (differentia number two).” Or, as mentioned earlier: “A cactus is a plant (genus) that lives in the desert (differentia number one) and has prickles (differentia number two).”

As stated earlier, Figure 10 includes only those words that share more than one sense with *over*. The words that do not share more than one sense with *over* include the differentiating entries in each of its senses.

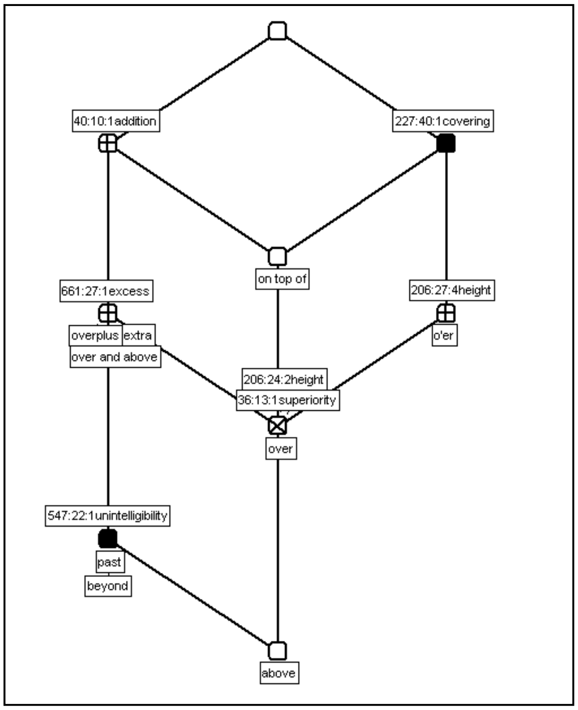


Figure 11*. Lattice of the semantic neighborhood of *above

So this lattice is a kind of “genus” topology, only. The missing words are what facilitate the discrimination of senses from one another in the same way that distinguishing features allow us to recognize and differentiate individual people, and living things are differentiated amongst in biological taxonomies. Examples of the missing differentiae are: 105:8:4 Time—“for the period of;” 70:5:1 End—“terminate;”183:29:1 Location—“here and there on;” and 661:27:1 Excess—“in excess of” (see Appendix B: Methods of Data Analysis, under *Formal Concept Analysis (FCA)*, for a simplified, prototypical example).

Moreover, there is a symmetric organization between the words and the senses. In the same way that senses can be read down the lattice (their constituent words identified), and words can be read up the lattice (their various senses can be identified), some senses act as differentiators for words and some words act in a “genus” capacity, gluing the senses together. No matter what size of neighborhood, or selection method is used to extract words and senses from RIT, the same symmetric genus and differentiae organization emerges—it is scale free.

This “genus-differentiae” facet of word-sense organization has implications for the conceptual organization of the brain. The organization seen in the lattice emerges naturally from the data—from the semantic relationships between synonyms, and from the transitional or transformational connections between senses of polysemous words. This organization provides a natural way to arrange information in a near optimal fashion—so that the pieces of information become neither isolated, nor too densely packed. These features are similar to those described for small-world networks in Section: *Small Worlds*, in Chapter 2. To the extent that concepts newly added to a small-world network act as semantic differentiators (Steyvers & Tanenbaum, 2001, p.5), and existing complex concepts (or hubs) act as genus classifiers, this genus-differentiae organization supplements the small-world model, with its global perspective, with a local model. The two perspectives together provide a unified model that may account the conceptual organization and classification of words in the mental lexicon. This is discussed further in Section: *Networked Words*.

Summary

A preposition is a word (or phrase). But in Roget's Thesaurus that specific word may be represented by many entries under separate prepositional senses. The same word, or string of characters (excluding homographs), may also have one or more entries classified under other, non-prepositional parts-of-speech. So, to say that *over* is a preposition is not to exclude it from being any other part-of-speech. Also, to say that *over* is a synonym of *above* is not to say that it is a synonym of *above* in all senses or, for that matter, for all parts-of-speech.

To say a word "means" something, or "is" a preposition, is misleading. Outside of usage (spoken or written context), the meaning of a word can only be understood in the context of the semantics of all of its senses, synonyms, and parts-of-speech, together. Despite this apparently overwhelming complexity, senses of words, in context, can be disambiguated[25] almost instantaneously by native speakers. It may not be "despite," but "because of" this complexity that we are able to do it. The patterns described here are implicit in RIT, and are consistent with the small-world model. Combined with the small-world model the genus-differentiae organization of words and senses, as viewed through

[25] And if not immediately disambiguated, at least identified as congruent with the current context.

the lattice structure, provide a new perspective on semantics.

Semantic Origins of Prepositions

Most words that are called prepositions also occur as other parts-of-speech. Their proto-Indo-European roots are frequently also the roots of verbs, adverbs, adjectives, and even pronouns. This hinders the separation and identification of early prepositional semantics, but insights into the etymological development of prepositions, as will be shown, also gives insights into the other facets of RIT semantics. This Section attempts to identify the semantics shared by the most important, or core, RIT prepositions by classifying the meanings of their hypothetical Indo-European roots. The model developed not only helps explain the semantics of prepositions, but probably accounts for the vast overlap in meaning between words in RIT from different word classes, or parts-of-speech—a phenomenon which supplies much of the connectivity within network models of language described in the review of literature on Roget's and related research (Chapter 2).

The classification, or semantic model, of roots proposed here has two main semantic groupings divided into four broad notions. This is explained in detail in the rest of this Section. Other interpretations are possible, and an attempt has been made to include all relevant roots (those which gave rise to core English prepositions in RIT) for further analysis by other researchers[26]. That includes those roots which are not prepositional in meaning, but from which some core prepositions are derived; some roots which are prepositional, but from which no known English prepositions are derived; and some Germanic roots for which there is no clear Indo-European ancestor.

[26] See Appendix C for the complete list of prepositions defined here as *core*, along with all related Indo-European roots.

Sources and Definitions

The English prepositions described here were selected based on the 411 words found in RIT labeled as prepositions. Some entries in the thesaurus are ambiguously classified as "Advs, Preps" some as "Preps, Advs." Only those labeled "Preps, Advs" were included in this study.

Approximately half of the thesaurus prepositions are also classified as other parts-of-speech. For example, as mentioned in the previous Section on *Semantic Overlap*, *after* has thirteen senses listed under five basic parts-of-speech. These are listed in Table 15, along with the number times each occurs in RIT.

Part-of-Speech	Sense count
Noun	1
Adjective	2
Adverb	5
Preposition	4
Conjunction	1

***Table 15.* Distribution of Part-of-Speech for the Word "after."**

Any words referred to in the following discussion will be, by default, their prepositional senses, unless otherwise stated.

Although there are 411 prepositions in Roget's Thesaurus, only about 35 are *core*. Core prepositions are defined here as those having a high frequency in usage, and excludes archaic prepositions such as *heretofore, fornenst, fornint,* and *anent*; compounds formed from other core prepositions, such as *without, into, upon*; participles such as *regarding,* *concerning*; and phrases combining core prepositions with words from other parts-of-speech, such as *in respect to, in connection with, over and above* and *on top of*. Examples of core prepositions are *at, up, out, of, by, on, in, with* and *to*. Of the non-core prepositions approximately half are phrases.

There is a danger in classifying any preposition as core. Some of the important prepositions discussed here occur more frequently in text as other parts-of-speech. Examples of prepositions that occur more frequently as adverbs are *off, up, down,* and *below. Since* occurs most often as a conjunction, and *past* more often as an adjective. The term *core* is used here in an attempt to constrain the complexity of the analysis, and to avoid issues with prepositions that combine the semantics of two or more distinct roots.

The Indo-European roots used in the following discussion are taken from Claiborne (1989) and the *American Heritage Dictionary* (AHD, 4th Edition).

Classification of Prepositional Roots

The Indo-European roots from which core RIT prepositions are derived are generally prepositional or adverbial in nature—that is, the words used in the sense definitions (meanings) assigned to the roots in Claiborne and the AHD often also occur in Roget's Thesaurus as core prepositions, as adverbs, or both. Root meanings may also be nominal, verbal or adjectival. For example *below* comes from a root meaning *to lay,* or *lie* (down), and *except* comes from a root meaning *to grasp*.

The four notions described below focus on the semantics (sense definitions, or meaning) of the Indo-European roots, rather than the semantics of their modern English descendant prepositions. Where the classification is ambiguous or unclear core descendant prepositions are used to clarify the semantics.

The first and main semantic group is a directed, forward-moving, separation, or transfer notion. The second is a stationary notion, often involving location, proximity to, or a relation between places or objects. That includes the concepts of being, containment, adjacency, and possession or having. The first notion could be summarized as *leaving* or *going* and the second as *being* or *staying*.

Verbs were chosen to characterize these prepositional notions as the Indo-European root sense definitions include terms describing a change of location, not just stationary positions and spatial relationships. Examples of such terms are *to, from, forward, away, pass* and *apart*. Even AD-, the Indo-European root of *at* (perhaps the prototypical English preposition for describing a stationary, point location), as well as meaning *at*, also means *to*. This sense was retained in early Indo-European languages. The descendant Latin root *ad* (the source of the prefix ad-), means *to, toward,* while another descendant, the Celtic root *ad-, means both *to* and *at* (*American Heritage Dictionary*, *ad-*).

A third notion found among the Indo-European roots, although more fully formed in modern descendant languages, could be called the *up-down* or *high-low* axis. Prepositional Indo-European roots contributing to the high-low axis extend from, or at least overlap with, those classified under the *leaving/going* notion and emphasize (the semantic emphasis is on) being or going up, more than being down.

The fourth and final notion extends from the first notion as a consequence of it. The notion may be characterized as *at a distance, elsewhere*, *set apart*, or *gone*.

The four notions are demonstrated below using those Indo-European roots that have related adverbial/prepositional senses, and the core RIT prepositions derived from them. The following format is used to present the roots:

- ROOT- (root meaning) >> ***core*** *derivatives, other derivatives …*

The symbol ">>" represents the phrase "whence we get," referring to the derivative, or descendant, prepositions. Core prepositions that do not have prefixes from other roots are emphasized in bold-type.

The Leaving or Going Notion

Roots that support this first notion include:

- PER-1 (forward, through) >> ***for****, before,* ***from****,* (also *far, forth, farther, further*);
- (A)PO- (away, off) >> ***of, off, after***;
- DE- (this, that, that way, to) >> ***to****, into* (also *too*);
- TERh-2 (cross over, overcome, pass) >> ***through****, (throughout, trans-)*;
- UD- (up, out) >> ***out****, about*;
- UPO- (under, up from under) >> ***up*****,** ***above****, hypo-, sub-*.

(A)PO- is also the source of the Latin prefix *ab-* (away from) in words such as *about.*

Above, perhaps counter-intuitively, is not amongst these words. It is derived as follows:

UPO > (Teutonic) ufan > be-ufan > bufan > a-bufan > *above* (OED, *over*).

The Being or Staying Notion

Roots that support the second notion include:

- AD- (at, to, near) >> ***at***;
- (A)MBHI- (around) >> ***by****, abaft*;
- ANT- (front, forehead, before, opposite) >> *<u>a</u>long, (ante, anti-, un-);*
- AN-1 (on) >> ***on***;
- EN- (in) >> ***in***;
- MEDHYO- (middle) >> *amid;*
- WI- (in half, apart) >> ***with***.
- DWO- (two) >> *between, betwixt.*

(A)MBHI- is also the source of the prefix *be-* (by, near) in words like *between*.

(A)MBHI- is "Probably derived from *ant-bhi,*from both sides*" (AHD, *ambhi*), but like AD-, this *stationary* preposition also harbours some of the semantics of motion—the descendant Sanskrit root, *abhi,* mean *to, toward* (Webster's, *by*).

There are a few important English-preposition-producing roots that are strictly Germanic:

- DHUNO- (Gc) (fortified place, enclosure) >> ***down*** (also *dune* and *town*);
- *GAGINA-* (Gc) (in a direct line with) >> *against* (also *again.* But *gain* comes from a different root, WEIh-);
- *NEHW-IZ-* (Gc) (near) >> ***near****,* (*nigh, neighbour, next*).

DHUNO-, being a noun, seems to be an odd root to supply us with such an important concept as *down* but there may have been an older, more prepositional, Indo-European root that gave rise to the concept. Vincent (1999, p. 30) hints at residues of it found in the Romance languages as the "pan-Romance preposition *de/di*, which once had a sense of 'down, away from' (c.f. Latin *de-scendo* 'I come down')." There is also an old Germanic root, DAN-, meaning *low ground*, from whence we get *Dane* and *den*, which could support the plausibility of an ancestral Indo-European root. The Indo-European root DANU-, meaning river, and from whence we get *Danube* and *Donau*, may also be related, as a nearby river is always the local geographic low point.

Two prepositional Indo-European roots which have influenced many English words, although produced no core prepositions, are:

- KOM-1 (beside, near, with, by) >> *co-, con-, com-, contra-;*
- KSUN- (with, together) >> *syn-, sym-*.

Some other important prepositions derive from Indo-European roots that are not themselves prepositional. The prepositions derived from the first three roots below could be classified with the *going* notion, and the last, the *staying* notion.

- GWRES- (fat, thick) >> *cross, across*;
- WER-3 (turn, bend) >> *toward,* (also *forward, seaward, skyward, ...);*
- RET- (run, roll) >> *round, around*;
- MAG- (knead, fit, fashion) >> *among, amongst* (also *make* and *mingle)*;

Toward inherits its *prepositionality* from *to*-. The *a*- prefix(es) that contribute prepositionality to the other prepositions, despite their similarity, do not necessarily share a common ancestry.

The High-versus-Low Notion

The third notion consists of prepositions that form a clear axis, even among their Indo-European roots. Those not listed as belonging to the first, *going* notion are:

- UPER- (over) >> ***over**, (super-, hyper-);*
- LEGH- (lie, lay) >> *low, below;*
- NI- (down) >> *beneath, (nether, Netherlands);*
- NDHER- (under) >> ***under**, (infra-, inferior).*

A related root, but which does not have any descendant English prepositions, is NER-1:

- NER-1 (under, on the left, left, left of eastward) >> *north, northern, Norman, Nordic, Norse, Norwegian.*

North was *left* because the primary Indo-European compass point was the East, the direction of sunrise, from AUS-1 (shine, direction of sunrise). West was WES-PERO- (evening, night), where the sun went down; and South was SAWEL- (sun) >> *sun, solar, helios…*

The Notion of Being Gone, or Set Apart

The fourth notion, that of *being* or *having become set apart*, is reflected in the following roots, several of which are not prepositional, and most of which do not contribute core RIT prepositions:

- AL-1 (beyond, other) >> (*ultra-, alter-, ultimate, utter, alias, alien, other, else* … also *alarm, alert*);
- PETh- (spread, stretch out) >> *past*;
- KAP- (grasp, hold) >> *except;*
- (a pronominal stem) >> *beyond* (*yon, yonder;* also *ilk, id, if, yes, yet*);
- EGHS- (out) >> (*ex-, ecto-, extra-, exotic, exterior, extreme, strange* …);
- ETI- (above, beyond) >> *eddy, eider* (duck)*, et* (Latin, meaning *and*).

Except gains its prepositionality from the semantics of the prefix *ex-* (from the Indo-European root EGHS-).

The *gone or set apart* notion has only weak support. Either the roots reflect the notion, but have no descendants, or there is a descendant core preposition but the root does not reflect the notion.

A summary of the proposed four-notion model for classifying the meanings of the prepositional Indo-European roots is presented in Figure 12. The definitions of the roots are used to label the elements. Figure 13 shows the actual roots mapped to the same schema. Figure 14 shows the core prepositions that are derived from these roots. The classification model accounts for 22 of the 33 core prepositions—and all of the most important ones.

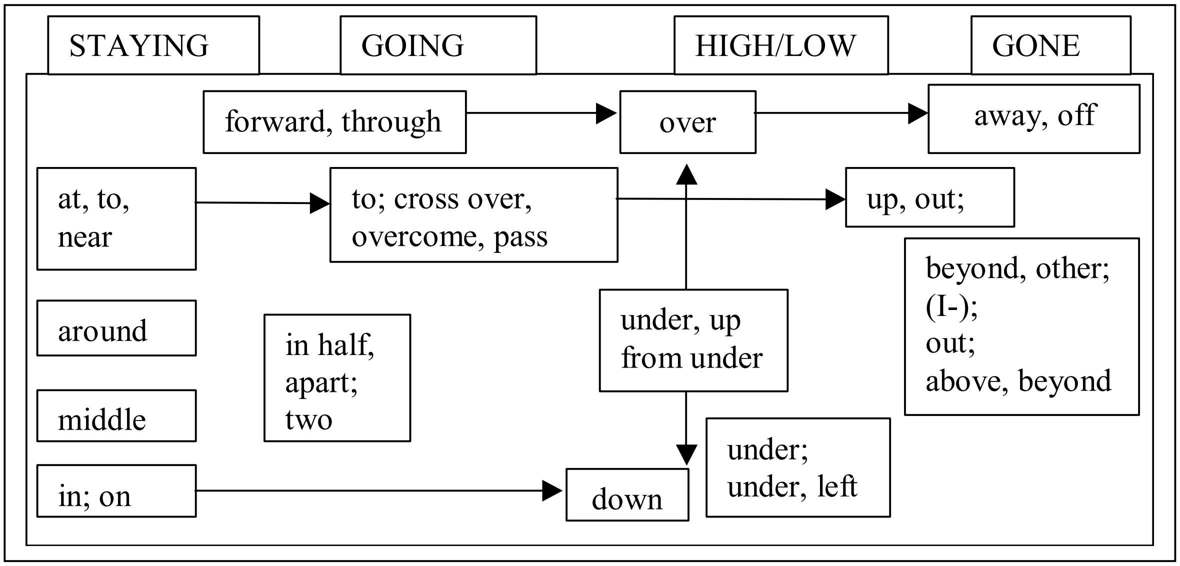


Figure 12. **Sense definitions of prepositional Indo-European roots.**

Complex Issues

The proposed model does not incorporate the important dimensions (discussed below) of *behind* (to the back) and *beside* (to the side) found among modern English prepositions. They were omitted from the four-notion model because they are not strongly represented in the sense definitions of the prepositional Indo-European roots given by Claiborne and the AHD.

The model also glosses over several complications in order to narrow the core notions. These *complications* include the fact that many of these prepositional Indo-European roots have other descendant words (and sometimes prefixes that form parts of words) that are not themselves prepositional. Furthermore, some apparently simple non-prepositional English words have been, over time, elaborated into a wide range of Modern English prepositions.

The first root listed under the *going* notion, PER-1, has many tones of meaning and has contributed to hundreds of English words. The *American Heritage Dictionary* (AHD) includes this description of PER-1:

> [it is the base] of prepositions and preverbs with the basic meanings of “forward,” “through,” and a wide range of extended senses such as “in front of,” “before,” “early,” “first,” “chief,” “toward,” “against,” “near,” “at,” and “around.” (AHD, *per-1*)

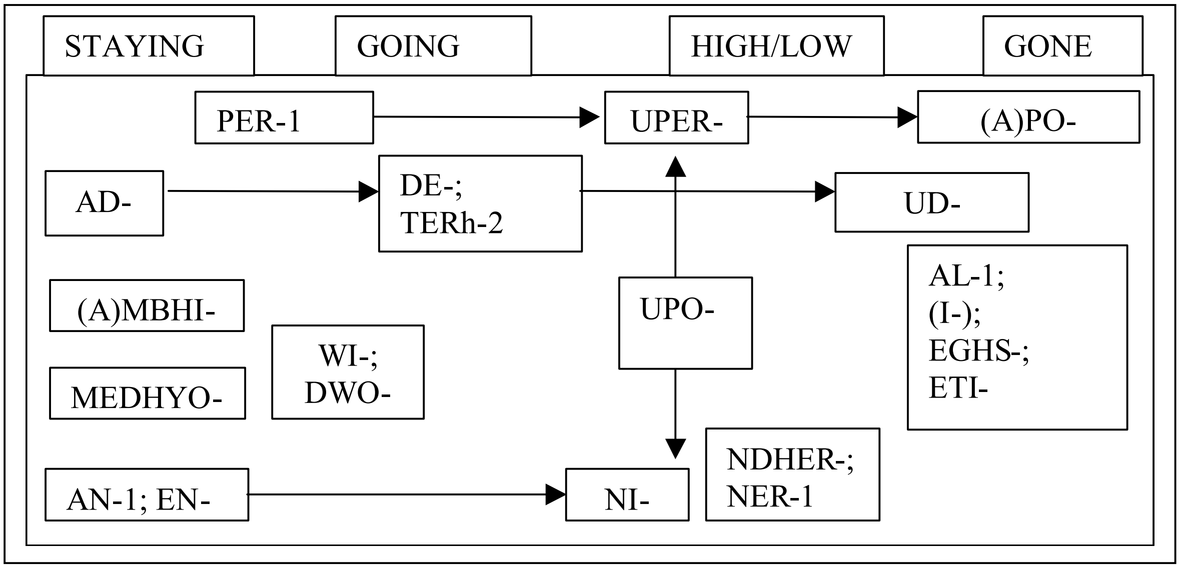


***Figure 13.* Classification model of prepositional Indo-European roots**

PER-1 is also the source of some of our most common prefixes: *for-, fore-, para-, pro-, pri-, pre-, peri-,* and *proto-*.

As more descendant words are included the original semantics becomes more clouded. Common cognates (words derived from the same root) of for and from are: far (adj), farther (adv), further (adv, vb), first (adv, adj), and forth (adv). Compound cognates include: afford (vb), approach (vb), appropriate (vb, adj), approve (vb), approximate (vb, adj), deprive (vb), forbid (vb), forget (vb), foremost (adj), forward (adv), furnish (vb), furniture (n), improve (vb), paradise (n), paradox (n), ... precede (vb), proceed (vb) … premier (n, adj), priest (n), primary (adj), primate (n), ... prince (n), principal (n), principle (n) … professor (n), prompt (vb, adj), propel (vb), propose (vb) … purchase (n, vb)… There are more than 120 descendants in all (not including derivatives such as inflected forms).

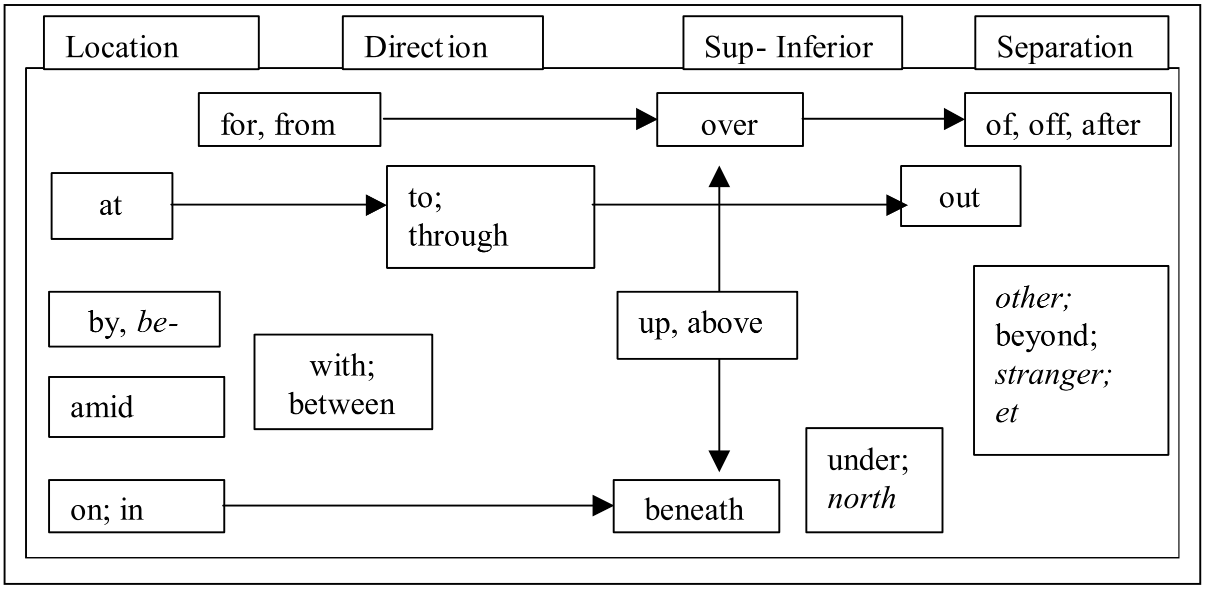


***Figure 14.* Core prepositions mapped to the classification model**

PER_1 is an extreme case, but illustrative of how the senses of words come to overlap in subtle ways, crossing part-of-speech boundaries and carrying their essential (meaning, *of the essence*) semantics into associations with words of different etymological origins.

The preposition *beside* comes from *side,* which is derived from the root SE-2:

- SE-2 (long, late) >> *side, beside, besides, sidelong, alongside…*

The full list of words classified in Roget's Thesaurus as prepositions related to *side* is: *beside/besides, aside, inside, outside, alongside, along the side of, aside from, aside of, inside of, outside of, on one side, on the side, by the side of, side by side, sidelong.* Clearly, including all related forms of a preposition also would make the identification of semantic commonalities more complicated.

The preposition *behind,* is derived from the root KO-1:

- KO-1 (this, that) >> *hind, behind, hinterland, here—(*also, *he, her, him, it*).

Beside could still be classified with the *staying-being* notion. *Behind* could either be classified with the fourth (*gone and apart*) notion, or used to introduce a new notion with an axis something like *before-behind* or *in front of-to the back of;* but there are no apparent Indo-European roots supporting this axis or notion. An interpretation of the roots, by others, might support it.

Time and Temporal Prepositions

A major prepositional concept has been omitted from the discussion so far—that of time. Time has been omitted from the model as modern references to time are usually made via a spatial analogy, such as to a path—in a sense we travel or pass through time—and expressed using spatial prepositions. We think about short or long durations of time (*through; throughout; during*), and points in time (*before* sun-up; *after* sunset; *past* 10 o'clock). *During* and *duration* are derived from the root DERU- (firm, solid). The only root that is both temporal and prepositional and has descendant prepositions is: SE-2 (long, late) >> *side, beside,* and since.

The other prepositions related to time, *in, on* and *at,* are used in the same general-to-specific order as in their spatial usage: *in* the first millennium (*in* the garden); *on* Saturday (*on* the path); *at* 3pm (*at* the apple tree).

Many temporal words are adjectives. *Late* comes from LE-2 (slacken, let go). *Now* comes from a root of its own, but which is closely related to the root of the adjective *new* (AHD *nu-* and *newo-* respectively) and probably related to the Germanic root of *near*, NEHW-IZ-. *Then* has the same root, TO-, as *the, this, there, that, they* and *those*; all of which are indexical (and suggest location) rather than temporal. *When* comes from KWO-, the same source as *who, what* and *where.* Other Indo-European roots with temporal senses refer to nouns such as *morning*, *day*, and *season* (from a root meaning *to sow*), more closely related to the concepts of repeating periods and cycles of the sun.

Related Prepositional Roots

PER-1, UPER-, (A)PO-, and UPO- are probably mutually related roots—having a common origin. Although Claiborne and the AHD list UPER- (the root source of *over*) and UPO- (the root source of *up* and *above*) separately, they are clearly different forms of the same root. The *Oxford English Dictionary* (OED) gives the following etymological information under the headword ***over***, which makes this explicit:

> Skr. *upari* adv. and prep., locative form of *upara* adj. 'over, higher, more advanced, later', comparative form of *upa*, in Teut, *ufa-, uf-,* whence the adverbial *ufan* (see OVENON, ANOVEN) and *be-ufan, bufan,* with the compound *a-bufan,* ABOVE. Over was thus in origin an old comparative of the element *ufa*, *ove,* in *ab-ove*. (OED, *over,* Volume X, p. 1056)

PER-1, the strongest prepositional root generatively, is probably the precursor of these related roots, and is perhaps at the heart of prepositional semantics in RIT. The only root not included in this analysis that could compete with PER-1 in terms of numbers of English descendants, of all parts-of-speech is:

- STA- (stand) >> *stand, stem, state, stable, stay …*

The semantics of STA- fits tightly with the second notion, that of *staying*. As noted elsewhere (Old, 2000) it is likely that the ancient concept or notion underlying the formation of prepositions was also the basis of a wide range of concepts, having a variety of parts-of-speech, all to do with the nature of being or staying, and to ways of leaving or going.

NDHER-, NI-, and NER-1 also appear to be mutually related roots, as do EN-, ETI-, AN-1 and ANT-; DE-, TERh-1, DEL-1 and UD-; and WI- and DWO-.

A template for the classification of the meanings of the core RIT prepositions, based on this model, and utilizing the spatial concepts of *direction*, *position* and *location* as classes, is proposed in Table 16a. The classification refines the notions developed in the model, through the addition of differentiating features such as *adjacency* and *proximity* to discriminate among the different functions of the prepositions in modern English usage. Table 16b shows the corresponding Indo-European roots for comparison.

As stated earlier, (A)PO- is also the source of the prefix *ab-* and (A)MBHI- is the source of the prefix *be-*. These and other prefixes, through the semantics of their roots, give many of those miscellaneous and often-unexpected roots discussed throughout this Section their prepositional (and *position-al*) senses. Those roots and their descendant prepositions are listed in the columns labeled *Position* in Tables 16a and 16b. An insight into the distinction intended here between *position* and *location* can be gained by comparing *up, above* and *over*. Location is more specific than position, and direction is more general than position.

	Direction	**Position**	**Location**
Possession	**for**	about	**of**
Containment	**through**	between amid among	**in**
Adjacency	**with**	across along against	**on**
Situation	**to** **from**	around	**at**
Proximity	**by**	beside before behind	**near**
Separation	**out** **off**	beyond	**after**
Superiority	**up**	above	**over**
Inferiority	**down**	beneath below	**under**

Table 16a. **Core prepositions arranged by features.**

Conclusion

The *gone and apart* notion is quite weak with regard to prepositional roots. It could relate to *the territory beyond* (or *outside of*) *ours*. Or it could relate instead to the concept of strangers (aliens, outlaws, or *auslanders*) rather than the positional-locational concept we usually associate with prepositions. In support of this, the roots of *yonder* and *behind* are also the roots of pronouns; EGHS- (out) is the root of stranger; and AL-2 (beyond, other) is the root of *alien, alias, else* and *other*.

The *going* or *staying* notions can also be interpreted as “giving or keeping,” “there and here,” or “then and now.” These interpretations are not mutually exclusive. There is sufficient semantic similarity among the notions for mappings such as metaphor and analogy to occur among them. Their uses could exist concurrently and be discriminated based on context.

	Direction	**Position**	**Location**
Possession	PER-1	UD-	**(A)PO-**
Containment	**TERh-2**	MEDHYO- MAG- DWO-	**EN-**
Adjacency	**WI-**	GWRES- DEL-1 GAGINA-	**AN-1**
Situation	**DE- PER-1**	RET-	**AD-**
Proximity	**(A)MBHI**	SE-2 PER-1 KO-1	***NEHW- IZ- -***
Separation	**UD- (A)PO-**	I-	**(A)PO-**
Superiority	**UPO-**	UPO-	**UPER-**
Inferiority	***DHUNO-***	NI- LEGH-	**NDHER-**

***Table 16b.* Roots of core prepositions**

A more attractive model might have been one that matched with the *X, Y, Z* axes of a three-dimensional space (up-down; front-back; left-right), and which classified the prepositions according to those dimensions, but the Indo-European roots do not appear to support this. Such a three dimensional, abstract world-model may not have been fully conceptualized until the development of geometry. It is certain that the early Indo-Europeans would have been at least capable of giving directions to, and specifying the location of, remote resources. The first two prepositional notions (*going* and *staying*) may be interpreted in terms of the concepts of *direction* and *location*, respectively.

The last two prepositional notions (*high-low* and *set apart-gone*) may be interpreted as having to do with *position* alone, but as shown in Table 16a, their semantics overlaps with all three concepts (*direction, position and location*).

RIT has many words that have multiple parts-of-speech. In the senses in which these words occur are found many words from other semantic backgrounds. This contributes to complex semantic connectivity crossing POS boundaries within RIT. The model of prepositional semantics proposed here could be extended to account for these other parts-of-speech. For example, as illustrated using the IE root, PER-1, many non-prepositional words share the roots used here to define the model.

There are two important issues that have not been addressed in this Section. They are polysemy (*over,* alone has five different prepositional senses and many more non-prepositional senses, as discussed in the previous Section: *Overlap among Parts-of-Speech*); and the antonymy relation as observed in the following pairs of prepositions {*in-out*; *on-off; to-from*}. Polysemy, as explained in Chapter 2, probably derives from words acquiring new meanings[27] over time through metaphor/analogy—prepositions are no different from other words in this way.

Antonymy is not straightforward for prepositions, and is inconsistent with the model proposed here. The semantics of antonymous prepositions possibly derives from their related adjectival senses. As Miller, et al. (1993) suggest (discussed in Chapter 2),

[27] As opposed to new words coming into the language through such mechanisms as coinage, derivation, borrowing, compounding, blending, clipping, acronym, conversion, and backformation.

adjectives naturally form antonymous pairs. Further research is needed to identify how this relationship has developed within prepositions.

As will be seen later in this Chapter, the core notion of *change of location*—from *staying* to *gone*—identified in this model, is at the heart of semantic connectivity within Roget's Thesaurus as a whole.

Visualization of Implicit patterns

Visualization allows patterns to be viewed that cannot be otherwise detected. Spatial representations allow metaphors to be drawn between abstract entities and the their relationships, and metaphors of distance (proximity) and direction between graphical entities. Using dimensions of color, size and symbols, other attributes of the abstract entities can be added to or overlaid on the space created.

Analysis using visualization is interactive and recursive, as stated in Appendix B: *Methods of Analysis*. Feedback through interaction facilitates the development of theories, which may be confirmed or confounded by modification of, or generation of new visualizations.

There are some drawbacks. The results, if any, are difficult to explain to others through a single graphic meant to encapsulate the results. Without animation, interaction with the data (such as querying and zooming), and often even without color, obvious relationships become opaque. In general, however, by using these methods hidden, or implicit information in RIT can be made explicit.

Local Views

Several methods for making explicit views of subsets of the implicit information in RIT are shown below. The basic unit is the semantic neighborhood of a word.

Multi-Dimensional Scaling of a Neighborhood

Figure 15a shows a multi-dimensional scaling (MDS) map of the semantic neighborhood of the word “over.” The relevance metric was derived from formal “Concepts” generated by the Formal Concept Analysis algorithm, using data from the restricted neighborhood of *over*. A restricted neighborhood includes only those words that share more than one sense with the topic word—in this case, *over*. The points in the MDS map represent Concepts, the words represent Objects of the Concepts, and the Synset index numbers represent Attributes of the Concepts. Similar results could be achieved by defining a relevance metric between words (for example, based on the number of shared senses, or Synsets), but this method has the advantage of combining both the Synsets and the words on the same map.

It can be seen that the word “above,” labeling the Concept at the top of Figure 15a, is situated near to the semantically similar Concept labeled by {on, upon}, but far away from the dissimilar Concept labeled by {remaining, leftover}. Words like “past” and “beyond” (to the right), and words like “all over” and “through” (to the left) are also arranged in a way that reflects their semantic similarity or differences. The concept labeled by the word “over” is central. This may be compared to the lattice of *over* presented in the discussion in Section *Overlap among Parts-of-Speech*, which was derived from the same data set.

MDS has limitations. Data is forced into two, or at most, three dimensions. If the data has an inherent high dimensionality then information may be lost which is relevant to the relationships within the data. For simply four points in a two-dimensional plane it is impossible to place the fourth point at an equal distance from three existing equidistant points, without using the third dimension. So any MDS of data of any significant dimensionality will necessarily be a compromise. None-the-less insights can be gained from such representations.

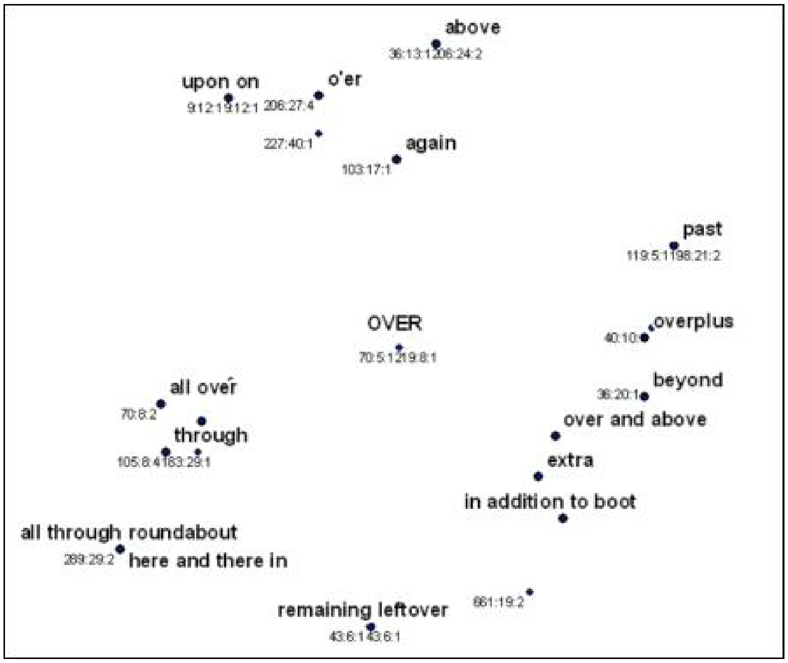


Figure 15a. **Multi-dimensional scaling of the restricted neighborhood concepts of "over"**

The data from Figure 15a can be imported into a Geographic Information system (GIS) to create an "information map," and the points color-ramped (colored according to a scale) to show the relevance of each concept to *over*. Such information maps retain the spatial relationships that represent the semantic similarity between "over" synonyms from Figure 15a, but allow the addition of other data dimensions (using color, in this case) to the semantic information without detracting from, or distorting the original spatial information.

Formal Concept Analysis

As discussed in Section: *Genus and Differentiae*, above, and the explanation of FCA in *Methods of Data Analysis*, Formal Concept Analysis (Wille, 1982) is an automatic method of classifying and arranging the relationships between formal Attributes and formal Objects. Any pair of sets with a relation between may be substituted for Objects and Attributes and relation automatically transformed into a lattice representation.

Figure 15b shows a "concept lattice" of the semantic neighborhood of the word "agitate" in RIT. The words are ordered reading down, and the senses are ordered reading up. For example the word "bother" is found in two senses, **888:3:1** and **864:15:1.** *Agitate* is found in all of the senses. *Torture a question* (found in the top right corner) is found only in sense **481:16:1**, and characterizes it. Sense **481:16:1** contains only three Entries, the Synset {agitate, agitate a question, torture a question}. Sense **888:3:1** contains the Synset {concern, bother, trouble, disquiet, disturb, agitate}. The words found labeling only the Concepts that have Synset index numbers attached are differentiae within the "agitate context." The Category labels, such as 888 Anxiety and 323: Agitation, have been omitted here but their inclusion can also add semantic context.

This lattice makes explicit the complex set of relationships between the topic word, "agitate," and its synonyms and senses in RIT. Concept lattices can be used to represent any aspect of RIT where there exists a relationship between two sets, whether words, senses, categories or other objects of interest.

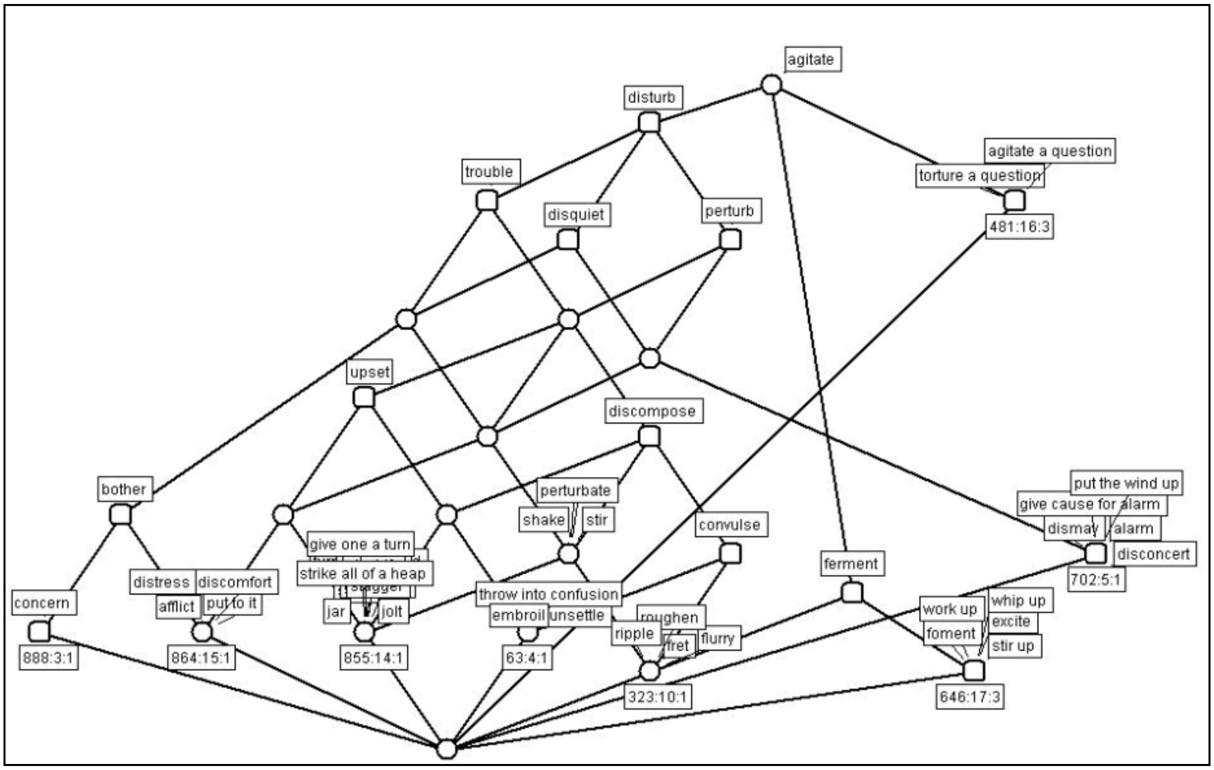


Figure 15b. **Formal Concept Analysis lattice of the neighborhood of the word "agitate"**

FCA can also be used to combine elements of RIT with elements from other data sources. By selecting the descendants of an Indo-European (IE) root, for example, a lattice can be generated showing the semantic influence of that root within RIT. Figure 15c shows a much-simplified lattice—a *restricted neighborhood lattice*—of the descendants of the IE root, SEK-, which means "cut."

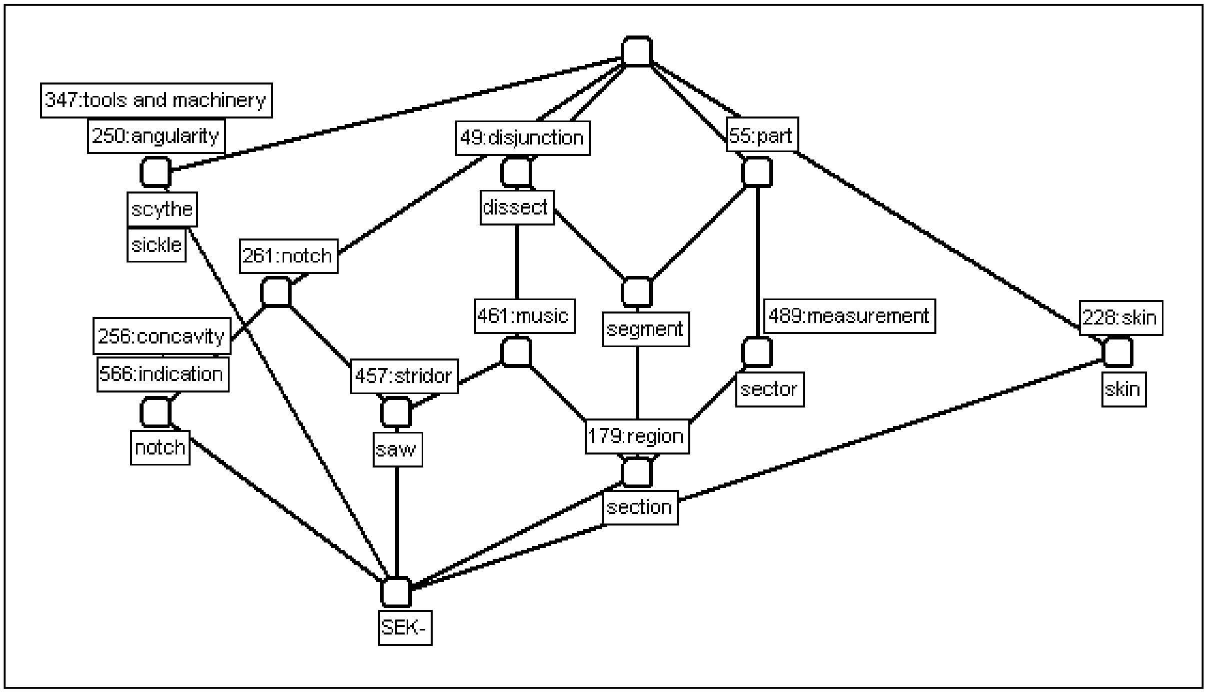


Figure 15c. **Restricted neighborhood lattice of the descendants of IE root "SEK-"**

A restricted neighborhood shows only those words that occur in more than one sense, or senses that contain more than word (or both) depending on the context. "Word" here means string, by part-of-speech, so that the sentence "I will skin [vb] the skin [n] off my shin" contains two different words spelled "skin." The full lattice would contain 54 senses in 36 Categories.

By converting data into spatial form, spatial processing software can be used to manipulate the data using spatial metaphors. Figure 15d shows the lattice from Figure 15b displayed in mapping software, a Geographic Information System (GIS). The color index (color ramp, GIS terminology), from red to blue, indicates polysemy of the words in RIT as a whole.

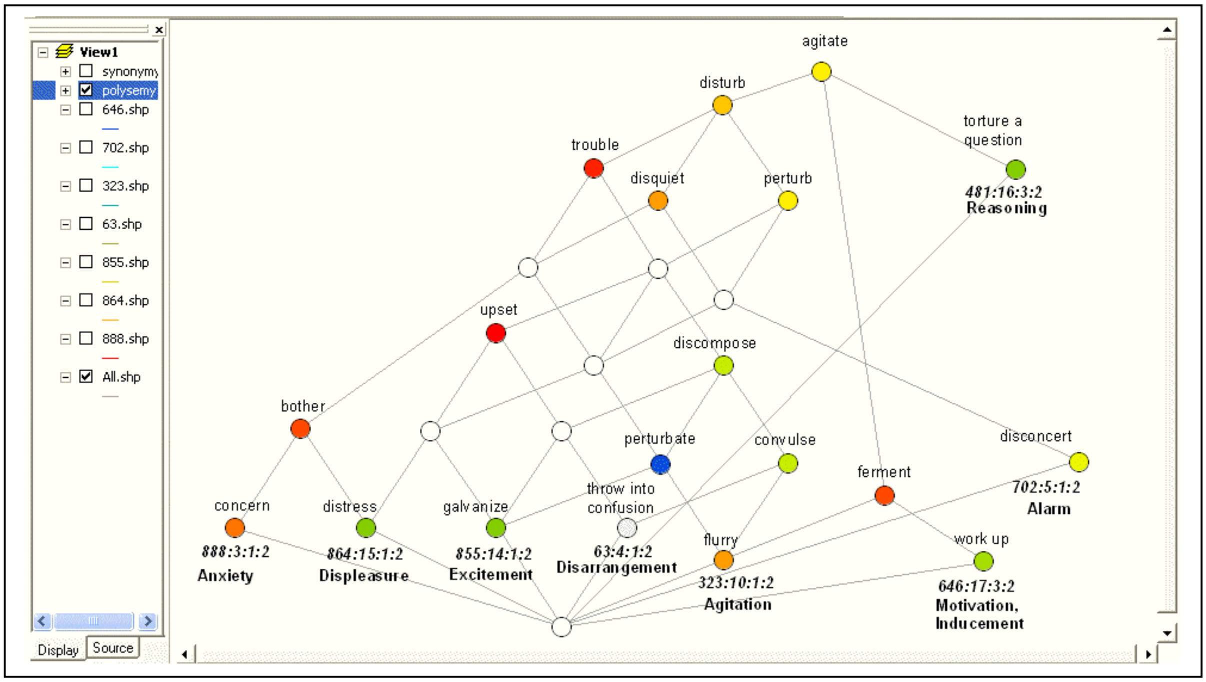


***Figure 15d.* Lattice of "agitate" in a GIS. Color ramp reflects polysemy**

Local views of data can aid in the development of intuitions about the data, especially when multiple local views are compared, but say nothing about the global structures of the data set from which the local subset was drawn. The following Section deals with this issue.

Global Views

By generating X and Y coordinates from values, such as RIT index numbers, a map display can be made of facets of the entire thesaurus. The coordinates may also be derived from cross-references, word order, or part-of-speech.

Figure 16a shows an information map derived from the Category index numbers in RIT. The values for the points here are derived from a count of the number of words shared between Categories. In other words, the points represent *semantic sharing*, in the same way that a cross-reference represents a semantic link between one Synset and one or more paragraphs.

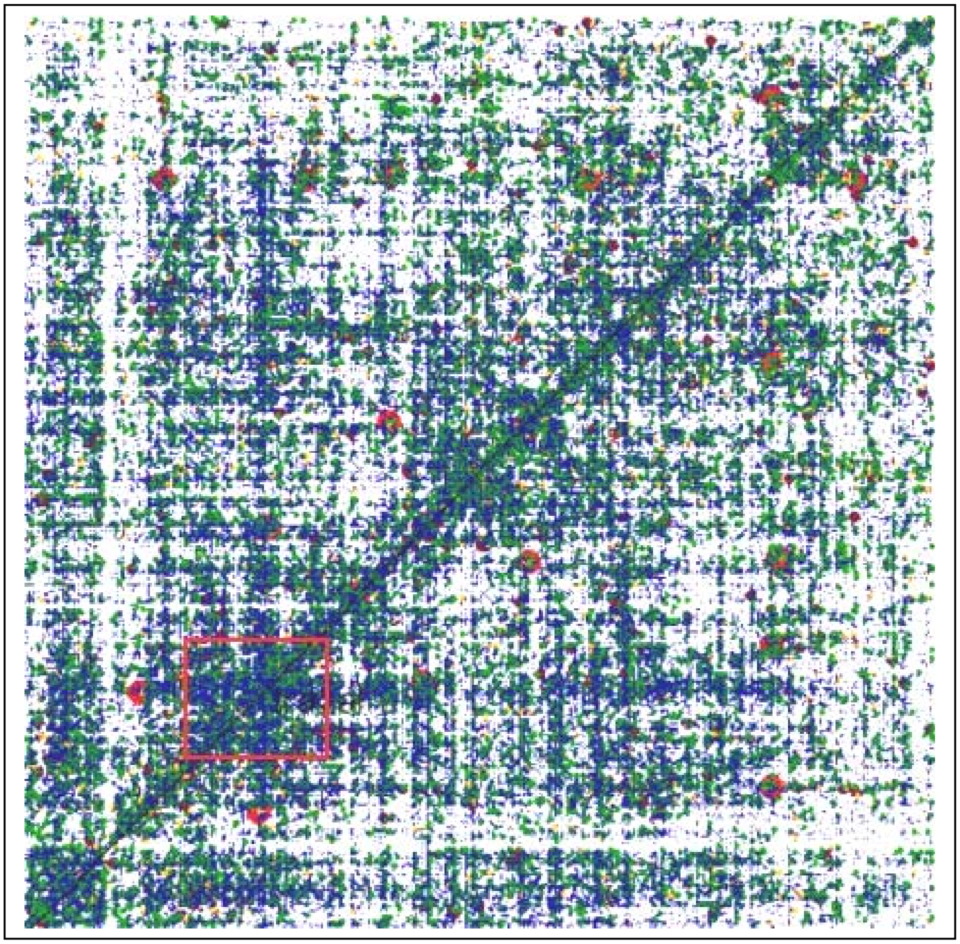


***Figure 16a*. Overview of RIT: Number of shared words between Categories**

A scale generated by the GIS program assigns numbers to a color ramp (shades of red, orange, yellow, green, blue). Points representing large numbers are colored red; small numbers are colored blue. Although a point-location on the map exists for every member of the Category-Category relation, for Categories that share no words, there is no actual point displayed.

The identity diagonal (running from top right to bottom left) is simply the relationship between the Category and itself (equivalent to a SelfRef-type cross-reference). The map (also known as a scatter plot, dot plot, or two-way contingency table) is symmetric about the diagonal axis (the identity diagonal), as the relation between Categories is symmetric ("Category 3 shares words with Category 270," means the same as "Category 270 shares words with Category 3").

Clusters (denser areas of points) near the diagonal indicate coherence, or internal word sharing within Classes. Points further out indicate words shared between Categories at a distance from each other, often crossing Class boundaries.

It can be seen that there is a "confluence," or higher density of points, forming a cross, the center of which is highlighted by the red box. The cross is formed by a group of Categories that share words with many other Categories throughout RIT, and whose X and Y co-ordinates are centered near the location where the cross intersects the diagonal. Figure 16b shows a close-up view of this group, with some of the points labeled. Although the labeled points here are those Categories with the highest counts of shared words, it is the overwhelming number of Categories with lesser counts (green and blue points) that make the cross visually prominent.

The Categories shown in Figure 16b are in an area of the thesaurus where there are high densities of very polysemous verbs. These verbs have to do with motion, movement, travel—or “change-of-place.” Examples of the Categories with high numbers of shared words are 855: Excitement / 323 Agitation (182 words); 707: Haste / 268: Velocity (86 words); and 173: Tendency / 289: Direction (38 words). Categories 323, 268, and 289 are located in the part of the thesaurus where the cross intersects with the identity diagonal in Figure 16a.

As noted, it is the density of word sharing that accounts for the visibility of the cross in Figure 16a. Not all Categories with high numbers of shared words are located in this group. Category 182: Town and Category 180: Country, share 256 words; 410: Plants and 306.1: Food, share 125 words; and 578: Language and 417: Peoples share 124 words.

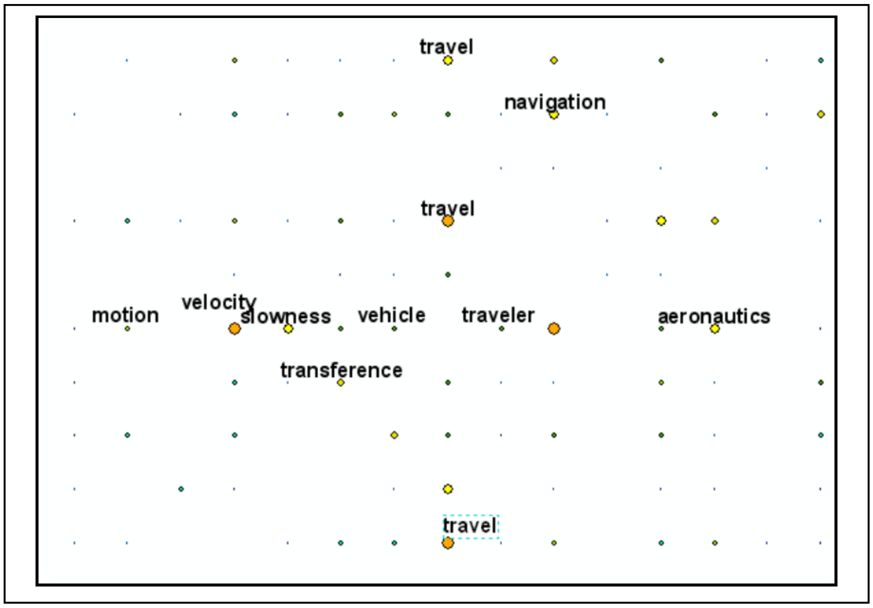


***Figure 16b*. Close-up view of information map “Number of shared words between Categories,” sparsely labeled with Category names**

Figure 17 shows the distribution of shared words (Word-Word relation) by part-of-speech throughout the thesaurus, overlaid by a grid showing the Class divisions. Verbs, displayed as green points, appear prominently and in approximately the same pattern as that of the areas where Categories sharing high numbers of words commonly occur. That is, where the *cross* appears.

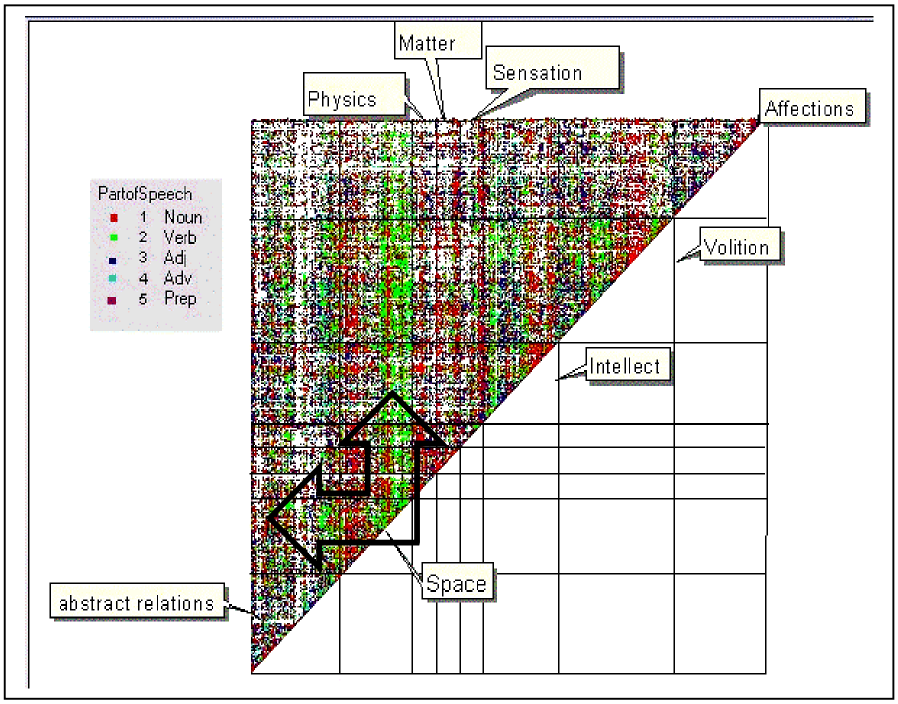


***Figure 17.* Distribution of words shared between Categories, colored-coded by part-of-speech, and overlaid by a grid showing Class divisions. (Arrows indicate a high-density area of high-polysemy verbs)**

An overview of the Category level and the higher levels (Letter Class and Roman Class) in the Hierarchy can be obtained by using a three-dimensional display. Figure 18a displays the same information as seen in Figure 16a, along with equivalent information maps from the Letter level Classes (middle layer) and Roman level Classes (top layer). Locations on a map at one level correspond to equivalent thesaurus locations in the maps at the other levels. So a point in the middle layer not only represents the number of shared words between a particular pair of Letter level Classes, but also represents the summation of shared words for the lower level Categories located below that Letter Class in the hierarchy (or *in* that Letter Class, in the body of the thesaurus).

The color ramp in Figure 18a differs from that in Figure 16a in that it is based on the standard deviation of word-counts for the Classes and Categories represented. The confluence of red points (> 2.5 Std. Dev.) highlights the same area of RIT as was identified in Figure 16a.

Figure 18b shows the top layer from Figure 18a. The call-out labels are Class labels and the labels with white backgrounds are Roman Class labels. For presentation clarity, only one of the pair of labels that would be attached to each point is used.

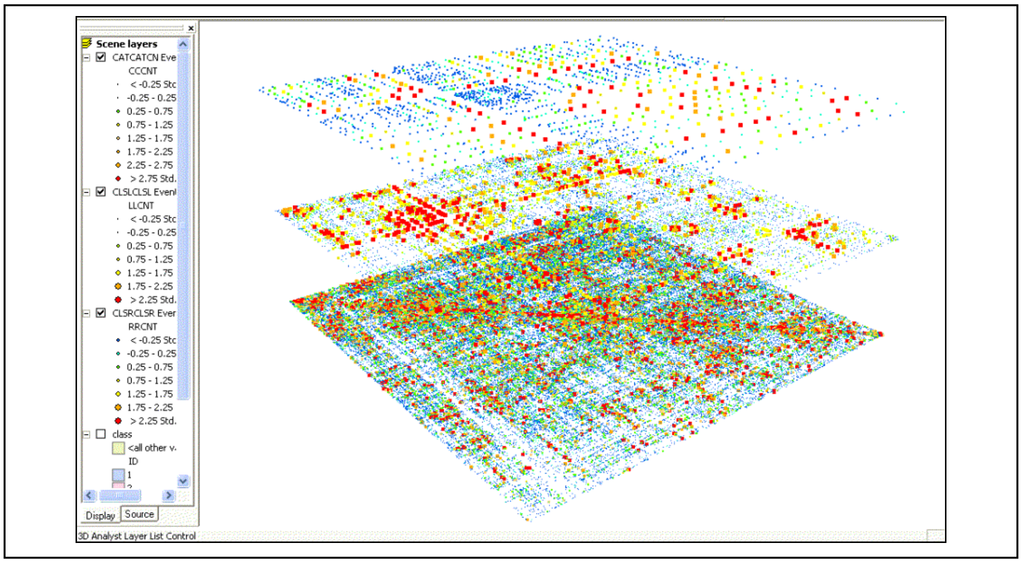


***Figure 18a*. Standard deviation of number of shared words between Categories, Letter Classes, and Roman Classes**

The red point closest to the label “motion,” represents 1,219 words (more precisely, instances of words represented as Entries) shared between Roman Classes 2:IV: Motion, and 2:III: Structure, Form. The blue points immediately above the “motion” label represent such relationships as between 3:III: Light, and 3:IV: Electricity and Electronics—40 shared words, only. The red point to the right of “voluntary action” represents 2,025 words shared between Classes 2:IV: Motion, and 7:III: Voluntary Action. Examples of some of the 2,063 words shared between Classes 2:IV: Motion and 6:III: Communication of Ideas are (in order of polysemy, or number of senses in the thesaurus as a whole): *cut, run, set, turn, head, charge, pass, close, line, beat*… The list includes the most polysemous words in RIT.

The main Cross-reference “authorities” (discussed in Section: *Cross-reference Patterns*) are 6:I: *Intellectual faculties and processes*, 349 links; 8:I: *Personal affections*, 348 links; 6:III: C*ommunication of ideas*, 281 links; 7:I: *Volition in general*, 231 links; and 2:IV: M*otion,* 213 links. The similarity to “most connected” Roman Classes, in Figure 18b, is striking.

The areas of high connectivity between Roman level Classes, and, going down the thesaurus hierarchy, between Letter level Classes, Categories, and Synsets, is due to such highly polysemous words as those given in the example. Except for those rare words that are homographs (having the same spelling but etymologically unrelated, such as “lead” [to command] and “lead” [the metal]), words carry with them some part of their essential semantics, no matter what company, in terms of synonyms, they keep, and no matter in what sense they are being used. If this were not so ambiguity would arise in word usage in text and speech. Context alone would not be sufficient to connect them to the particular sense in which they might presently be used. This assumption given, the connectivity among the groupings of words at different levels of the hierarchy reflects semantic connectivity.

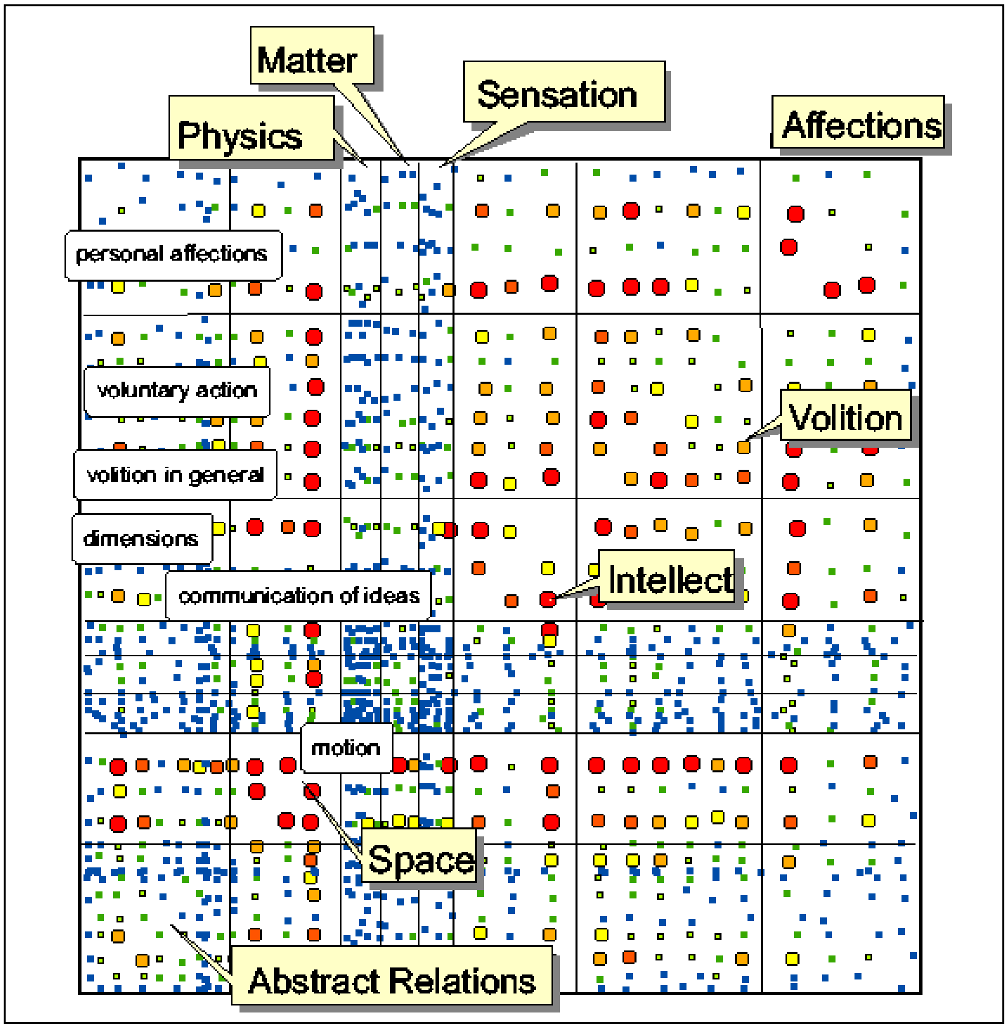


Figure 18b. **Standard Deviation of number of shared words between Roman level Classes**

Figure 19a shows the cross-references between Categories in RIT. Both the X- and Y-axes represent RIT Categories, and each point represents a cross-reference. A link between Category 1: Existence, and Category 3: Substantiality, will be in the bottom left corner; and a link between Category 1002: Lawsuit and Category 1007: Penalty will be found in the top right corner.

The points are graded in shade according to the type and number of cross-reference. Whole-Category reciprocated cross-references are darkest brown. Cross-references between adjacent or nearby Categories fall along the identity diagonal.

The Cross-references are not symmetric about the diagonal. Cross-references are directed from X to Y, so those linking to Categories with higher numbers will fall to the left of the diagonal, and those linking to Categories with lower numbers will appear below the diagonal. For example, a cross-reference between Category 4: Unsubstantiality and Category 1015: Specter, will appear in the top left corner; and a cross-reference between Category 1010: Atonement and Category 554 Disclosure will occur about halfway down the right hand side. A symmetric, reciprocated cross-reference from Category 554 Disclosure to Category 1010: Atonement, will occur in exactly the same position on the opposite side of the identity diagonal, halfway across the top of the figure.

The coherence *within* Classes (clustering of points indicating links or connections), seen in Figure 16a, is also evident in Figure 19a.

The plot of cross-references has been overlaid on a representation of the eight Classes (colored columns), following the X-axis orientation. The Class columns enable the identification of the clustering of cross-references within their own Classes. This clustering shadows the clustering in Figure 16a of words shared between Categories, supporting the hypothesis that cross-references reflect the overall semantic structure of RIT.

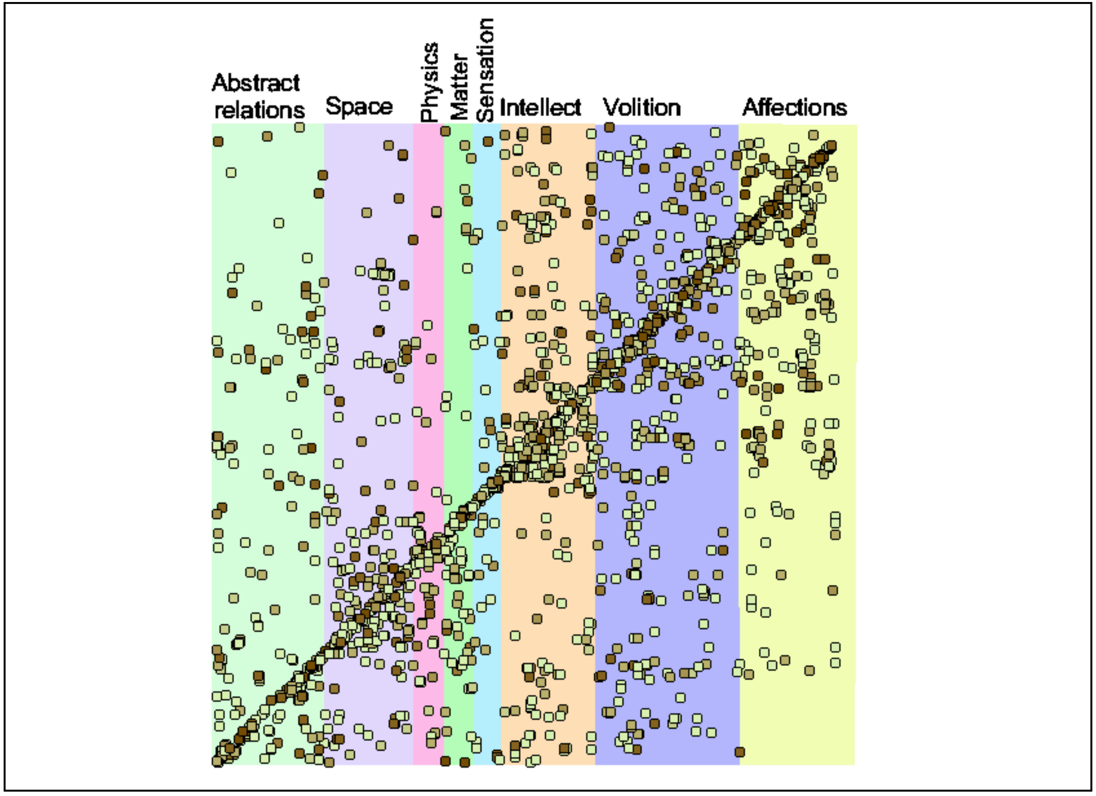


Figure 19a. **Cross-references in RIT. Colored columns show Classes**

Figure 19b is comparable to Figure 16b, a close-up of the cross phenomenon formed by high densities of shared words between Categories. Because there are points for only those members of the Category-Category relation for which there is at least one cross-reference, and cross-references are much sparser than polysemous words (those having more than one sense, and therefore more frequently shared between Categories), there are fewer points in this figure. The distribution of labeled Categories is similar to that found in Figure 16b.

A major difference between Figures 19a and 16a is that the cross is not evident. This may be because, although the Categories forming the cross in Figure 16a share many words, the senses of the words (the way they are used in speech and text) are so different that do not warrant a correspondingly high number of cross-reference links. Also, as noted in Section: *Cross-reference Patterns*, cross-reference sources are most frequently noun Synsets, and refer to senses or sets of senses that are nouns—including whole Categories, of which the labels are always nouns. The *cross* on the other hand, is dominated by verbs relating to notions of action, motion, volition, and travel.

Figure 20 shows the prepositions shared between Categories (312 words). This graphic contrasts with those given in the previous figures in this section in that it does not follow the motion/action semantic pattern that was evident in the other figures. Here the distribution suggests that the highest density of prepositions occurs in Class 1: Abstract Relations. The colors indicate the polysemy of the prepositions—grouped and color-ramped from red (high polysemy) to blue (low polysemy). Values falling on the identity diagonal are omitted—the points represent instances only of prepositions that are shared between Categories. Points occurring in the column labeled by "abstract relations" and in the row labeled by "Volition" represent words that occur as prepositions in both Class 1: Abstract Relations and Class 7: Volition.

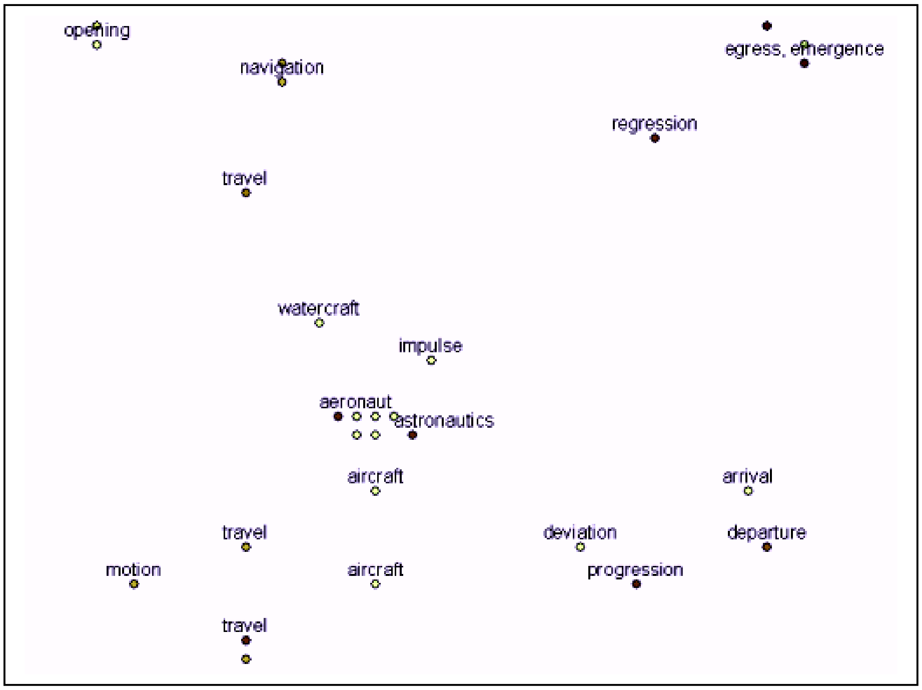


***Figure 19b*. Close-up of RIT cross-references in Figure 19a.**

The points in the square labeled “Abstract relations” represent words that occur as prepositions, and are shared between Categories occurring within Class 1: Abstract Relations, only.

The actual word that each point represents will be found in both Categories, but this does not preclude it from being found in other Categories (and, in this case, also as a preposition in other Categories). So point “X_1Y_1” may represent the word “over,” occurring as a preposition in two Categories, X_1 and Y_1. Point X_1Y_2 may also represent the word “over” occurring as a preposition, but between Categories, X_1 and Y_2, where CategoryY_1 may be in the same Class as X_1 (local, and so occurring near the diagonal, in the same grid) and Y_2 may be in a very different Class from X_1 (remote, and so not occurring in the same grid).

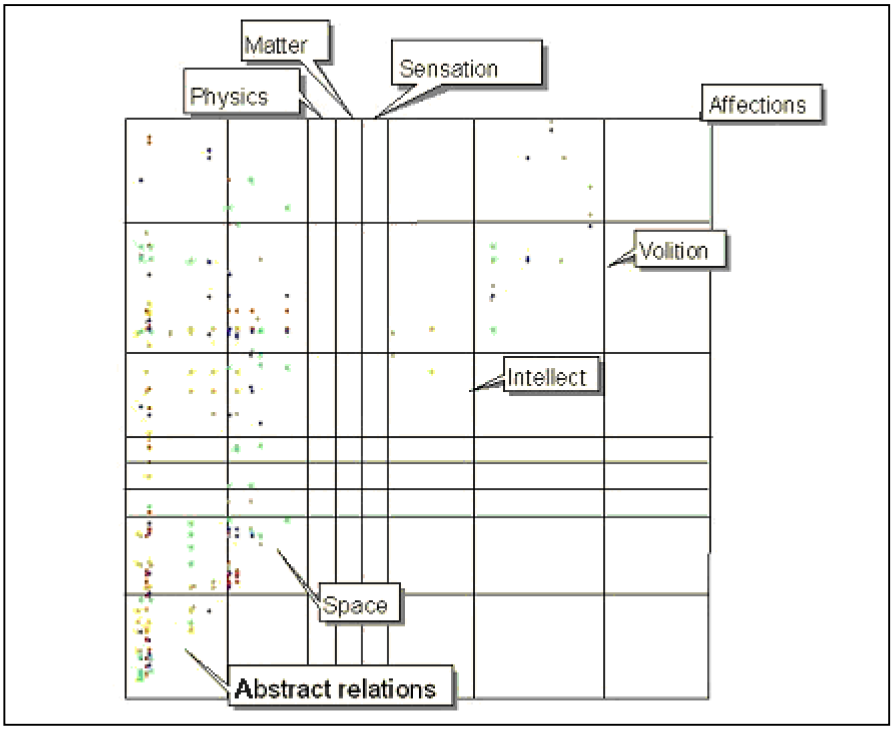


Figure 20. **Distribution of prepositions in RIT, color-ramped by polysemy**

Adverbs, although a much larger group (6,346 words), follow a similar pattern, probably reflecting the strong etymological ties between prepositions and adverbs. 44 percent of polysemous words that have prepositional senses also have adverbial senses, and many typical prepositions (such as off, up, and below) occur more frequently in usage, as adverbs.

Figure 20 shows that, despite the high density of words in other parts of the thesaurus, the distribution of this word class, or part-of-speech, follows semantic lines and is not drawn in by weight of numbers. This supports the earlier statement that connectivity among the groupings of words at different levels of the hierarchy reflects semantic connectivity. The high density in some areas of the thesaurus, or clustering, is unlikely to be the reflection of an industrious (or lazy) editor adding large numbers of words to localized areas of the thesaurus, while ignoring others.

WordNet

The WordNet model of the "mental lexicon" (Miller et al., 1993) has a quite different organization from that of Roget's Thesaurus, yet, as can be seen in Figure 21, the "change, move" notion still predominates, at least in the Verb section of WordNet. Figure 21 represents 60,000 verbs. The color-ramp is based on standard deviation of frequency counts (number of senses). Red indicates a standard deviation of greater than *2.5*. Points representing the most frequent verbs only, are labeled.

The relevance of the words and concepts in RIT related to "movement" and "change-of-place" is discussed elsewhere, for example in Old (2002), and the Section: *Cross-Reference Patterns*, above. The top cross-reference "authorities" (by fan-in) were 6:I: Intellectual faculties and processes, with 349 links; 8:I: Personal affections, with 348 links; 6:III: Communication of ideas, 281 with links; 7:I: Volition in general, with 231 links; and 2:IV: Motion, with 213 links.

WordNet's 25 top concepts, or *unique beginners* do not correspond directly to Roget's eight, top level Classes, or to the 43 Roman level sub-classes. However they do include *knowledge and cognition; feeling and emotion; communication; motive;* and *act, action, and activity*, which correspond closely in their semantics to the top cross cross-reference authorities and the top Roman Classes identified using counts of shared words, discussed in Section: *Global Views*.

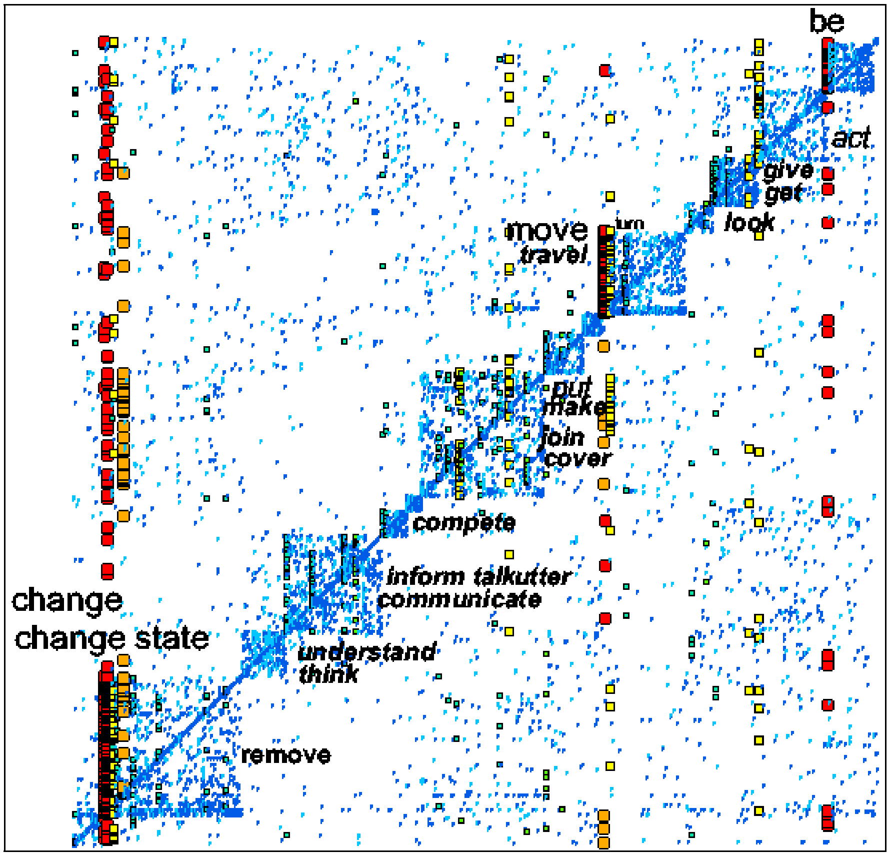


Figure 21. **Information map of the Entailment relation (Verbs) in WordNet**

Word Associations

Word association data is loose and probabilistic. Responses to cue words can vary by subject, as can the frequency of the words or *targets*, given as responses. The target words are not a fixed set as are the cue words (5,018, in the data set used for this research). Individual subjects may give unique (and sometimes puzzling) responses (see Appendix A: *Data Sources*, for statistics). However the weight of associations resulting from total subject responses can provide convincing data to support, or not, hypotheses about the structure of language semantics, and to act as a mirror to compare contrast the semantics found in the organization of RIT.

RIT has been described as an association dictionary and the common intuition gained from studying the arrangements of words associated by synonymy is that this true. In addition, as reported in Chapter 2, antonymy, common in word associations, is also an important relation in the thesaurus; at least in Dr. Roget's original structure. The following discussion compares the observations and conclusions about RIT made so far, with information and observations gained from analysis of the word association data.

Using the notion introduced by Bryan (1973) of Type-10 chains, or constraints on the connections allowed between words when modeling semantic connectivity, the word association data was processed to identify similarities and differences with Roget's Thesaurus. By selecting cue words and their target words as links, quartet-like quadruples were formed using the following constraints: A cue with at least two targets was matched with a second, different, cue that shared both targets. An example resulting from the selection of *over* as the first cue is shown schematically in Figure 22a.

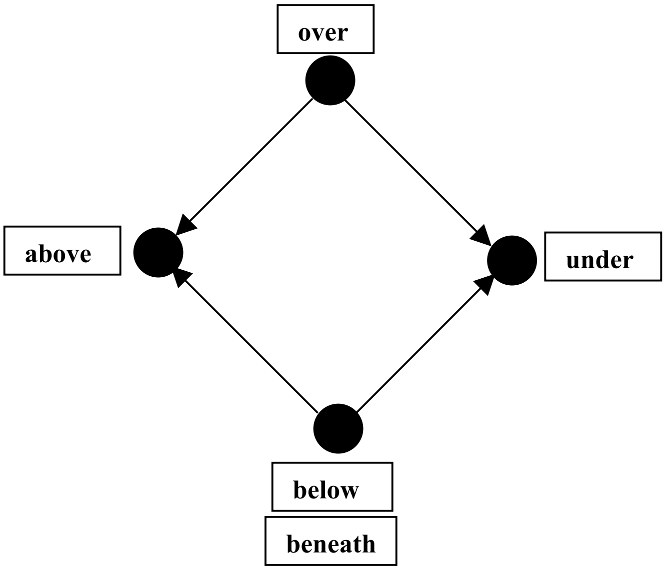


Figure 22a. **Word association quartet**

Over and *above* are clearly synonyms. This is a single pair—one record in the word association data and one link in the quartet. *Over* and *under* are clear antonyms. These two links constitute the first requirement or constraint—a cue with two targets. We are not concerned about reciprocal links, although they frequently do exist (about 25 percent of the cue-target associations are also found reciprocated by equivalent target-cue pairs).

Below and *above* are clearly antonyms; and *below* and *under* are clearly synonyms. These two links constitute the second requirement, or constraint.

Beneath is a third cue word, along with its two targets, *above* and *under*, which also satisfies the second constraint. Consequently the quartet formed by *beneath* with *over* overlaps with the quartet formed by *below* with *over*.

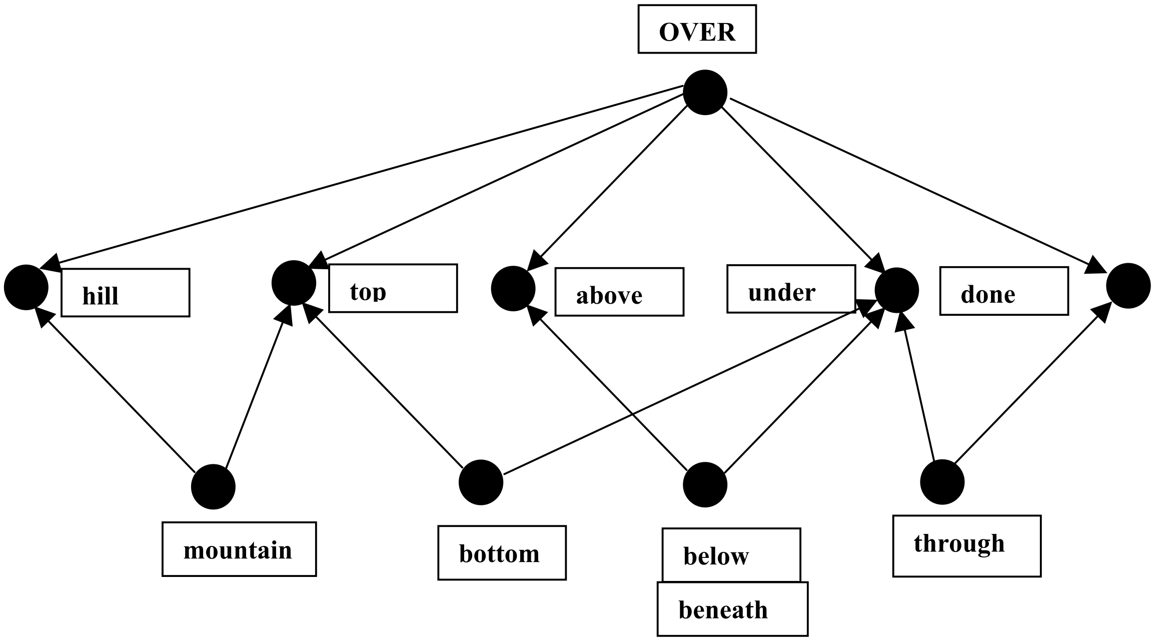


Figure 22b. **Cue-Target quartets formed with *over***

The (not exhaustive) collection of quartets, seen in Figure 22b, is formed by selecting *over* as the primary cue word, and illustrates some of the differences between the internal, or implicit, organization of the word association data, and that of RIT. Association is a single relation—no discrimination is made in the data between antonyms, synonyms, or meronymy (part-whole relations); all of which occur in RIT. Nor are word completions, phrase completions, alliteration, or other sound-similarity relations—none of which are found in RIT—discriminated in the word association data. Any lay, native English speaker can discriminate these relations, but they cannot be authoritatively discriminated by automated methods.

Hill is most likely a target of *over* because it forms part of the compound “over-hill.” Likewise, *top* forms a common phrase with *over*, “over the top.” These two words also share something in their essential semantics to do with “up-ness.” Also, “on top of” occurs in several instances as a synonym of “over” in RIT, as do “at the top,” “on the top,” “atop,” and “topside.” But although *over* is semantically associated with top in this way, there is no direct relation in RIT between *top* and *over*. *Top* instead occurs generally as a single word[28], classified as a noun. This is common for all compounds and phrases in RIT—they tend to be classified by their composite sense, and not by the meanings of their constituent word elements.

Over occurs 49 times in the word association data—eleven times as a cue (that, is as a cue word that elicited eleven different responses) and 38 times as a target (that is, it was the response to 38 different cues). Of the associated words, *above, beyond, through, again, finished, superior,* and *around* occurred as synonyms in RIT. These are just seven of the 21 synonyms of *over* in RIT that occur with *over* in more than one sense (excluding rare words that are idiosyncratic to particular senses). That is, about 15 percent of the word associations involving *over,* involve the synonym relation. The rest of the cues and targets associated with *over* were antonyms (three) and compound word completions such as *bent-over, hung-over,* and *over-turn*.

[28] Despite the fact that *top* occurs more often in RIT in nominal senses (as a noun), its Indo-European root, TAP(P)-, is verbal and means “strike”—hence “a *tap* on the *tuft* of the *toupee* where the *tip* of the *top* is, will *topple* him” (all of the italicized words are cognates). A British slang sense of *top* is to kill. *Top* possibly evolved its sense of “up-ness” from metaphoric use to do with “the place that is struck.”

For the word association data in general, about 23 percent of the word associations involve the synonymy relation. If words that <u>always</u> form word association pairs and that occur in RIT as synonyms are excluded (leaving only words such as *over* that have a variety of responses), 15 percent of the remaining associations involve the synonymy relation. Again, the rest of the associations are, excluding a small number of antonyms, completions.

Word-, compound-, or phrase completion is evident throughout the word association data. A sample of completions starting with “f,” follows:

- …
- fishing-boat
- first-place
- flower-garden
- fruit-juice
- …

Cues are often also targets of other cues. For example *over* is found in the response pairs *hung-over* and *start-over*; it is also found in *ending/over, adjourn/over,* and *through/over,* illustrating that there are other relations at work in the word association data. An example of two (overlapping) quartets containing *over* as a target are shown in Figure 22c.

“Over-turn and *turn-over”* is an example of a reciprocated (though not semantically symmetric) completion, where *over* is the cue in one pair and the target in the other. Two of the examples given above as completions also have reciprocal associations: *garden-flower* and *place-first*. Again, they are not semantically symmetric—order is important to

their meaning. *First-place* is an adjective-noun combination; *place-first* is a verb-adverb combination. There are many other examples in the database.

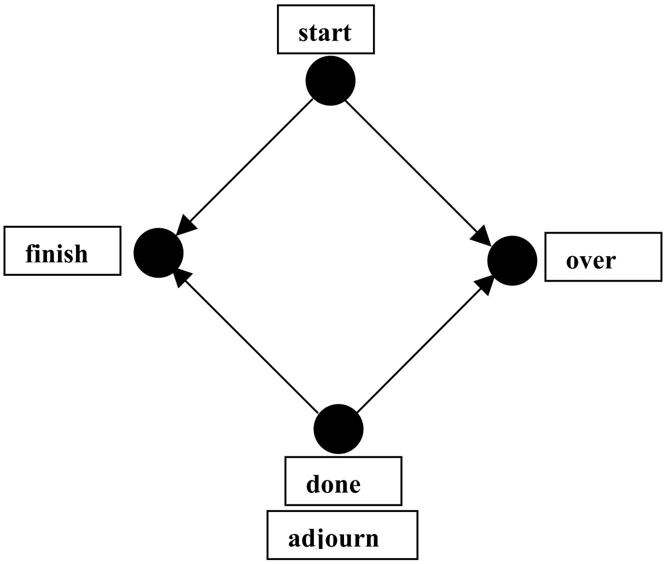


***Figure 22c.* Quartets containing *over* as a target response**

A novel relation found by comparing the word association data with RIT is between words that form parts of a compound, and that are also synonyms of each other:

- bath-tub
- class-room
- drum-beat
- together-with

From a global perspective, Steyvers and Tenenbaum (2001) have already shown that word association data form a small-world network, as does RIT. The brief analysis given here suggests that that the two networks are, however, not equivalent. The relations of synonymy, and to some extent, antonymy are shared in common, but a much richer set of relations is evident, implicit in the word association data. That is not to say that there are no other relations in RIT that can be derived; just that there is no explicit connection in the organization as there is in the cue-target links. By taking all of the words in RIT that occur as compounds (such as *street corner*, and *garden fence*), it is possible, by comparing the members of the pairs with instances of those words elsewhere in the thesaurus, to identify many of the relations found in the word association data.

The connectivity in Figure 22b is probably indicative of the small-world pattern at the local level. The target word with the most links in Figure 22b is *under*, which is the most direct antonym of *over* according to WordNet. As new quartets are added, for example (Cue 1) *over-top / under-on* (Cue 2), and (Cue 1) *over-top / under-bottom* (Cue 2), the links to *under* and *top* increase—the rich get richer, in the way described by Albert & Barabasi (2002) in describing the development of small-world networks.

Two final, minor observations: The cue words given to subjects did not get equal numbers of target responses. The word *field* received 34 different targets as responses, and *left*, only one (the response to *left* was always *right*). Similarly, target words as responses were not given equally frequently. The word *food* was the response given to 324 cues, while *shots* was the response to only one cue. The cue words listed by frequency, although full of highly polysemous words, bears no obvious similarity to RIT. The target list on the other hand, does. Many of the most frequent target responses were RIT Category labels. For example, among the top 20 targets (of 10,469) by frequency the first three are *food, money* and *water*; *animal* comes fourteenth, and *clothes* eighteenth. The top twenty RIT Categories, ordered by the number of words classified under their notions, begin with Animals, Food, Plants, and Clothing. Money, second in the word association list, occurs forty-ninth in the Category list.

Salton (1968) in his discussion on automatic thesaurus generation, has suggested that: "when words of unequal frequency are included in a thesaurus or represented on an association map…a hierarchical arrangement results almost inevitably, since frequent words can be made into categories, and words of lesser frequency into subcategories" (p. 57). His observation supports the hypothesis that the thesaurus hierarchy, at least to the Category level, could result naturally from the structure of word associations in the English language.

The final observation is that, by taking the quartets of associations for the whole word association data set and selecting the most strongly connected words (connected beyond their immediate targets to Type-10 chain-like connected words) the top words appear to have to do with what one might expect university students, the source of subjects in the word association study, to find important in their daily lives. The top twenty are: *rape, condemn, homework, study, class, student, family, school, essay, move, trauma, unpleasant, soup, violence, pastry, victim, medicine, desk, cake, cookie*… The word association data clearly reflects something about the experiences, mindset, or priorities of students from an American university. It is possible that these associations were uppermost in the minds of these subjects simply because they were being tested in this context, or that and they were "primed" for these associations by recent experiences such as conversations with their peers (in addition to the cue words that had been used in preceding cue-target probes).

The thesaurus “associations” found within the connectivity of the RIT Synsets possibly likewise reflect the mental workings and experiences of a nineteenth century British scientist, citizen, and Christian, Dr Roget. The use of archaic (to us) terms and concepts in the original edition, arranged in association with words that we consider “modern,” are one such example. The explicit organization, Classes and Categories reflect Roget’s classicist training, but many of the subjects clearly also reflect Roget’s world. Besides including such topics as *parts of sailing ships*, Roget also contrasts the *Word of God* and *Bible* (985: Revelation) with the *Koran* and *Vedas* (986: Pseudo-Revelation); and categorizes *Christianity, divinity* and *true faith* under 983: Religious Knowledge, while categorizing *Judaism, Buddhism*, and *Mahometanism* [sic] along with *heresy, idolatory, superstition* and *paganism* under 984: Heterodoxy (all examples are from the original 1852 edition).

Later editors of Roget’s Thesaurus have attempted to remove bias and prejudice, and to use more-politically-correct terms and Category labels, but will also inevitably have left residues of 20th Century beliefs and attitudes.

Core Connectivity Patterns

The small-world model suggests that data sets, or systems, that satisfy the small-world model are organized into a network topology of denser and sparser connections between elements such that the statistical distribution has characteristic values. Frequencies follow power curves and the paths connecting disparate points in the network are, on average, short because intermediaries can usually be found to connect any two points. Clearly, low connectivity points will tend to contribute only to local cohesion—though a single connected point may form a vital link joining two disparate areas of high density. Conversely, the most highly connected nodes and their associates can tie together vast areas, supporting or even creating global coherence, while providing no observable local cohesion—although a single highly connected point may be predominantly connected to points whose only connection to the network is that point. Strogatz (2001, p. 272) describes a small-world network as lying "somewhere between the extremes of order and randomness:" it has clusters and it has just a few "random" links that connect the clusters.

As an illustration, the genus and differentiae in lexical definitions, discussed under Section: *Part-of-Speech Patterns*, follow this pattern at both local and global levels. To the extent that Entries are associated as synonyms locally and as words globally; and to the extent that all frequency distributions in RIT follow power curves; the explicit connectivity in RIT also displays the features of the small-world model. Cross-references, which have dense linkages within divisions of the RIT hierarchy and sparse but significant linkages between divisions, also probably fit the small-world model, but this has not been tested statistically. The word-sharing illustrated between Categories in earlier Sections (but also existing between all levels of the RIT hierarchy) also demonstrates the phenomenon of local density, or cohesion, but with significant global sharing.

This Section examines local clustering and global patterns for the automatically derived, highly constrained, exhaustive *and implicit* Type-10 chains and components of Bryan (1974), Mooney & Talburt (1990), Talburt & Mooney (1990), and Jacuzzi (1991), discussed in Chapter 2, Section: *Models of, and Research on, the Structure of Roget's Thesaurus*.

TMC and VJC Components

Talburt and Mooney found that the majority of semantically strong connectivity within RIT formed one large component network of sense-sense and word-word quartets, partitioned from other component networks; and which is dubbed TMC-69. Jacuzzi fractured this crystal maize by applying a further constraint—that a quartet could not participate in a component if it shared only one entry with that component. While TMC-69 was a tightly inter-connected network of word and sense associations, the resulting derived VJC-184 is an extremely densely bound network—a core of the core connectivities of the largest Talburt-Mooney component.

To see what happens at the sub-component, Category, and Entry level a simple Jacuzzi component is chosen and analyzed first, after which the largest, most complex component is examined in less detail, but using the principles learned and drawing more general conclusions.

Talburt-Mooney Component 73 (TMC-73) has 246 entries composed from combinations of 60 words and 48 senses, all classified under 36 Categories, and all occurring as nouns. After processing by Victor Jacuzzi, 17 new components were derived from TMC-73. In Diagram 5 (loosely based on a graph from Jacuzzi, 1991) the structure, or inner topology of TMC-73 can be seen. The derived VJC components label the boxes. All of the TMC-73-derived VJC components are quartets, having four Entries, except those where the number of component Entries, in parentheses, follows the component label. The most important TMC sub-component of TMC-73, VJC-203, lies at the heart of the cluster with 142 Entries. To the right of VJC-203 and connected to it, is a ring of components {8416, 8540, 8493, 8737}. Below and to the left of VJC-203 is a second important TMC sub-component, VJC-3060, with 36 entries. Below VJC-3060 is a chain including VJC-77, with 8 Entries. VJC-5732, to the left of VJC-203, is an example of a single-quartet sub-component of TMC-73.

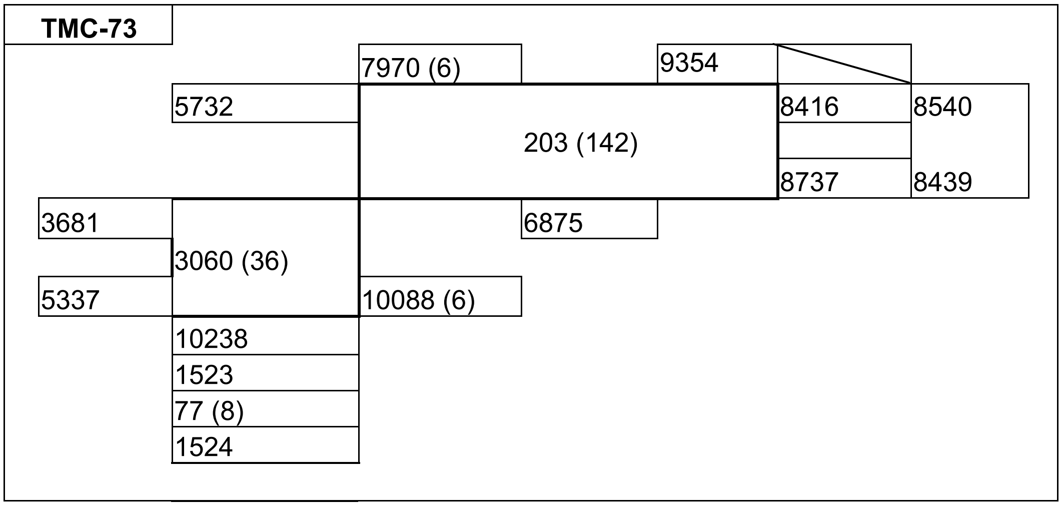


***Diagram 5*. Talburt-Mooney-derived-Component-73 with Jacuzzi-derived subcomponents**

To give a flavor of the semantics involved in Diagram 5, Diagram 6 shows the same components labeled by the words from the “corner-joins” that link the subcomponents together to form TMC-73. VJC-3060 and VJC-203 are labeled, in parenthesis, with a differentiating word indicating that they retain many words which do not participate in corner-joins and which are idiosyncratic to those components. VJC-203 and VJC-3060 form a corner-join, not labeled here. The corner-join involves Entry 988:3:1-*grossness* (a different instance of the word *grossness* from that labeling the VJC-5337 box).

Table 17 shows a cross-bar table (identical to a Formal Context) of VJC-77, an example of the structure of one subcomponent. VJC-77 has eight entries, as indicated in parentheses after the label in Diagram 5, or by simply counting the crosses; and six overlapping quartets—found by taking all combinations of four sets of crosses at a time.

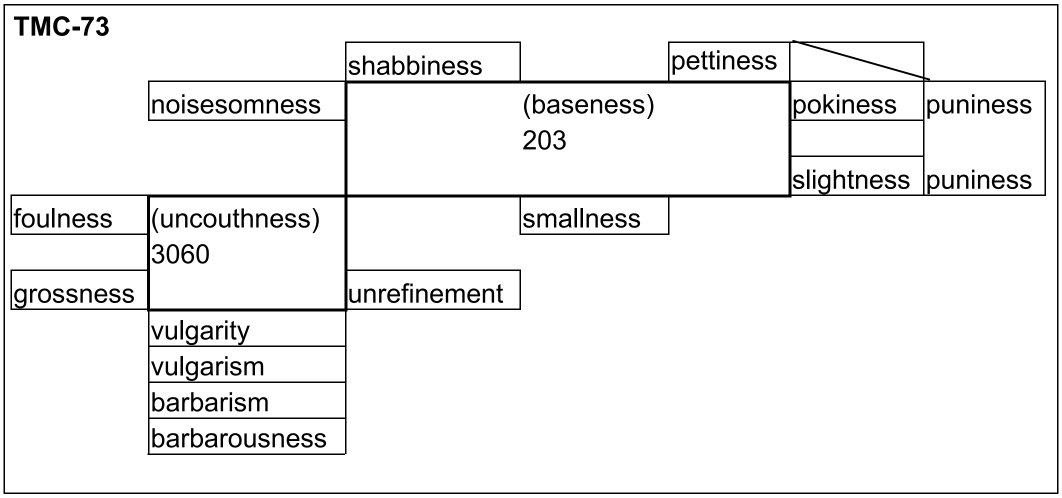


***Diagram 6*. Talburt-Mooney-derived-Component-73 with Jacuzzi-derived subcomponents, labeled by words from corner-join Entries**

Entry 588:1:6-barabarism (upper, bold-type **X**) is shared as a corner-join with Component 1523 (above VJC-77 in Diagram 5), and Entry 896:3:3-barbarousness (lower, bold-type **X**) is shared as a corner-join with Component 1524 (below VJC-77 in Diagram 5). *Gothicism* in all its senses is idiosyncratic to VJC-77; that is, it characterizes and differentiates VJC-77 from the other components. Synset 476:8:3, likewise, is found in only VJC-77.

VJC-77	476:8:3	588:1:6	896:3:3
barbarism	X	**X**	X
Gothicism	X	X	X
barbarousness		X	**X**

Table 17. **Cross-bar table of VJC-77 entries**

Similar sub-component structures, or topologies formed by corner-join connections of entries between Jacuzzi components, are found where ever a Talburt-Mooney component has been processed—with the same gradation from simple quartet-components to highly complex components; with an inverse relation between the number of entries found in a component and the number of components of that particular size; the same pattern of shared Entries versus differentiating Entries; and approximately the same probability of finding chains, rings, or more complex structures. That is, components display the same patterns and characteristics found among other elements of the Thesaurus, such as cross-references.

A Concept Lattice was purposely not used to illustrate this example to show that these features emerge independently of the representation used, and no matter what type of data set is selected from RIT. The TMC-73 sub-structure, represented in a lattice, can be seen in Figure 23. The chain (with Component 77) is in the upper left quadrant, the ring is in the upper right quadrant, the single-quartet components attached to Component 3060 are in the bottom left quadrant, and Component 203 with its attachments is in the bottom right quadrant. The Synset indexes of the Entries are displayed, rather than the words, and to simplify the lattice a restricted neighborhood is used; but the structure is the same. Synset Index 988:3:1, which contains the word *grossness* as part of a Quartet shared between Components 3060 and 203, links or holds together the two major components of TMC-73. The Categories range across the whole Thesaurus spectrum, from Category 35: Smallness to Category 988: Indecency, yet retain a kind of semantic coherence.

VJC-203, at the heart of the internal structure of TMC-73, has been pared down to a core by the Jacuzzi algorithm, yet still retains more than half of the Entries of the original TMC-73. The 124 Entries consist of words and senses found in 19 of the original 36 TMC_73 Categories. The top five VJC-203 Categories (and, incidentally, the top five Categories of all VJ components derived from TMC-73) by frequency of Entries, are: 913: Disrepute (24 Entries); 673: Badness (15 Entries); 863: Unpleasantness (14 Entries); 671: Unimportance (13 Entries); and 428: Unsavoriness (10 Entries). The remaining 31 Categories follow the theme of Vulgarity and Indecency. In other words there is a semantic coherence among the notions represented by the labels of the highest frequency Categories, and they collectively encapsulate the same broader theme that was evident in TMC-73.

The top words contributing globally to Entries in TMC-73, the original Talburt-Mooney component, were: *grossness* (13 Entries); *smallness* (11 Entries); *meanness* (10 Entries); *foulness* (10 Entries); and *vileness* (9 Entries). The top words for VJC-203, alone, were *meanness, smallness, vileness, baseness, and foulness.*[29] The highest frequency Entries derived from this main VJC sub-component, like the highest frequency Categories, encapsulate the broader theme of the original TMC-73 component.

***Figure 23.* Internal structure of TMC-73 (concepts are labeled below by Jacuzzi components, and above by Synset index numbers)**

The highest frequency words and senses found in VJC-203 provide the links that make VJC-203 highly externally connected with the other components by corner-joins, and highly internally connected by Type-10 chain links (internally, every sense has at least

[29] The reason that *grossness* is missing from the VJC-203 high-frequency list is that it occurred only three times within VJC-203—once as a bridge holding VJC-3060 to VJC-203.

two shared words, and every word has at least two shared entries—and there are no “weak” corner-joins).

The conclusion to this discussion is that the *highest frequency* Categories and words taken from a Talburt-Mooney component represent the semantics and connectivity of the component. They also retain the essential semantic flavor; and this is most concentrated in the central Jacuzzi component, if one exists, derived following the application of Jacuzzi’s constraint (the elimination of corner-joins). The relationship between the original Talburt-Mooney components and the derived Jacuzzi components is not so important as the fact that the Jacuzzi components are semantically concentrated, semantically coherent, and tightly bound internally, so that high frequency samples taken from elements of the component can be assumed to truly reflect the component as a whole. This is relevant to the discussion in the next Section.

VJC-184

As stated earlier and discussed in Chapter 2, Section: *Models of, and Research on, the Structure of Roget’s Thesaurus*, VJC-184 was the largest resulting Jacuzzi component derived from the largest Talburt Mooney component, TMC-69, after it was reduced to 2,507 smaller components. The second largest component was VJC-1501. This is illustrated schematically in Diagram 7.

While there are 5,960 components in the full TMC set, there are 10,341 in the final VJC set.

VJC-184 is comprised of 1,490 entries, composed of 324 senses and 312 words. The senses and words are classified in 153 Categories, or 253 Paragraphs, under 3 parts-of-speech. The Entries are divided between 868 nouns, 621 verbs, and 2 adjectives (*upset* and *put out*). This is a rich and mixed bag, but, if the implications hold from the analysis of TMC-73's reduction to VJC-203, VJC-184 contains the core of the core of RIT semantics.

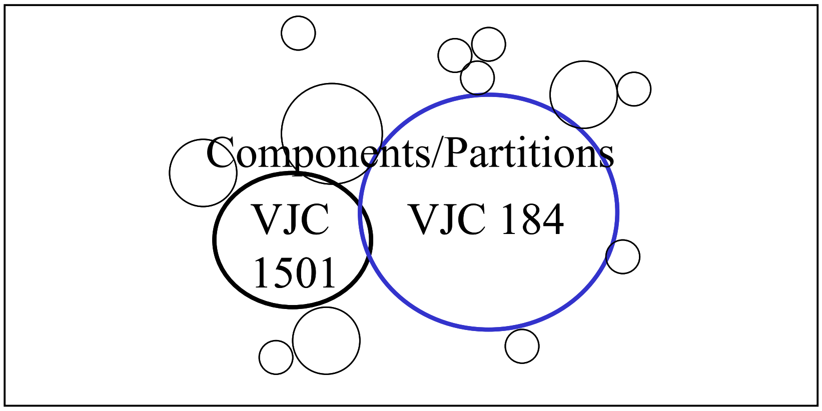


***Diagram 7*. Relationship among TMC-69 subcomponents (schematic)**

The most prominent semantic features of VJC-184 emerge not from the numbers but from the labels and words. The top labels of the top (most frequent) Paragraphs in VJC-184 are *agitation*, and *commotion*. The top (most frequent) Categories are 855: Excitement, 394: Stream, and 323: Agitation. "Stream" appears to be semantically incongruent with *excitement* and *agitation*. It lies between Categories 393: Rain and 395: Channel in the RIT Synopsis of Categories, so is certainly a classifier of "water" words. The semantic relationship becomes clearer on closer inspection; some of the 29 words occurring in VJC-184 that are derived from senses in Category 394: Stream, in descending order of frequency, are *flood, gush, run, surge, flow, deluge, rush, race,* *course*… These words, used in their (probably metaphoric) non-liquid senses are in fact congruent with *excitement* and *agitation*.

The top two synsets—those contributing the most words to Entries (and therefore to the Quartets, and connectivity) in VJC-184—are: 323:1:1 {fume, bluster, bustle, churn, commotion… (of 30 Entries)}, and 62:4:1 {row[argue], bluster, bother, brawl…(of 28)}. The overall most frequent words (out of the total of 312 words) begin, in order of frequency: *turn* (30 Entries)*, course* (20)*, run* (18)*, clash* (15)*, bother* (15)… These words correspond closely to the top words, by polysemy, in RIT.

The top (most polysemous, and therefore most frequent) 20 words in RIT are listed in Table 18. Eight are found most frequently in VJC-184. Three are found most frequently in VJC-1501 (the second largest sub-component of TMC-69, with 705 Entries[30]). The rest of the words are found most often in other components. The word "beat" is found in both VJC-184 and VJC-1501 as it occurs in a Quartet serving as a corner-join bridging the two components.

Not all instances (as Entries) of the most polysemous words in RIT occur in VJC-184. Only 30 of the 45 senses of *turn* occur here. It could be suggested that *statistically* the most polysemous words would inevitably find themselves into most components. This is possible but there are only 312 words in VJC-184, of 113,000 available in RIT. The concentration of these particular words, and the fact that they are the glue providing the

[30] VJC-1501 is characterized by words such as *cut, crack, hit, bust, gash, split, break…*

connectivity binding the component together suggests that this is their "home base" semantically.

Word	Polysemy	VJC Component
cut	**64**	1501
run	**54**	184
set	**51**	184
turn	**45**	184
head	**43**	184
pass	41	
charge	41	
close	39	
line	**38**	184
beat	**37**	184 + 1501
sound	37	
break	**36**	1501
check	36	
discharge	**36**	1501
drop	35	
cast	34	
go	**34**	184
lead	**34**	184
light	34	
form	**34**	184

***Table 18.* Top twenty words in RIT by polysemy with Jacuzzi components they occur in most frequently.**

An intuition about the semantics may be gained from listing these most frequent elements of VJC-184, but a different method is necessary to gain insights into the relationships among the elements.

Figure 25 is a lattice using only words and senses that occur in at least ten Entries of component VJC-184. There is a clear left-right division within VJ-184, connected in the middle by Synset 707:1:1 Haste (the label above the top black Formal Concept), through the sharing of *rush* (left black Formal Concept) and *flurry* (right black Formal Concept).

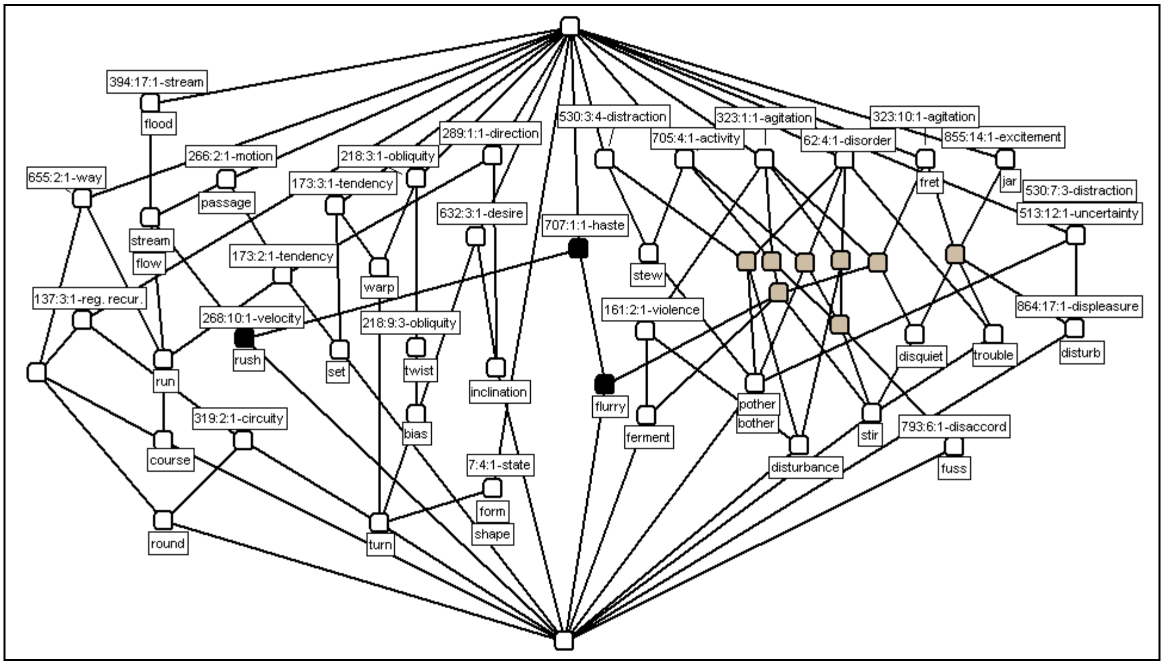


***Figure 25*. Lattice of VJC-184 restricted to senses and words with ten or more instances (contributing to at least ten Entries in RIT). Also visible are the "motion" cluster (left) and the "commotion" cluster (right).**

The left collection in Figure 25 has semantics characterized by *turn, run, course*/Stream, Motion, Direction. The right hand collection has semantics characterized by *fuss, bother, trouble*/Agitation, Excitement, Disorder. For brevity they will be referred to respectively as the "motion" and "commotion" groups, or clusters.

Table 19 shows how Synset 707:1:1 ties together the motion and commotion groups. Only the words shared between the two groups are used here. Not all of the relevant Quartets are included. In order to make explicit the Quartet formed by Synsets 268:10:1 and 394:17:1 with *rush*, a second word, *run*, which they both share, would need to be added to the table. *Run* does not occur on the *commotion* group so it is omitted from Table 19.

	Agitation	Agitation	Activity	Haste	Motion	Velocity	Stream
	323:1:1	323:10:1	705:4:1	**707:1:1**	266:2:1	268:10:1	394:17:1
bustle	X		X	**X**			
Dash				**X**		X	
Flurry	X	X	X	**X**			
Flutter			X	**X**			
Rush				**X**	X	X	X
Scamper				**X**		X	
scramble				**X**		X	
Scurry				**X**		X	
Scuttle				**X**		X	

Table 19. Cross-bar table showing the connections (via Synset 707:1:1 Haste) between the *commotion* (left) and *motion* (right) groups of VJC-184

Unlabeled Concepts

The *commotion* group (right side of the lattice) displays a feature hidden by any other form of representation. When two Objects share two Attributes in a Formal Concept lattice, but in addition each of the four elements is differentiated by further Objects and Attributes that are not shared among the other three elements, an "unlabelled Concept" emerges. These are the Concepts colored in gray, in Figure 25. While unlabeled Concepts are always rare in lattice representations of semantic data, an entire cluster of unlabeled Concepts is extremely rare[31].

In Boolean lattices, where every permutation of Attributes is used, a trellis-like structure appears with only the outer-most edges labeled. The central array of Concepts represents

[31] Unlabelled concepts appear in the Concept Lattice of *agitation*, illustrated earlier. But it will be seen that that lattice is simply a facet—a slice—of VJC-184, and that the unlabelled Concepts are the same.

the exhaustive, individual permutations. In more day-to-day examples, unlabeled Concepts are an indication that an Object (label) does not exist for every combination. In terms of the semantics of RIT it means that very fine discriminations have been made, but not all combinations have been given an individual label. This method has been used to identify "missing concepts," or lexical gaps between languages by creating lattices where the words in one language are taken as Objects, and the words of the other language are taken as Attributes (Janssen, 2002, p. 179). Where a concept exists in one language and not in the other an *unlabelled Concept* emerges.

An analogy for what might produce a cluster of unlabeled Concepts would be to classify blue and green colors by fine-but-overlapping divisions of the spectrum, and to specify the shades, tones, and hues of colors by name, in detail. For example, {… *green, blue-green, green-blue, blue, azure, algae, aqua-marine, light aqua-marine, duck-egg, sea-blue*...}. It would be possible to uniquely identify any color by wavelength, but if no name for it exists, or can be constructed using an adjective or noun compound, it remains an unlabelled concept—possibly below our perceptual threshold. The cluster of unlabeled concepts in the commotion group suggests a notion that is not fully reified, but is represented by a large number of words with overlapping hues and tones of meaning.

Four of the words in the "commotion" group that contribute to emergent unlabeled Concepts are *flurry, ferment, fuss,* and *stir*. These all classify under 705:4:1 Activity, and 3:1:1 Agitation They words are also classified under other senses, but in different combinations (subsets); or independently of each other, alongside other words. *Flurry* is also found in 707:1:1 Haste alongside *rush*; and *ferment* is found in 161:2:1 Violence, alongside *disturbance,* for example. This is explicit in the lattice in Figure 26, which represents the *commotion* group, separated from the rest of VJC-184 at Synset 707:1:1 Haste (and still restricted to senses and words contributing to at least ten Entries in VJC-184).

The Concepts in Figure 26 are numbered. For example, Concept #17 represents 3:1:1 Agitation, 705:4:1 Activity, and the words they share, {*bustle; turmoil, tumult, bluster, embroilment, commotion; hubbub; disturbance; ferment; fuss; flurry; stir*}. What is "Agitated Activity?" What is the difference between that and "Agitated Disorder" (Concept #19) involving most of the same words but omitting *turmoil, tumult, bluster, embroilment,* and *commotion*? It appears that the underlying concepts, represented by these words in their various combinations, are somehow non-specific or malleable. Like minute shades of blue-green, the blue-green is easily perceived, but the tones and hues are difficult to discriminate or identify except when in contrast to each other. And yet they are there.

It can be seen in Figure 26 that Synset 323:1:1 Agitation holds sway over the majority of unlabeled Concepts. Synset 3:1:1 contributes the most Entries to VJC-184 (all nouns). Three of those Entries link to other Jacuzzi components in the original TMC-69. Synset 323:1:1 is, however, not the key to the nest of unlabeled Concepts which is further linked to many, many other words and senses omitted from Figure 26, and which also hold this mesh together. If Synset 323:1:1 (or any individual Synset) is removed, other senses such as 161:2:1 Violence, 705:4:1 Activity, 62:4:1 Disorder, and 323:10:1 Agitation (the verb contribution from Category 323) would continue to hold the structure in place. Like a single strand plucked from a spider's web, the web distorts but mostly holds in place—similarly if words are removed.

Some of the dense connections seen in the VJC-184 (and other components) are comprised of apparently etymologically unrelated words that in fact share common Indo-European roots. Examples from VJC-184 are: *flood, fluster, flutter, flight,* and *flow*, which all derive from the root, PLEU-, meaning "flow." *Warp, pervert, wring,* and *wrench* all derive from the root, WER-3, meaning "turn, bend"—and there are others. Such ancient etymological threads may explain why some Synsets are so large, and why they interconnect so readily. Etymology alone cannot explain the cluster of unlabeled Concepts, however, as no single Indo-European root pervades that group. The underlying concept perhaps can be explained by proposing that it is an ancient concept at the root of human conceptual organization—not the *central source*, although it can be traced out to connect to more than 70,000 entries, but one of several primitive concepts possibly more felt than intellectualized, and a facet of consciousness connected to many other areas of thought.

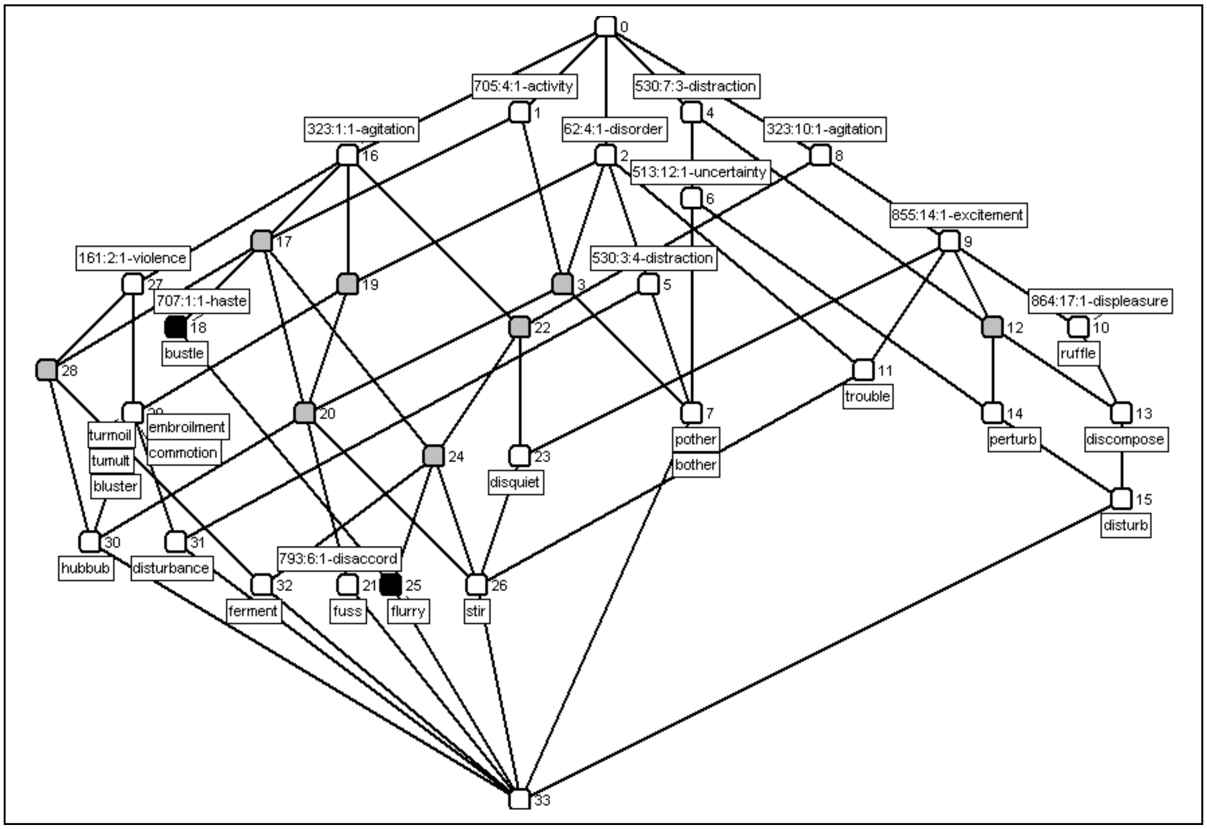
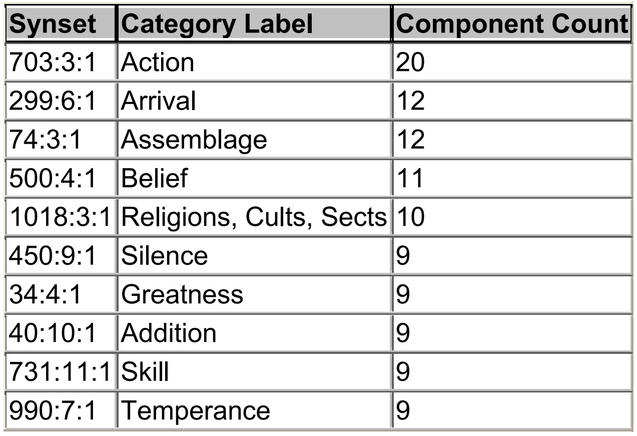


Figure 26. **The *commotion* group, only, of VJC-184**

Conceptual Switching Centers

It is not the hubs or authorities that necessarily provide the short path lengths in the small-world of RIT. Discussion in the previous Section showed that it can be the *almost random* links–the corner-joins such as Entry 988:3:1-*grossness* of TMC-73, and the internal linking Quartets such as [707:1:1 Haste - *flurry* - 323:1:1 Agitation] and [707:1:1 Haste - *rush* - 268:10:1 Velocity], in VJC-184–that tie the network together. These *almost random* links are not referred to here as simply *random* links because they cohere semantically with their neighbors.

On the other hand Synsets with large numbers of words (high synonymy), and highly polysemous words are the most likely to form Quartets, the links forming local clustering. High polysemy words are also the most likely to form the *anchor text* of the semantically-strong, whole-Category cross-references. The highest-polysemy words were listed in Table 18 and discussed in relation to the largest, most tightly knit Jacuzzi components. The top ten Synsets that contribute the most connectivity among the components (as opposed to connectivity that weaves and binds the elements of a component together internally), by *component count*, are given in Table 20.



Synset	Category Label	Component Count
703:3:1	Action	20
299:6:1	Arrival	12
74:3:1	Assemblage	12
500:4:1	Belief	11
1018:3:1	Religions, Cults, Sects	10
450:9:1	Silence	9
34:4:1	Greatness	9
40:10:1	Addition	9
731:11:1	Skill	9
990:7:1	Temperance	9

***Table 20.* Top ten inter-component linking Synsets, by component count.**

Some examples from the Synsets in Table 20 illustrate that these Synsets contain highly polysemous, common words: 500:4:1 Belief {…*feeling, idea, notion, concept, thought*…}; 1018:3:1Religions, Cults, Sects {… *group, body, community, society, faction, order*…}; 450:9:1 Silence {…*mute, kill, smother, drown, choke, stifle, throttle*…}.

40:10:1 Addition was discussed in Section: *Part-of-speech Patterns*, under *Genus and Differentiae*. It is an adverbial sense that shares words with many of other senses. In the thesaurus it has 37 Entries. Of the 37 Entries, 24 are words that have more than one part-of-speech, and 31 are polysemous. Of those with more than one part-of-speech, 14 double as adjectives, 12 double as prepositions, 4 as nouns, and 3 as verbs. It contains most of the synonyms of *over*, and many of the other core prepositions.

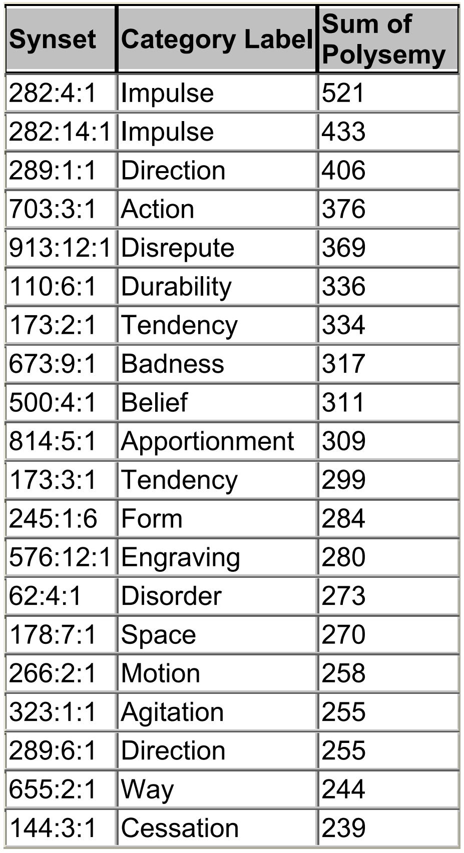
These Synsets, like the high-density components, bind the RIT small-world together. They may be viewed as acting as bridges between areas of high connectivity. But unlike cross-references they have no inherent directionality—they cannot be called hubs or authorities—so they are called here, *conceptual switching centers.*

Switching centers gain their power from the polysemy of their associated words. A Synset that contained a large number of words, where each word had only one sense, would contribute nothing to RIT connectivity. Calculating the power of a Synset (the degree to which it may be considered a *switching center*) from the sum of the polysemies of its words results in an ordering on the Synsets independent of the number of components that they link together. The top 20 switching centers, by summed polysemy, are shown in Table 21.

There are 71,398 Synsets altogether in RIT. Their summed polysemies range between 1 and 258. The top ten component-linking Synsets given in Table 20 are all in the top 100 switching centers in Table 21. Synset 703:3:1 Action, top of the list in Table 20, is fourth on the list of switching centers in Table 21.

The top two Synsets in Table 21, both from Category 282: Impulse, are from the core of the dense *cross* formed by large numbers of shared words between Categories, discussed earlier in Section: *Global Views*. Synsets 289:1:1 (third) and 289:6:1 (18th) from Category 289: Direction, are also from the *cross*.

Synset 323:1: Agitation, overlord of the *commotion* cluster from component VJC-184, is 17th on the switching centers list, and Synset 62:4:1 Disorder, which shares most of the words with 323:1:1 is 14th. Synset 913:12:1 (fifth on the list) is a lynchpin Synset in VJC-203 (Figure 23). Seventh and eleventh on the lists are Synsets from Category 173: Tendency—both lynchpins in the *motion* cluster of VJC-184.



Synset	Category Label	Sum of Polysemy
282:4:1	Impulse	521
282:14:1	Impulse	433
289:1:1	Direction	406
703:3:1	Action	376
913:12:1	Disrepute	369
110:6:1	Durability	336
173:2:1	Tendency	334
673:9:1	Badness	317
500:4:1	Belief	311
814:5:1	Apportionment	309
173:3:1	Tendency	299
245:1:6	Form	284
576:12:1	Engraving	280
62:4:1	Disorder	273
178:7:1	Space	270
266:2:1	Motion	258
323:1:1	Agitation	255
289:6:1	Direction	255
655:2:1	Way	244
144:3:1	Cessation	239

***Table 21.* Top 20 Synsets by summed polysemy—the *conceptual switching centers*.**

The central adverbial/prepositional Synset, 40:10:1 Addition, is “only” at position 41. But this is relative to the full list of more than 71,000 Synsets.

Switching centers are an indicator and probably (although untested statistically) predictor of high-connectivity features in RIT, such as components. They are also frequently, as seen in Table 20, bridges between areas of high connectivity. They are the explicit evidence of probably what makes Roget’s Thesaurus display the characteristics of a small-world network

Networked Words

Not much has been said thus far regarding *words*, the substrate of Roget’s Thesaurus, specifically. The following, some of which has been mentioned in passing in previous Sections, summarizes the characteristics of words in RIT. Words exist in the Thesaurus as Entries—each senses or instance of the word being a separate entry—each instantiation potentially having a different part-of-speech. Words co-occur in senses (or *Synsets*), usually as synonyms, but occasionally as members of a list. They occur predominantly at the Synset level, but also as labels for Paragraphs, Categories, and Classes. At the Synset level they represent a shared concept—a specific idea defined somewhere at the intersection or overlap between the members of the Synset. At higher levels of the hierarchy they represent more abstract notions—the higher the level the more general the notion.

Words co-occur as antonyms in adjacent Categories. They are classified within Categories by part-of-speech. They serve as anchor text for cross-references, indicating the essential semantics of the *source* end of the cross-reference. They are character strings of a particular set and order, or spelling, that may (though rarely) coincide with that of a second, etymologically unrelated, word known as a *homograph*. They may share Indo-European root ancestry, in which case they are called *cognates*. They are the complement of senses in the organization of RIT.

Finally, words relate to each other in some ways that can only be made explicit by automated means. The rest of this Section looks at some of those ways that words come together to form the conceptual web of RIT.

Implications between Words

Words can form implications. As discussed in Section: *Set Implication*, an implication or inference exists between words that share Synsets, such that the set of Synsets of one word is a subset of the set of Synsets of the second word. In such cases the word with the subset of senses is more specific and implies the second, more general word—implications flow from the specific to the general.

The examples given in Section: *Set Implication,* were of simple pairs of words between which implications existed. The implicit organization of the Thesaurus is, however, an elaborate semantic topology, not a set of lists. The implications between words are, likewise, much more elaborate. For example *twaddle*, which has four senses occurring in two Categories, 545: Meaninglessness and 594: Talkativeness, shares these senses with *babble* and *jabber*. *Babble* and *jabber* have other senses in Categories 472: Insanity, Mania, and 578 Language, respectively (amongst others). They have more senses than *twaddle,* and those senses are a superset of the senses of *twaddle*.

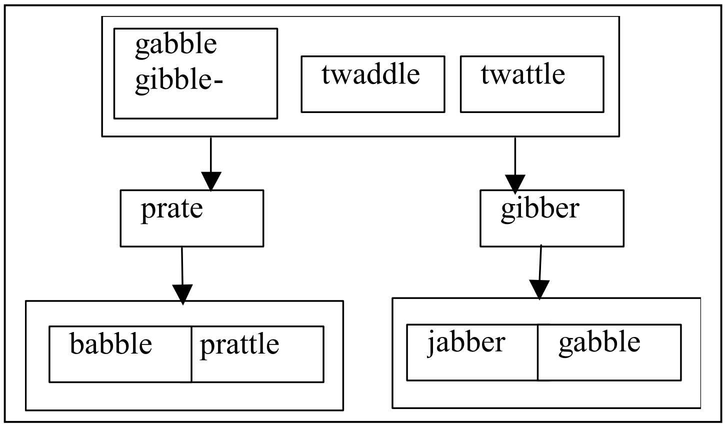


***Figure 27.* Transitive implications among a sample of RIT words**

Furthermore, there are words, such as *gibble-gabble* and *twattle* that share exactly those senses of *twaddle*; there are words that share an intermediate number of senses between *twaddle,* and *babble* and *jabber,* such as *prate* and *gibber*; and there are words that share senses with various combinations of the given words, although this example will omit those from the discussion in order to reduce the complexity of the relationships, and the discussion.

The implications between these words are shown in Figure 27. *Twaddle* (or *gibble-gabble*, or *twattle*) implies or infers *prate,* which in turn implies *prattle*, or *babble*—or both. This is a transitive relation (if *A* implies *B* and *B* implies *C,* then *A* also implies *C*), so *twaddle* also implies *babble*.

Figure 28 shows the words from Figure 27, elaborated with their eighteen Synsets from RIT.

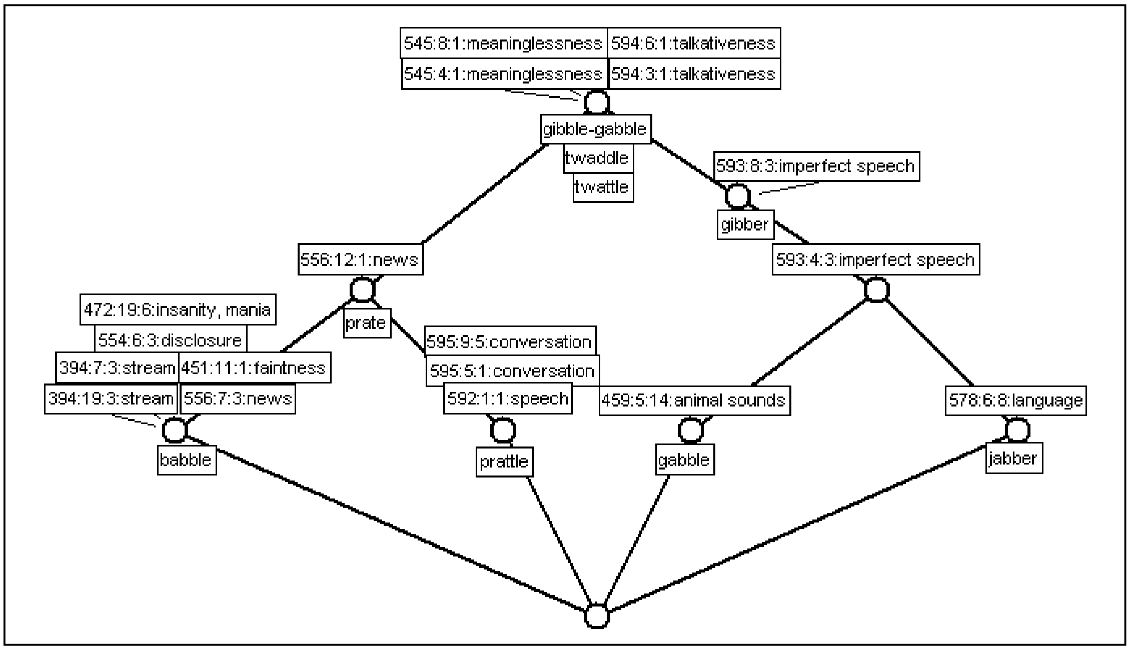


***Figure 28.* Transitive implications among a sample of RIT words, along with their related senses**

The most general words—those with the most senses—are found at the bottom. The Synsets shared by all of the words (found in Categories 545: Meaninglessness and 594: Talkativeness) label the top Concept.

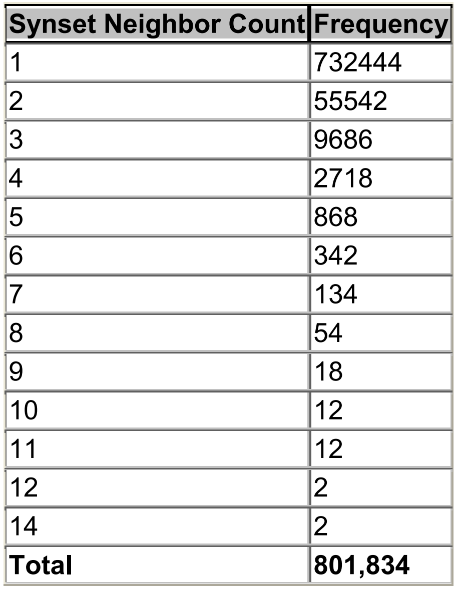
There is a symmetric organization among the senses compared to that of the words. The Synsets at the top are the most general. The most specific senses, those shared by the fewest words, are found at the bottom. In other words, the Synset Indexes of the most specific senses label the same concepts as the most general words, and Synset Indexes of the most general senses label the same concepts as the most specific words. This symmetry between words and senses (or Objects and Attributes, in Formal Concept Analysis terminology) is known as the *Galois connection*. It is beyond the scope of this discussion to elaborate on the Galois connection, but it can be stated that anything that is said about the organization of the RIT words, in the context of neighborhoods, is symmetrically true for the senses.

There are a further 51 words involved in the full lattice of Figure 28 if the neighborhood of *twaddle,* alone, is taken into consideration. Those words, such as *blether, chatter, palaver, gush, spout,* and *gab* form more-complex relationships, where for example, *pairs* of words can form implications.[32] If all of the words from the eighteen senses in Figure 28 are included, 93 words in all are pulled in as neighbors. In short, the complex relationships, interconnections or associations among words in RIT may be summarized in principle, but the complexity is pervasive in every facet.

Word Co-occurrence, Genus and Differentiae

The complexity observed at the word level throughout the Thesaurus emerges from a relatively simple relationship—co-occurrence of words in a Synset. Table 22a shows the frequency of word co-occurrence counts (*Synset neighbor count*) with other words in RIT Synsets.

[32] All of the examples given were single-premise inferences. This does not exclude two or more premises, but the discussion here is limited to the former.

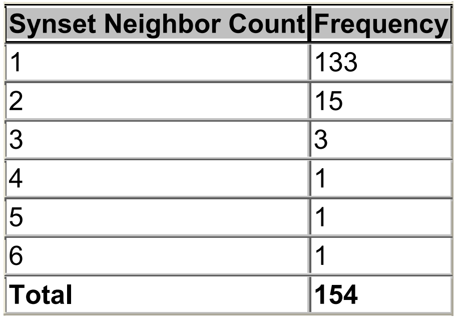


Synset Neighbor Count	Frequency
1	732444
2	55542
3	9686
4	2718
5	868
6	342
7	134
8	54
9	18
10	12
11	12
12	2
14	2
Total	**801,834**

***Table 22a*. Word co-occurrence (*Synset Neighbor* Counts) frequency. Most words co-occur together only once. Two words occur together fourteen times.**

In explanation of how Table 22a was calculated, word co-occurrence for *over* was counted in the following way: *over*, which has 154 synonyms, is counted 154 times, in total, in the Frequency column of Table 22a. *Over* shares six of the seven senses of *above,* so *it* occurs once as one of the 342 words counted in the Frequency column as participating in a word pair that shares six Synsets (having a Synset neighbor count of 6). For the same reason *above* also occurs as one of the 342 words sharing six Synsets with a second word. *Over* co-occurs only once with 133 of its synonyms, so *over* is counted 133 times among the 732,444 single pairings (the first row of Table 22a). Table 22b shows the Synset neighbor count frequencies of *over*, specifically.

It is important to note that the once-only co-occurrence of some words with *over*, or of any pair of words, does not necessarily mean that one or both is a singleton (a word with a polysemy of one). Many of the 133 words that co-occur only once with *over* are very polysemous words. Examples are, *end*, with a polysemy of 18; *after*, with a polysemy of 13; and *close*, with a polysemy of 39. None-the-less, within the *semantic context* of *over*, they act as differentiae. About half of the single co-occurrences of words do involve singletons, and about 25% involve *only* singletons.



Synset Neighbor Count	Frequency
1	133
2	15
3	3
4	1
5	1
6	1
Total	**154**

***Table 22b. Synset Neighbor* Count frequencies for the word *over*. *Over* occurs with one other word, *above*, six times.**

The significance of the frequency count in Table 22a is that most words (91%) occur together only once; a further 7% occur together only twice; and 1.2% occur together only three times. That accounts for more than 99% of the relationships between words in RIT. The last row in Table 22a represents one pair, *keen* and *sharp*, which occur together fourteen times—the most of any word pair—yet *keen*, with a polysemy of 20, has another 6 senses that do not include *sharp*; and conversely, *sharp* has a further 17 senses.

These figures exist in the context of the fact that, between the 113,000 words in RIT there are almost 13 billion possible combinations, or pairings. The 801,834 actual pairings could have been produced by fewer than 900[33] words paired in all possible combination.

[33] 900 X 900, less pairings of words with themselves = 809,100

In this light the Word-Word network implied by the Synset neighbor count frequencies is very sparse.[34]

The two keys to the mystery of how these simple word combinations produce the complexity observed in RIT are *genus and differentiae* and the properties of small-world networks. As was shown in the lattice of *over* in Section: *Part-of-Speech Patterns: Overlap among Parts-of-Speech*, the senses of a word can be mostly discriminated by simply using those words that co-occur more than once with the *topic word*. This is because there is redundancy in the pairings. For *over* this means that the 21 repeating Synset neighbors (154, minus the 133 single-instance differentiae) are able to completely differentiate 14 of *over's* senses. A further two pairs of senses (four Synsets) are partially differentiated: 36:13:1 Superiority with 206:24:2 Height, differentiated by {*above, over*}; and 119:5:1 Past with 198:21:2 Distance, by {*past, over*}.

In general, four Synset neighbors alone could unambiguously differentiate 15 senses for any *topic word*. This is illustrated in Figure 29. The topic word in Figure 29 labels the bottom Concept because it occurs in every sense. Sense 1, labeling the top Concept, is a Synset consisting of all of the words in the lattice: {*topic word*, *word 1, word 2, word 3, word 4*}. Sense 3 contains only the topic word, and Words 2, 3, and 4; Sense 4 contains only the topic word and Words 1, 3, and 4; and Senses 12, 13, 14, and 15 are Synsets each containing the topic word and only one of the other four Words. The idea that polysemous words (and symmetrically, synonymous Synsets) alone are sufficient to

[34] Even if the co-occurrence frequencies are expanded, there is little increase. For example the 342 Synset Neighbors occurring together six times, expanded, would be 2,052 pairings. The total for all such pairings is still only 891,686.

disambiguate senses was also illustrated in the lattice of the *commotion* cluster of VJC-184, where all words had multiple senses and all senses contained multiple Entries.

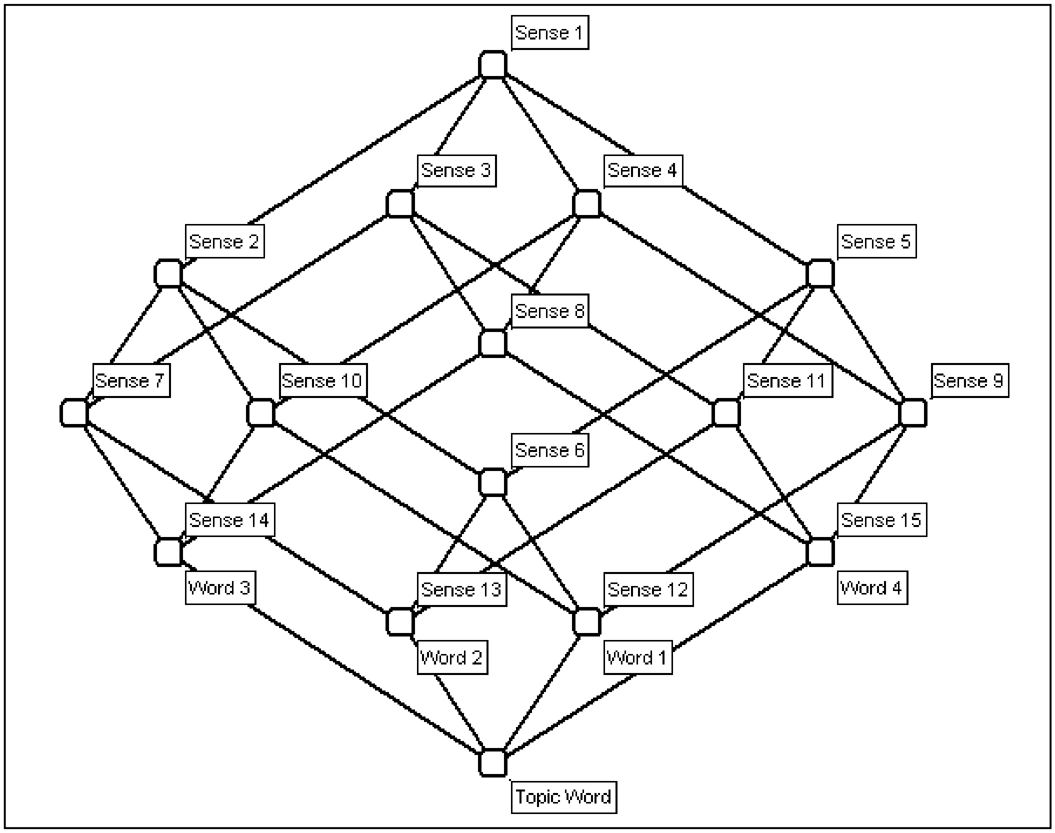


***Figure 29.* Lattice of a hypothetical topic word (four Synset neighbors are sufficient to differentiate 15 senses of a topic word unambiguously)**

When speaking of the sense of a common, polysemous word it is sufficient to use one other word, however polysemous, to distinguish, discriminate, or differentiate it. For example, “*over*, in the sense of *above*”; “*over*, in the sense of *across*”; or “*over*, in the sense of *extra.*” Combinations can just as easily differentiate, or resolve the ambiguity of the meaning of a word—perhaps more so. For example, ”*over*, in the sense of *beyond* or *past*.”

Of course the semantic context, or, in the case of RIT, the Category label, also effectively differentiates the sense of a word. For example, *over* or *above*, in the sense of Excess (Category: 661); and *over* or *above,* in the sense of Covering (Category: 227). The Category labels are hypernyms (broader terms), or the genus of the words used here. They are the "plant" in the definition of a cactus, rather than the "has prickles" or "lives in the desert."

On average, those Synsets in RIT that do contain monosemous, single-instance differentiating words in addition to combinations of polysemous Synset neighbors, contain 1.8 of them. If two differentiae, or singletons, were added to each sense in Figure 29, a further 30 words would be added to the lattice. If in addition there were words labeling each of the senses, together with the differentiae, this would yield a Synset neighbor distribution similar to that seen in Table 22a. The distribution for averages of all RIT words by ranges of polysemy, including counts for words co-occurring with the topic word up to six times, is shown in Table 22c. The hypothetical topic word in Figure 29 has 15 senses and would fit in the middle range. *Over*, whose distribution is shown in Table 22b, has a polysemy of 22 senses, so would fit in the high range.

Synset Neighbor Count	Polysemy < 10	Poly.14-17	Polysemy > 20
1	7	52	100
2	2.5	8	15
3	1.5	2.5	5
4	1	1.5	2
5	1	1	1.5
6	1	1	1
Total	**14**	**66**	**124.5**

Table 22c. **Average Synset Neighbor Count frequency for polysemy ranges**

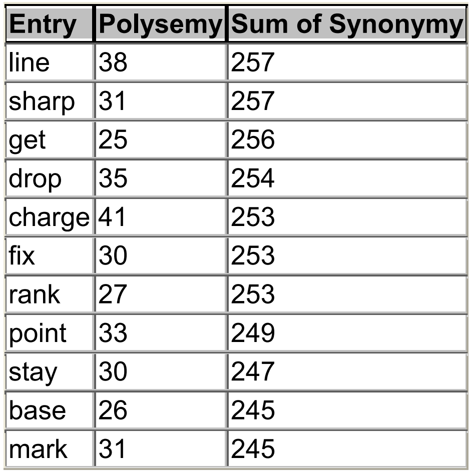
The Thesaurus is not divided into semantic neighborhoods, of course. When semantic neighborhoods are composed together they form topic-word-independent structures that are essentially Type-10 chain components, but including all the loose ends—the many odd senses and words that were trimmed off by the Type-10 chain and Jacuzzi algorithm constraints. These could be called *semantic topologies.*

The principles discussed here are, symmetrically, also true for *topic senses*.

Semantic Switching Centers

The word-equivalents of Synset switching centers described in Section: *Conceptual Switching Centers* are, approximately, the words in RIT that are the most polysemous. The conceptual switching centers were ordered based on their *power*, which was calculated based on the sum of the polysemies of their Entries. Table 23 shows the top twenty words ordered by the sum of their synonymies (the number of Entries co-occurring in each Synset), along with their polysemy. The words listed are called here, *semantic switching centers*. *Cut*, which has 64 senses, or Synsets, has a summed synonymy of 595, which is on average about 9 synonyms, or Synset neighbors, per Synset.

Entry	Polysemy	Sum of Synonymy
cut	64	595
set	51	379
turn	45	368
run	54	366
pass	41	345
beat	37	300
go	34	272
crack	23	264
dash	25	261



Entry	Polysemy	Sum of Synonymy
line	38	257
sharp	31	257
get	25	256
drop	35	254
charge	41	253
fix	30	253
rank	27	253
point	33	249
stay	30	247
base	26	245
mark	31	245

Table 23**. Top twenty semantic switching centers by summed synonymy**

Synonyms of *cut*, for example, may be counted multiple times by this method if the synonym occurs with *cut* in multiple Synsets. However a similar result is obtained even if the top twenty semantic switching centers are selected on the basis of the number of words, rather than the number of Entries that they each co-occur with. The numbers are reduced by about one-third in each case, but the order remains largely the same. In other words, if *above* is counted as co-occurring with *over* only once instead of the six times it actually co-occurs with *over*, *over* is still a top ranking word (at 114th). This means that it is the number of words that a word shares senses with (and thereby is associated with), not the size of the Synsets it participates in that gives it its semantic power, even though they will be strongly correlated. In terms of small-world social networks, by analogy, this would imply that person A knowing person B through work, as a friend, and as a relative, does not change the fact that A knows B, and that B serves as a connection in A's social network.

The top semantic switching centers are classified as Entries in RIT under multiple parts-of-speech. *Cut* occurs 30 times as a verb, 27 times as a noun, and 7 times as an adjective. If the top 100 semantic switching centers (of 113,665 words) are counted for their instantiations as Entries by part-of-speech there are 260 Entries (of 199,423 in RIT), and of these 99 are verbs, 96 are nouns, and 43 are adjectives. The verbal senses of the *word* switching centers are disproportionately more frequent than for the Thesaurus as a whole, by about 250%.

As shown in Section: *TMC and VJC Components*, high-polysemy words along with high-synonymy Synsets bind the components together. The most polysemous words occur in the largest, most complex (in terms of network connections) components. The longer a word is around the more senses it can acquire and the more fellows it can become associated with. To this extent the semantic switching centers may represent the oldest concepts in the English language.

Chapter 6: Discussion

The entry point for this research was “What are the patterns of connectivity within Roget’s Thesaurus as they relate to the conceptual structure?” This study has identified the explicit and implicit patterns within the thesaurus. The explicit patterns result from the structure of the thesaurus itself—the organization imposed by Roget as he developed and organized words to classify them by their meanings. This involves primarily the Synopsis of Categories (conceptual hierarchy, noun tree, or classification structure) founded on the classification of words “according to the ideas which they express” (P. M. Roget, 1852, Introduction).

Roget’s stated goal has often been misunderstood. The Thesaurus has been viewed by some reviewers as merely a synonym finder, a crossword helper or a crutch for weak vocabularies. Even Simon Winchester, respected for his book on the history and development of the *Oxford English Dictionary*, misunderstood the purpose and is negative and critical of the result. In a response to Winchester’s 15,000-word article lambasting Roget in *The Atlantic Monthly*, Paul Vallely wrote: “Roget – like anyone who properly appreciates his linguistic reveries – understands that there are implicit semantic relations and moral hierarchies in concepts and the words that connect them” (Vallely, P. 2002). The *implicit* and the *connected* have been the focus of this research.

Implicit patterns or structures emerge from the semantic relationships between words and words, words and senses, between the attributes of the words (such as part-of-speech and cross-references), and between words and the organizational structure of the thesaurus. These implicit patterns are hidden from the Thesaurus user but made explicit by automation and visualization. The semantic (word) switching centers that cluster senses and their associated words, and the differentiating singletons, in concert, create the balance between a highly-connected and clustered network of words, and a sparse network of random links. These features are probably why the Thesaurus displays the characteristics of a small-world network. The conceptual (sense) switching centers, essential to so many of the connectivity or clustering features observed throughout the Thesaurus such as the Type-10 components, share many of the same characteristics as the semantic switching centers.

When words and senses are combined they form a duality that is evident at the local level as the semantic neighborhoods and the Type-10 components visualized in the lattice diagram examples; and that is evident at the global level as coherence within Classes, balanced by long distance semantic overlap such as the word sharing seen in the *cross* feature of two-way contingency tables displayed as information maps. A pattern of relationships emerges when words and senses are viewed at the local level, called here genus and differentiae. This organization exists implicitly in all areas of the Thesaurus, at all scales, and appears to be a potentially optimal arrangement of words and senses such that information is neither too dense nor too sparse, and at the same time retains the word-word, word-sense, and sense-sense relations. Genus and differentiae appear to be a facet of small-world networks as many of the characteristics, such as the type of frequency distributions and being scale-free, are the same.

The most frequent type of words and largest Categories in the Thesaurus relate to the names of things, such as animals, plants and holidays. Lists alone account for 20,000, or 10% of the Thesaurus Entries, and nouns generally account for more than half of all Entries. However the highest frequency elements of the word and sense networks, and the notions represented by the Category and Class labels of the classification hierarchy, indicate a pervasive semantic core related to agitation, survival, and motion, rather than *things*. The *cross* pattern observed in global views is specifically formed by high polysemy verbs classified in Categories broadly to do with motion, velocity and direction; or *change of place*. Furthermore, verbal senses (relating to actions, rather than things) are disproportionately represented among the senses of the top one hundred semantic switching centers.

Additional elements of the Thesaurus structure, the cross-references and part-of-speech categories, although having different characteristics and not being organized in the same way as senses and words, provide views of the Thesaurus semantics that support the observations made via the analysis of word and sense patterns.

Cross-references, while not strongly reflecting the global patterns of cross-class connectivity or overlap, do reflect the local intra-class coherence at various levels of the hierarchy; and aggregated cross-references, the cross-reference hubs and authorities, reflect the notions of threat, survival and motion. The most-connected semantic centers, by cross-reference counts, categorize predominantly verbs. Those coinciding with the *cross* phenomenon, observed in global visualizations of RIT (by summed polysemy of their words), relate to notions of agitation, direction, travel, navigation, ejection and transference (2.IV Motion). Those outside of the cross pattern have to do with notions of displeasure, excitement and fear (8.I Personal Affections); excess, motivation and inducement, and avoidance (7.I Volition in General); information, disclosure, and speech (6.III Communication of Ideas); and (low on the word connectedness scale, but highest in cross-reference counts) inquiry, discovery, and belief (6:I: *Intellectual faculties and processes*). The *Intellectual* notions that were high in cross-reference counts were uncertainty, insanity-mania, and foolishness.

Part-of-speech patterns, while not analyzed generally, were observed to reflect a pattern of *change of place* in the semantics of core prepositions. The concepts of *staying*, *going* and *gone* were seen, through the etymological roots of prepositions, to overlap with the semantics of words from other parts-of-speech, suggesting that location, position and direction are an important organizing concept across the range of semantics of language (at least the English language, and to the extent that RIT is representative of it).

Living things that move (change place) also have central nervous systems, and living things that do not move, do not have central nervous systems. Motion or change of place and the perception, coordination, navigation skills, and memory that go with it would be a satisfactory explanation of the importance of such notions in RIT, as implied by the results of this study. The relationship of this to notions of agitation, excitement, displeasure, and fear may possibly be accounted for, as proposed by Old (2000), by a model related to survival and the fight-flight response.

In comparison to the word association data and WordNet, RIT is qualitatively different. It lacks many of the connections between words, or relations found among word associations. It also lacks the formal organization among parts-of-speech, and the explicit relations such as meronymy, found in WordNet. It does however correlate and compare favorably at many points of intersection, providing confirmation of the validity of both itself and the compared data sources.

Dr Roget laid the foundations for a complete categorization of the concepts represented by words of the English language. Although culture, attitudes, philosophy and, especially, vocabulary have moved on in the 150 years since its first publication, and almost 200 years since its first inception, the structure is still valid and rich in ways that are not available in dictionaries or other lexical sources. A global model has yet to be developed to account for all of the patterns lying implicit in the associations, connections, implications, networks and substructures, observed during this research, but the information is now at least explicit.

Future Research

The goal of developing intelligent machines, proposed in the 1950s, has been largely redirected after 50 years of disappointment to the pursuit of systems that potentiate human intellectual capabilities. Insights and theories derived from the study described in this dissertation could potentially support either goal. Future research directions should relate findings of this study to current research in cognition and cognitive processes. This includes comparing magnetic resonance imaging studies of the organization of the brain to structures identified in Roget's Thesaurus. Research into the organization of the mental lexicon and classification structures in the brain could be guided by feedback between results and patterns observed, and patterns found in Roget's Thesaurus. For example, Williams, among others, has suggested that word recall speed is faster for native English speakers for words of Anglo-Saxon origin, than for words of foreign origin (Williams, 1975, p. 125). High frequency words also have a shorter recall time, and are recognized faster. Both observations have implications for research in cognition, and both could be related to Roget's Thesaurus concept patterns.

Griffiths and Steyvers (2002) have shown that probability distribution over the allocation of words to topics (derived from corpus data) is consistent with the statistical properties of Roget's Thesaurus, where words are allocated to thesaurus Categories. Griffiths and Steyver's model has predictive and generative properties for concepts arising in a particular given context. Future research should be directed toward evaluating whether word distributions in Roget's (such as those found among the genus and differentiae distributions) have the same predictive and generative properties.

Cross-references probably reflect the small-world structure of the Thesaurus semantics. This remains to be tested.

Only the top levels (by counts of polysemy, synonymy, connectedness, and so on) of words, senses, cross-references, and components were studied here in detail. Further work is needed to characterize the singletons (monosemous words and monolexic senses); the list-like synsets; the patterns among the regular, Xref type cross-references; and the relationships among the smaller Type-10 chain components.

Network visualizations were not used in this dissertation. Free-form graphs and networks, rather than the formal, constrained and structured lattices and contingency tables used for this study, could provide more creative insights into the RIT data. These visualizations could be used to explore the *semantic topologies* of combined semantic neighborhoods in order to develop an understanding of how the neighborhoods relate to each other in a *topic-neutral* way. In addition, components could be treated as nodes and the Jacuzzi *corner-joins* could be used as arcs to provide a network view of the relationships between components. Preliminary multi-dimensional scaling shows components predominantly representing notions of “stridor,” *impulse, excitement, displeasure* and *agitation* clustered around a notion characterized by *impairment* (death).

Implicational structures between cross-references and between words were not developed here to their full potential. While implications may be read directly from lattices, global networks of inferences, including multiple-premise implications could be displayed and explored using network visualizations. Sequences of implications, or inference chains, would provide strongly connected, low complexity shadow structures of the Thesaurus data, and possibly provide supplementary insights into the semantic patterns within RIT.

The preliminary study carried out prior to this dissertation research found that conceptual patterns among synonyms translated from Chinese differed from those in RIT. For example there was a greater emphasis, as measured by word connectedness, on concepts such as {mark, sign, seal, image, display} associated with {talk, say, speak, speech, explain, information}—possibly reflecting the use of iconic representation in Chinese written language. In general, further research should compare the patterns observed in this study of English to patterns of non-Indo-European languages.

The origins of humankind, human language, human culture, and human technology[35] are fascinating to modern humans despite our future-orientated civilization. But theories about how humans came to be as they are, from linguistics, archaeology and anthropology, have been contradictory at times. Recent primate genetic studies have contradicted some of the traditional theories about our origins, but also endorsed others. Only a trans-disciplinary approach can resolve apparent contradictions and provide coherent models. Modern culture is carried largely by language, in the concepts and metaphors represented by the words of the language(s) and cultural evolution is probably reflected by the evolution of language. Future research should be directed toward comparing the implicit patterns found in this study to data accumulated by etymologists, geneticists, primatologists, social anthropologists, and archaeological anthropologists, with a view to developing a model that accounts for these patterns.

[35] The qualifier "human" is repeated because, while culture, language and technology may be more developed among humans, they may not set humans apart.

The man is not wholly evil- he has a Thesaurus in his cabin.

(Reputedly a quotation from J. M. Barrie's children's book, *Peter Pan*, in his description of Captain Hook—commonly quoted on the Internet, and included in the Editors Preface to the 2002, 150th Anniversary Edition of the British Roget's Thesaurus).[36]

[36] In fact the correct quote is, "The man was not wholly evil; he loved flowers"—but this one is more appealing to thesaurophiles.

References

Albert, R. & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*, 47-97.

Alford, R. R. (1998). *The craft of inquiry*. New York: Oxford University Press.

AHD, The American heritage dictionary of the English language (4th ed.). (2002). Bartleby.com. Available at http://www.bartleby.com/61/

Barrett, M. D. (1982). Distinguishing between prototypes: The early acquisition of the meaning of object names. In S. A. Kuczaj (Ed.), *Language development*, vol. I: *Syntax and semantics* (pp. 313-334). Hillsdale, NJ: Lawrence Erlbaum.

Berrey, L. (Ed.). (1962). *Roget's international thesaurus* (3rd ed.). New York: Crowell.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual (Web) Search Engine. *Computer Networks and ISDN Systems. 30*(1-7), 107-117.

Brown, T. L. (2003). *Making truth: Metaphor in science*. Urbana-Champaign: University of Illinois Press.

Brugman, C. (1981). *The story of* "*over.*" Unpublished master's thesis, University of California, Berkeley. Available from the Indiana University Linguistics Club.

Brugman C. & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. I. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical ambiguity resolution* (pp. 477-508). San Francisco: Morgan Kaufmann.

Bryan, R. M. (1973). Abstract thesauri and graph theory applications to thesaurus research. In S. Y. Sedelow (Ed.), *Automated language analysis, report on research 1972-73* (pp. 45-89). Lawrence, KS: University of Kansas.

Bryan, R. M. (1974). Modeling in thesaurus research. In S. Y. Sedelow (Ed.), *Automated language analysis 1973-74* (pp. 44-69). Lawrence, KS: University of Kansas.

Claiborne, R. (1988). *The roots of English: A reader's handbook of word origins*. New York: Anchor books, Doubleday.

Crystal, David (1987). *The Cambridge encyclopedia of language.* Cambridge: Cambridge University Press.

Denisowski, P. (2001). *CEDICT: Chinese-English Dictionary*. Available at http://www.mandarintools.com/cedict.html

Dodds, P. S., Muhamad, R., & Watts, D.J., (2003). An experimental study of search in global social networks. *Science*, *301* (5634), 827-829.

Dornseiff, F. (1970). *Der deutsche Wortschatz nach Sachgruppen*. New York: Walter De Gruyter.

Dutch, R. A. (Ed.). (1962). *Roget's thesaurus of English words and phrases*. London: Longman.

Ellman, J., & Tait, J. (1999). Roget's thesaurus: An additional knowledge source for textual CBR? *Proceedings of the 19th SGES International Conference on Knowledge Based and Applied Artificial Intelligence* (pp. 204-217). London: Springer-Verlag.

Elsen, H. (2001). The structure of meaning: Semasiological and onomasiological aspects of development. *Onomasiology Online 1 (2000).* Available at http://www.ku-eichstaett.de/SLF/EngluVglSW/elsen1001.pdf

English Linguistics 1500-1800: Texts on microfiches - arranged by subject categories and then by date within each category. Bergamo, Italy: Universita Degli Studi Di Bergamo. Available at http://wwwesterni.unibg.it/siti_esterni/anglistica/m-f10.htm

Garfield, E. (2001). "*From Bibliographic Coupling to Co-Citation Analysis via Algorithmic Historio-Bibliography" A Citationist's Tribute to Belver C. Griffith.* Presented at Drexel University, Philadelphia, November 27, 2001. http://www.garfield.library.upenn.edu/papers/drexelbelvergriffith92001.pdf

Girard, Abbé Gabriel (1762). *A new guide to eloquence: Being a treatise of the proper distinctions to be observed between words reckoned synonymous.* London: Scolar Press, 1974 (See *English linguistics 1500-1800*).

Google (2003). *Our search: Google technology*. Available at http://www.google.com/technology/index.html

Goldstone, R.L. & Kersten, A (2003). Concepts and categorization. In A. F. Healy & R. W. Proctor (Eds.), *Comprehensive handbook of psychology, Volume 4: Experimental psychology* (pp. 599-621). New York: Wiley.

Griffiths, T. L. and Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray (Ed.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*. Fairfax, VA: George Mason University. Available at http://www-psych.stanford.edu/~gruffydd/papers/semrep.pdf

Herman, I., Melancon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, *6*, 24-43.

Hofstadter, D. R. (1999). Analogy as the core of cognition. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational and neural sciences*. Cambridge, MA: MIT Press.

Ipsen, G. (1924). *Stand und Aufgaben der Sprachwissenschaft*. Heidelberg: Winter.

Ito, K., Ed., (1987). *Encyclopedic Dictionary of Mathematics, Second Ed.*, Cambridge, Massachusetts: The MIT Press.

Jacuzzi, V. (1991, May). Modeling semantic association using the hierarchical structure of Roget's international thesaurus. Paper presented at the *Dictionary Society of North America Conference*, Columbus, Missouri.

Janssen, M. (2002). SIMuLLDA *: a multilingual lexical database application using a structured interlingua* (Doctoral Dissertation, University of Utrecht, 2002). Utrecht, NL: Digitaal Archief Universiteit Utrecht. Available at http://www.library.uu.nl/digiarchief/dip/diss/2002-0905-111545/inhoud.htm

Jones, K. S. (1964). *Synonymy and semantic classifications*. Cambridge: Cambridge Language Research Unit.

Kleinberg, J. M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys*, *31*(4es): 5.

Kochen, M. (Ed.), (1965). *Some problems in information science*. New York: Scarecrow Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lanbridge's concise Chinese-English dictionary (1985). Taipei, Taiwan: Lanbridge Press.

Lenat, D. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), 33-38.

Lenat, D., Guha, R., Pittman, K., Pratt, D., & Shepherd, M. (1990). CYC: Toward programs with common sense. *Communications of the ACM*, *33*(8), 30-49.

Liddy, E., Hert, C., & Doty, P. (1990) Roget's international thesaurus: Conceptual issues and potential applications. In S. Humphrey & B. Kwasnik (Eds.), *Proceedings of the 1st ASIS SIG /CR Classification Research Workshop* (pp. 95-100). Silver Spring, MD: American Society for Information Science.

Mawsom, C. O. S. (Ed.), (1911). *Roget's thesaurus of English words and phrases*. New York: Crowell.

Mawsom, C. O. S. (Ed.). (1922). *Roget's international thesaurus* (1st ed.). New York: Crowell.

Mawsom, C. O. S. (Ed.). (1935). *Roget's international thesaurus*. New York: Crowell.

McHale, M. L., & Crowter, J. J. (1994). Constructing a lexicon from a machine-readable dictionary. *In-House Report RL-TR-94-178, Rome Laboratory, Air Force Materiel Command, Griffiss Air Force Base.* Washington DC: U.S. Government Printing Office.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., & Tengi, R. (1993). Five papers on WordNet. *Technical Report*. Princeton, N.J: Princeton University.

Mooney, D. M., & Talburt, J. R. (1990). Homograph discrimination for intelligent interfaces via thesaural lexicons (Abstract). *ACM Conference on Computer Science 1990*, 449.

Motter, A. E., de Moura, A. P. S., Lai, Y.-C. & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review, E, 65,* 065102.

Muller, C. (1997). *World Wide Web CJK-English dictionary database*. Chiba, Japan: Toyo Gakuen University. Available at http://www.uoregon.edu/~felsing/wired/generalindex.htm

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Available at http://www.usf.edu/FreeAssociation/

Old, L. J. (1991a). Analysis of polysemy and homography of the word "lead" in Roget's international thesaurus. In R. Gamble & W. Ball (Eds.), *Proceedings of the Third Midwest AI and Cognitive Science Conference,* (pp. 98-102). Carbondale, IL: Southern Illinois University.

Old, L. J. (1991b, May). Image schemas and lexicons: A comparison between two lexical networks. Paper presented at the *Dictionary Society of North America Conference*, Columbus, Missouri.

Old, L. J. (1993). Image schemas and lexicons: A comparison between lexical networks. In T. E. Ahlswede (Ed.), *Proceedings of the 5th Midwest AI and Cognitive Science Conference* (pp. 31-35). Chesterton, IN: Southern Illinois University.

Old, L. J. (1995, June, 22). *Lattice representation of cross-lingual semantics*. Guest lecture, Algemeine Algebra, Fachbereich Mathematik, Darmstadt Technische Hochschule, Darmstadt, Germany.

Old, L. J. (1996b). Synonymy and word equivalence. Online Proceedings, *Midwest Artificial Intelligence and Cognitive Science Society Conference* (MAICS96). Available at http://www.cs.indiana.edu/event/maics96/Proceedings/old.html

Old, L. J. (1999, April, 24). Spatial representation of semantic information. Paper presented at the *Mid-West Artificial Intelligence and Cognitive Science Society Conference* (MWAICS'99), Bloomington, Indiana.

Old, L. J. (2000, October). Core concept patterns in English semantic networks and Indo-European roots. Paper presented at *Connections 2000: The Sixth Great Lakes Information Science Conference* in Knoxville, TN. Abstract: C*anadian Journal of Information and Library Science 25* (1), 42.

Old, L. J. (2002). Information cartography applied to the semantics of Roget's thesaurus. In S. Conlon (Ed.), *Proceedings of the 13h Midwest Artificial Intelligence and Cognitive Science Conference* (MAICS'02), (pp. 65-70). Chicago, IL: Illinois Institute of Technology.

Old, L. J. (2003). An analysis of semantic overlap among English prepositions in Roget's Thesaurus. In P. Saint-Dizier (Ed.), *Proceedings of the Association for Computational Linguistics SIG Semantics Conference* (ACL-SIGSEM) (pp. 13-19). Toulouse: IRIT

The Oxford English Dictionary (2nd ed.). (1989). Burchfield, R.W., (Ed.). Oxford: Clarendon Press.

Piotrowski, T. (1994). British and American Roget. In W. Hüllen (Ed.), *The world in a list of words* (pp. 123-135). Tübingen: Niemeyer.

Piozzi Lynch, H. (1794). *British synonymy; or, an attempt at regulating the choice of words in familiar conversation*. (See *English linguistics 1500-1800*).

Priss, U. (1996). *Relational Concept Analysis: Semantic structures in dictionaries and lexical databases*. (Doctoral Dissertation, Technical University of Darmstadt, 1998). Aachen, Germany: Shaker Verlag.

Priss, U., & Old, L. J. (1998). Information access through conceptual structures and GIS. In C. M. Preston (Ed.), *Proceedings of the American Society for Information Science Conference (ASIS '98),* 35, 91-98.

Project Gutenberg (2002). *Project Gutenberg Official Home site – index – free books on-line.* Available at http://promo.net/pg/

Quillian, M. R. (1967). Word concepts: a theory and simulation of some basic semantic capabilities, *Behavioral Sciences*, *12,* 410-430.

Roget, P. M. (1852/1992). *Thesaurus of English words and phrases, classified and arranged so as to facilitate the expression of ideas and assist in literary composition* (Facsimile of the First Edition). London: Bloomsbury Books.

Roget, S. R. (Ed.), (1933). *Thesaurus of English words and phrases (Authorised American Edition)*. New York: Grosset & Dunlap.

Sabbage, L. (2003, March, 9). Take my word for it. *The Sunday Times*, *37* (41). Available at http://www.sundaytimes.lk/030309/mirror/2.html

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Saussure, F. de (1916). *Cours de linguistique generale*. Paris: Payot.

Sedelow, S. Y. (1961). *The narrative method of Paradise Lost* (copyrighted microfilm edition). Ann Arbor: University Microfilms.

Sedelow, S. Y. (1972). Language Analysis in the humanities. *Communications of the ACM, 15*(7), 644-647.

Sedelow, S. Y. (1974). Brief overview of research under this contract. S. Y. Sedelow (Ed.), *Automated language analysis, report on research 1973-74* (pp. 2-5). Lawrence, KS: University of Kansas.

Sedelow, S.Y. (1991). Exploring the terra incognita of whole-language thesauri. In R. Gamble & W. Ball (Eds.), *Proceedings of the Third Midwest AI and Cognitive Science Conference* (pp. 108-111). Carbondale, IL: Southern Illinois University.

Sedelow, S.Y. (1993). Formally modeling and extending whole-language-scale semantic space. *Behavior Research Methods, Instruments and Computers*, *25*(2), 310-314.

Sedelow, S. Y., & Sedelow, W. A., Jr. (1969). Categories and procedures for content analysis in the humanities: The analysis of communication content. In G. Gerbner (Ed.), *Developments in scientific theories and computer techniques* (487-499). New York: Wiley.

Sedelow, S. Y. & Sedelow, W. A., Jr. (1986a). The lexicon in the background. *Computers and Translation*, *1*(2), 73-81.

Sedelow, S. Y., & Sedelow W. A., Jr. (1986b). Thesaural knowledge representation. In *Proceedings of the 2nd International Conference of the University of Waterloo Centre for the New Oxford English Dictionary: Advances in Lexicology* (pp. 29-43). Waterloo, ON: University of Waterloo.

Sedelow, S.Y., & Sedelow, W. A., Jr. (1992). Recent model-based and model-related studies of a large-scale lexical resource. In *Proceedings of COLING-92*, 1, 1223-1227.

Sedelow, S.Y., & Sedelow, W. A., Jr. (1994a). Graph theory, set theory, & order theory in semantic space: Analysis for use in knowledge representation. In J. Liebowitz (Ed.), *Proceedings of the Second World Congress on Expert Systems*. New York: Cognizant Communications Corporation. (*CD ROM - The World Congress on Expert Systems '94*. Cambridge, MA: Macmillan New Media).

Sedelow, S.Y., & Sedelow, W. A., Jr. (1994b). A topological model of the English semantic code and its role in automatic disambiguation for discourse analysis. In S. Hockey & N. Ide (Eds.), *Research in Humanities Computing 2* (pp. 18-31). Oxford: Oxford University Press.

Sedelow, W. A., Jr. (1957). Science and the language of history. *Behavioural Science: A Journal for the Society for General Systems Research, 2* (1), 80-82.

Sedelow, W. A., Jr. (1968). History as language. *Computer Studies, 1*(4), 183-190.

Sedelow, W. A., Jr. (1985). Semantics for humanities applications: Context and significance of semantic "stores." *Proceedings of the American Society for Information Science Conference, 22,* 363-366.

Sedelow, W. A., Jr. (1990). Computer-based planning technology: an overview of inner structure analysis. In L. J. Old (Ed.), *Getting at disciplinary interdependence*, (pp. 7-23). Little Rock, AR: Arkansas University Press.

Sedelow, W. A., Jr. (1991). The lay of the land. In R. Gamble & W. Ball (Eds.), *Proceedings of the Third Midwest AI and Cognitive Science Conference,* (pp. 88-92). Carbondale, IL: Southern Illinois University.

Sedelow, W. A., Jr. (1993). The formal analysis of concepts. *Behavioral Research Methods, Instruments and Computers 25*(2), 314-317.

Sedelow, W.A., Jr. & Sedelow, S.Y. (1979a). The history of science as discourse. *Journal of the History of the Behavioral Sciences*, *15*(1), 63-72.

Sedelow, W.A., Jr. & Sedelow, S.Y. (1979b). Graph theory, logic, and formal languages in relation to language research,. In W. A. Sedelow Jr. & S. Y. Sedelow (Eds.), *Computers in Language Research: Formal Methods* (pp. 7-17). The Hague: Mouton.

Sedelow, W. A., Jr. & Sedelow, S. Y. (1983). Science and human language. In W. A. Sedelow Jr. & S. Y. Sedelow (Eds.), *Computers in language research 2*, (pp. 1-23). New York: Walter de Gruyter.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*, 327-340.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, *28*, 9-12.

Soule, R. (1871/1948). *Dictionary of English synonyms & synonymous expressions designed as a guide to apt and varied diction*. New York: Tudor.

Steyvers, M., & Tenenbaum, J. B. (2001). *The large-scale structure of semantic networks: statistical analyses and a model of semantic growth*. Manuscript submitted for publication.
Available at http://www-psych.stanford.edu/~msteyver/papers/smallworlds.pdf.

Strogatz, H. (2001). Exploring complex networks. *Nature*, *410*, 268-276.

Talburt, J. R., & Mooney, D. M. (1990). An evaluation of Type-10 homograph discrimination at the semi-colon level in Roget's international thesaurus. *Proceedings of the 1990 ACM SIGSMALL/PC Symposium,* 156-159.

Taylor, S. R. (1974). Selected graph theory applications to a study of the structure of Roget's thesaurus. In S. Y. Sedelow (Ed.), *Automated language analysis, report on research 1973-74* (pp. 60-116). Lawrence, KS: University of Kansas.

Tooke, J. (1786/1840). *Epea Pteroenta, or the Diversions of Purley (1786-1805)*, (2nd ed.). London: Thomas Tegg.

Travers, J, & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*(4), 425-443.

Trier, J. (1931/1973). *Die Worte des Wissens*, in: A. Van der Lee & O. Reichmann (Eds.), *Aufsätze und Vorträge zur Wortfeldtheorie* (pp. 66-78). The Hague: Mouton.

Trusler, J. (1766). *The difference between words esteemed synonymous in the English language; and the proper choice of them determined*. (See *English linguistics 1500-1800*).

Tversky, B. (1999). What does drawing reveal about thinking? In J.S. Gero & B. Tervsky (Eds.), *Visual & Spatial Reasoning in Design*. Sydney, Australia: Key Center of Design Computing and Cognition. Available at http://www.arch.usyd.edu.au/kcdc/books/VR99/Tversky.html

Vallely, P. (2002). Peter Mark Roget: A man of words. *The Independent Portfolio*. London: Independent Digital. Available at http://enjoyment.independent.co.uk/books/features/story.jsp?story=309642

Van der Lee, A., & Reichmann, O. (Eds.), (1973). *Aufsätze und Vorträge zur Wortfeldtheorie*. The Hague: Mouton.

Vincent, N. (1999). The evolution of c-structure: Prepositions and PPs from Indo-European to Romance. *Linguistics*, *37*(6), 1111-1153.

Wang, Y-C., Vandendorpe, J., & Evens, M., (1985). Relational thesauri in information retrieval. *Journal of the American Society for Information Science*, 36(1): 15-27.

Watts, D.J. (1999). *Small worlds: The dynamics of networks between order and randomness*. Princeton, NJ: Princeton University Press.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442.

Webster's new collegiate dictionary. (1976). Springfield, MA: Merriam-Webster.

Webster's third new international dictionary (unabridged). (1965). P. B. Gove (Ed.), Springfield, MA: G. & C. Merriam.

Wille, R., (1982). Restructuring lattice theory: an approach based on hierarchies of concepts. In I. Rival, (Ed.), *Ordered sets* (pp. 445-470). Dordrecht: Reidel.

Wille, R. (1989). Geometric representation of concept lattices. In O. Opitz (Ed.), *Conceptual and numerical analysis of data* (pp. 239-255). Berlin: Springer-Verlag.

Williams, J. M. (1975). *Origins of the English language: a social and linguistic history*. New York: The Free Press.

Winchester, S. (1998). *The professor and the madman* (also published as The surgeon of Crowthorne). New York: Penguin.

Winchester, S. (2001). Word imperfect. *The Atlantic Monthly*, *287*(5), 53-72. Available at http://www.theatlantic.com/issues/2001/05/winchester-p1.htm

Wouters, P. (1998). The Signs of Science. *Scientometrics*, *41*(1-2), 225-241.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of COLING-92*, 454-460.

Young, M.P., (1993). The organization of neural systems in the primate cerebral cortex. *Proceedings of the Royal Society*, London, B., 252, 13-18.

Appendices

Appendix A: Sources of Information

This appendix gives more-detailed information on the data and information sources collected and organized into normalized relational databases in support of this research. Additional sources of information are text versions of Roget's Thesaurus, various editions, listed below.

Databases

1. Roget's International Thesaurus (RIT), 3rd Edition (1962) in database form, is the central data source. The semantics and word associations of Roget's Thesaurus have been "culturally" validated for 150 years. Partly for this reason RIT was chosen as the basis for the "lexicon in the background"—"a general purpose reference work of semantically-related words" (S. Y. Sedelow, 1974)—for a study of translational associations, funded by the Office of Naval Research, in support of natural language processing efforts to translate Russian Military Strategy in the 1960's and early1970's. RIT was converted to machine-readable form at that time, and converted to database form by this author in the early 1990's.

   The machine-readable form consisted of entries (one word may have several entries), each with 22 attributes. Attributes, coded as integers, consist of features associated with each entry in the actual text. Examples of these attributes are the entry's part-of-speech, its font type (bold or italic, and so on), polysemy (total number of senses in RIT for the word which the entry represents), cross-reference information, and the RIT Category in which the entry is found (its location in the text).

   The database version, in addition, includes lookup tables for the codes (what the integer codes mean in human-readable text); the structure of the full Roget hierarchy; labels for the classes and categories; cross-references, in a separate table; the antonymy relation between categories; and several tables to facilitate processing, such as a separate entry index. Following research by Talburt, Mooney and Jacuzzi, the derived components were also added to the database.

   RIT is still under copyright. Dr. Walter Sedelow Jr. and Dr. Sally Yeates Sedelow, Professors Emeriti, University of Arkansas, were granted rights to it for research purposes. They have allowed the thesaurus to be used for this dissertation research, and are informal consultants and committee members for this dissertation.

2. The 1911 edition of Roget's Thesaurus, the body of which is available as an electronic book (e-text) from Project Gutenberg (Project Gutenberg, 2002), has been converted to database form by this author, and is compared here with the 1962 RIT version. The 1911 edition is virtually identical to Roget's first (1852)

edition, and retains the synonyms-antonyms opposed category organization of the original Roget's hierarchy. Differences and similarities between the 1911 version and the 1962 version highlight changes and stabilities in technology, culture, and word usage; and editing philosophies.

(The Index of the 1911 edition, missing from the Gutenberg e-text edition, has been reconstructed/generated from the body as part of this research, and contributed for inclusion in the Gutenberg archive.)

3. Word association data from the University of South Florida (Nelson, McEvoy, and Schreiber, 1998) was converted to database form by this author. The USF data is available on-line for download in the form of multiple ASCII text files. The data consists of about 5,000 "cue" words, or prompts, and about 10,000 "target" words, or responses. This data has been extensively processed, analyzed, and researched by its developers at the University of South Florida. Almost all of the cue words are also targets, and can be found as entries in RIT. However 2,000 of the 10,000 target words are not in RIT. These "missing" words include plurals and other non lemmas[37] (25%); commercial products (12%); proper nouns (7%); slang, abbreviations, and interjections (6%); and misspellings (4%). The remaining missing words (46% of those missing) are numbers and dates; geographic locations; contractions, abbreviations, plurals, and conjugated verbs; partial phrases or compounds ("don't let [go]"; bumble [bee]); and some words which are genuine omissions from Roget's Thesaurus, such as the days of the week, the months, and many technical terms. These statistics were derived as part of the pilot study for this research.

4. Indo-European Roots and English Base Words Database: This is used to analyze semantic patterns in RIT and test theories of concept development. An example of an Indo-European root is "STA-."

> **STA-,** *L stare, stat-, Gc* STAND, whence UNDERSTAND, originally = stand under, whence comprehend ("have a good hold on") and STANDARD (originally, a rallying place where warriors came to stand). A *Gc* STOOL stands on its, legs and a plant stands on its STEM … (Claiborne, 1989, *STA-*).

Indo-European roots are, technically, proto-Indo-European (PIE) roots, as they are hypothetical constructs derived by linguists who analyzed the observed commonalities between languages. The roots are the common antecedents of the Indo-European languages that include Sanskrit and modern Indian sub-continent languages such as Hindi, Urdu, Punjabi and Bengali; Iranian (Persian, or Parsee); the Germanic languages (English, Dutch, Danish, and the other Scandinavian languages); Latin and Romance languages such as Romanian, Portuguese, and French; the Slavic languages such as Russian and Polish; Celtic languages such as Gaelic and Welsh; and several extinct languages such as Gothic. Indo-European

[37] A lemma is a dictionary entry in bold type—also known as a headword or morphological base. It is the main form of a word, as opposed to, for example, its plural form in the case of a noun, or past tense in the case of a verb.

roots are believed to represent the concepts important to the early Indo-Europeans at least 10,000 years ago.

The database contains about 1200 Proto-Indo-European roots and was developed by this author from several sources, primarily *The Roots of English* by Robert Claiborne (1989), *The American Heritage Dictionary* (AHD), and small lists of roots from the Internet. Agreement between sources was used to decide on the inclusion, or not, of specific roots in the data set; and to resolve conflicting opinions about the etymological origins of English words in the data. Cognates (words with a common PIE root) were, in addition, added based on etymological information in *Webster's (unabridged) Dictionary* (1965).

5. Chinese-English Association Dictionary: This is used as a non-Indo-European language source for comparison with patterns in RIT and for testing some theories developed during this research. This database contains 36,000 entries and was developed by this author from the CEDICT (Mandarin) Chinese-English Dictionary (Denisowski, 2001). The "associations" between Chinese words were derived based on Chinese words that share two English translations. For example the association between (using Pinyin notation) ke4 and bin1 was formed by using their shared (in common) English translation {guest, visitor}. Three native speakers of Mandarin have informally validated these derivations.

6. WordNet lexical database, available on-line from the Princeton University Cognitive Laboratory: This was converted to relational form by Dr. Uta Priss, School of Library and Information Science, Indiana University and Randee Tengi, Princeton University Cognitive Laboratory. WordNet is an alternative method of structuring synonym sets based on psycholinguistic theories of human lexical memory developed by Dr. George Miller at Princeton University. Connectivity patterns in RIT are compared to those of WordNet.

7. Word lists such as frequency data, homographs (developed by this author, and discussed in Old, 1991a), and common "base" words, from various sources, were used to supplement this analysis, for validation, and to provide bridges between data sets.

Texts

For ad hoc comparisons between versions of Roget's Thesaurus and as references for historical threads in the development of the thesaurus, the following editions have been used:

British editions of Roget's Thesaurus:

- 1852 (facsimile of the 1st Edition);
- 1933 (the "Authorised American Edition");
- 1936;
- 1962;
- 2002 (150th Anniversary Edition).

American editions of Roget's Thesaurus:

- 1911 (original text of the older electronic edition);
- 1935 (Roget's International Thesaurus 1st edition—first printed in 1922);
- 1946 (RIT 2nd); 1962 (RIT 3rd—the main data source for this study);
- 1977 (RIT 4th);
- 1992 (RIT 5th).

This is not an exhaustive list of traditional (direct descendant) Roget's Thesaurus editions, only editions owned by the author. They are however representative of the major revisions. The division between the British and American editions developed when in 1886 the publisher Thomas Y. Crowell adapted the thesaurus to include standard American terms, (considered by the British editors to be "Americanisms"), adding U.S. slang and colloquial terms, and classifying British terms as "Briticisms." This became the "international" edition in 1922.

The first American (1886) edition is virtually identical to the 1879 British Edition (edited by Roget's son, John Lewis Roget) "from which it was ultimately derived" (Berry, 1962, Publisher's Preface to the American/International edition). The succeeding 1911 American edition was claimed by the publishers to be "practically a new book" (Berrey, 1962, Publisher's Preface). The editor, Sylvester Mawsom, was Associate Editor of Webster's New International Dictionary, Webster's Collegiate Dictionary and "Consulting Specialist" (1935 edition frontispiece) to Sir James Murray, editor of the Oxford English Dictionary; so there is no doubt that any changes made were well informed. The publishers claimed it differed from its predecessors in many ways that every edition of a thesaurus claims—essentially that it was "bigger and better"—they added 22 Categories (containing about 680 words). They specifically list eight ways in which it differed but in one way it truly does differ. Terms were added that were "archaic or obsolescent in England, but of everyday usage in America" (Mawsom, 1911, Editor's Preface) and there was a "characterization of all obsolete, obsolescent, rare, archaic, colloquial, dialectal, and slang words, as well as British, foreign, and special terms, *as is done in all the best dictionaries*" [emphasis added] (Mawsom, 1935, Preface). The second quotation is taken from the 1935 edition in preference to the 1911 edition as Mawsom there revised the list to be more specific, and eliminated discussion and debate found in the 1911 edition about what constitutes an Americanism; and terms such as "Negroisms."

The annexation of Roget's Thesaurus by an unrelated American publisher did not meet with the approval of the British publishers so in 1933 an edition edited by Roget's grandson, Samuel Romilly Roget, was produced with the subtitle "Authorised American Edition." It was published by the American branch of the original London publisher, Longman, Green and Co.—Grosset & Dunlap, NY. In the foreword S. R. Roget comments,

> In the course of the years there have been several competing editions printed in America, all based on the London editions, but from none of these did the author or his representatives derive any pecuniary advantage (Roget, S.R., 1933, Foreword)[38]

Despite this attempt at an American version, the British publishers and Roget's heirs did not displace C. O. S. Mawsom's legacy—the "International" version prevailed and the two versions of the thesaurus developed thereafter independently[39]; the American branch ultimately resulting in the 1962 American/International edition used for this study.

[38] In other words it was hijacked and no compensation was paid.

[39] Though each claimed the copyright (Piotrowski, 1994, p. 135); and borrowed, copied, and stole the other's words and ideas flagrantly—and still do.

Appendix B: Methods of Data Analysis

The primary method of analysis for identifying the implicit structure of RIT in this research has been visualization. Most of the actual data processing for the research was done using Structured Query language (SQL), for example to generate the semantic neighborhoods, descriptive statistics, data for visual displays, coordinates for information maps, and contexts (matrices) for formal concept analysis (FCA) lattices.

Multi-dimensional Scaling (MDS)

In order to view localized data sets (fewer than 100 variables), where a proximity measure such as a relevance metric between words (Old, 1996) can be applied or a correlation matrix can be derived, multi-dimensional scaling (MDS) has been used (via the SPSS statistical software package). Proximity measures are used to define an index over pairs of objects (such as synonyms or Categories) that quantifies the degree to which the two objects are alike—MDS assumes an analogy between the psychological concept of similarity and the geometric concept of distance.

MDS is a standard multivariate statistical method but which involves expertise and to which many caveats apply. The natural dimensions of the data are reduced to only two or three—implying a loss of a great deal of data—and the procedure involves the automatic estimation of parameters and estimation of fit for various spatial distance models. Consequently MDS has been used sparingly and only where the fit has been good and the pattern of relationships illustrates what can be substantiated using actual (raw) data.

Formal Concept Analysis (FCA)

Where a topological rather than topographical model is appropriate (where distances between words or senses are not so important as relatedness or broad patterns of connectivity) network or graph visualizations are used. For example tree structures are used for viewing hierarchies, and formal concept analysis (Wille, 1989) is used to derive structural relationships of *semantic neighborhoods* (defined and described further, below) where both graph duals[40] are relevant—that is, where it is most useful to represent relationships between senses, relationships between words, and relationships between words and senses concurrently. Formal Concept Analysis (FCA) is based on the Galois connection, which results in a lattice—a type of partial ordering on two sets where a relation exists between the two sets (this is fully defined in Wille's 1989 paper). A type of graph called a Hasse diagram is used to represent the resulting lattice but, in general, the diagram is simply referred to as "a lattice."

As an example of a lattice, the cross-table, or "formal context" in Figure B1 represents the relation between three objects and seven possible attributes. The objects (animals in this case) could have been senses from RIT, and the attributes could have been words—or vice-versa, as the relation is symmetric. An "X" identifies that a relation exists (birds and

[40] A relation between two sets, such as Words and Senses, can be represented in a graphical network (a graph) using either the nodes to represent words and the links between them to represent the shared senses, or using the nodes to represent senses and the links between them to represent their having words in common. These are the duals.

fish lay eggs), while a blank shows that no relation exists (birds do not have teeth). Laying eggs, having warm blood and having teeth are shared attributes (more than one animal has each attribute), while flying, swimming, and walking are (in this simplistic example) idiosyncratic to a particular animal. The idiosyncratic attributes, or differentiating features, are the type that would be used to classify objects in a tree structure or hierarchical classification such as the RIT and WordNet hierarchies.

	Bird	Fish	Mammal
Alive	X	X	X
Lays Eggs	X	X	
Warm-blooded	X		X
Has Teeth		X	X
Flies	X		
Swims		X	
Walks			X

Figure B1. **Formal Context of animals**

The lattice in Figure B2 contains the same information as the context. By following lines down from an object, its attributes can be read. For example birds are warm-blooded, lay eggs and are alive. Symmetrically, by following lines up from an attribute, those objects that have that attribute can be read. For example, all three animals are alive; both birds and mammals are warm-blooded; and only a fish swims.

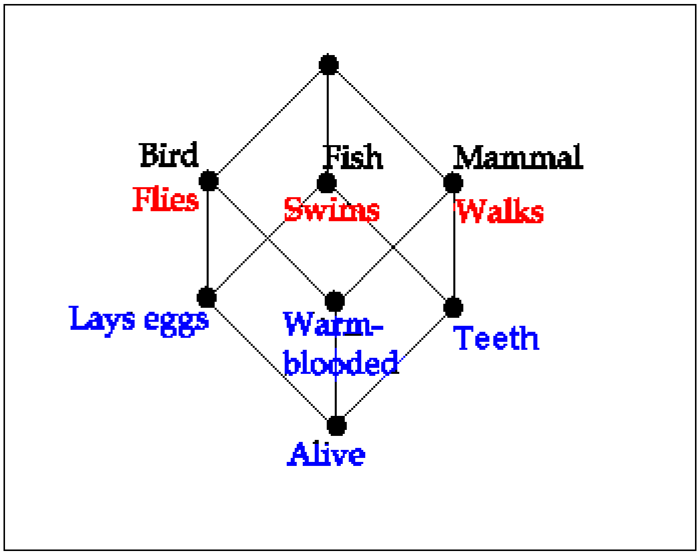


Figure B2. **Lattice generated from the Formal Context in Figure B1**

Each node is called a *formal concept* and consists of two sets, the set of objects (its extent) and the set of attributes (its intent)—these are the sets of all objects and attributes which can be read off the lattice, up and down, starting at that node. For example the concept labeled “warm-blooded” has the extent {bird, mammal} and the intent {warm-blooded, alive}. In other words, that concept represents the fact that (in this limited context) the animals (objects) that are (have the attributes of being) both warm-blooded and alive, are {bird and mammal}; and conversely and symmetrically, the attributes shared by the objects bird and mammal, are {warm-blooded and alive}.

A *semantic neighborhood*, like a word field (Wortfeld) or semantic field (Bedeutungsfeld)[41], refers to a set of semantically related words. For this research a semantic neighborhood differs in that both words and senses are included, and is defined here as the senses of a given word, along with the synonyms for that word for each of the its senses. A *restricted neighborhood* is defined here as a neighborhood which includes only those synonyms that occur in more than one sense of the word. A *sense neighborhood* can be defined symmetrically for senses. In either case "sense" may be replaced with any concept at any level in the RIT hierarchy, and "word" may be replaced by a set of words. The set of words can be from different Synsets. So, for example, the restricted neighborhood of the set {over, above} will include just those words which are synonyms of both words, and just those senses which are shared by both words. *Neighborhoods* are further defined, formally, in Priss, 1996 (pp. 38-39).

Information Cartography

Information cartography (Old, 2002) has been used for the concurrent visualization of multiple feature dimensions and global patterns within Roget's Thesaurus, and for comparisons between Roget's and relevant other lexical databases such as WordNet, Indo-European roots, or Chinese language data. Information cartography is a dynamic, interactive process, and conclusions (relevant configurations) or illustrative examples have been exported as information maps for inclusion in this dissertation.

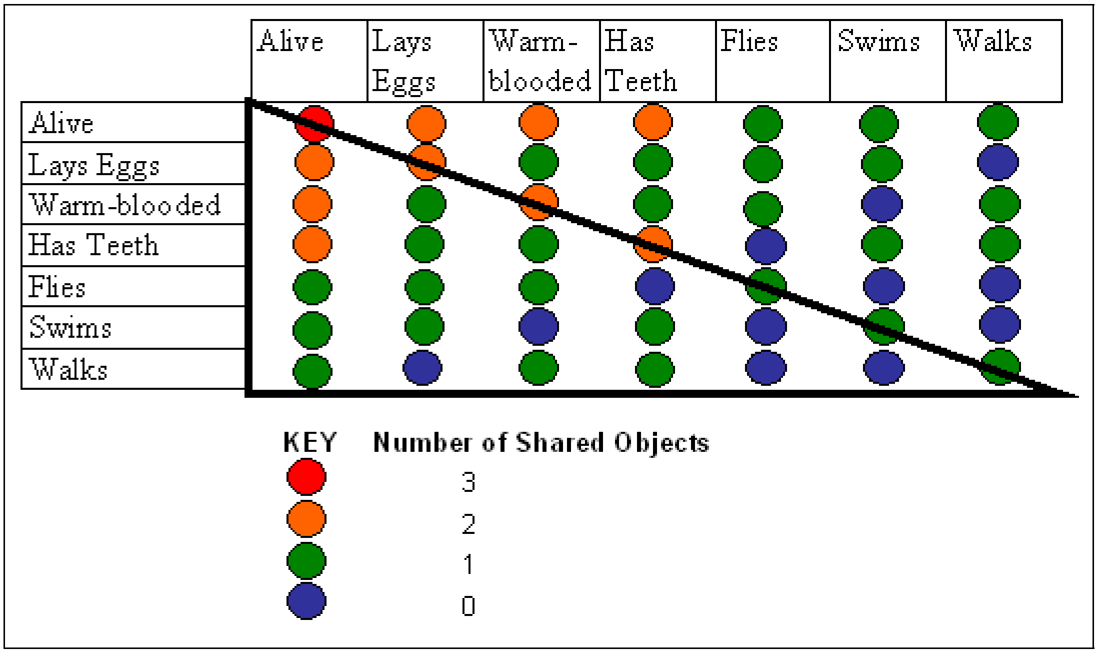
The most common information map display used in this research to illustrate the inner structure of RIT is the two-way contingency table (Snee, 1974). Like the formal contexts of concept lattices a contingency table is an incidence relation where the existence of a relation between an item on the x-axis (or a column) and an item on the y-axis (or a row) is indicated by the existence of a symbol. Where an "X" is used for a formal context, a point or colored disk is used in these information maps. The color ranges are chosen to represent value ranges and are automatically generated and indexed using Geographic Information System (GIS) software. Where the relation is between elements of the same set (a self-referencing relation), the set is duplicated on both axes. Because this is used in this research to represent large datasets (more than 1,000,000 points for the Category-Category relation), the labels are usually omitted.

Figure B3 illustrates a symmetric two-way contingency table-form information map of an attribute-attribute relation between the attributes in the *animals* context of the previous discussion on FCA. The points are colored to reflect the number of objects each attribute shares in common with its partner. The diagonal line of points (identity diagonal),

[41] According to Elsen (2001), Barrett (1982) was the first to use the term "semantic field;" The first to use "Bedeutungsfeld' was Ipsen (1924); and the first to us "Wortfeld" was Trier (1931).

representing the relation between attributes with themselves, is usually ignored. The triangle simply illustrates that the information is identical above and below the diagonal when the X and Y axes are the same i.e., as stated earlier, when the relation is between elements of the same set.

As the blue point represents relations between attributes which share no objects it can be seen that the attribute *walks* shares no objects with the attribute *lays eggs;* the attribute *swims* shares no objects with the attribute *warm-blooded;* and the attribute *flies* shares no objects with the attribute *has teeth.* The dimension of "number of shared objects between attributes" in this example may be of no particular interest, but the number or type of shared words, or number or type of shared attributes between Categories or Senses in RIT, globally, is of interest. In these cases patterns emerge analogous to the line of blue points in this information map which, if interpreted correctly, have implications for the tacit information contained in the thesaurus.



	Alive	Lays Eggs	Warm-blooded	Has Teeth	Flies	Swims	Walks
Alive	🟥	🟧	🟧	🟧	🟩	🟩	🟩
Lays Eggs	🟧	🟧	🟩	🟩	🟩	🟩	🟦
Warm-blooded	🟧	🟩	🟧	🟩	🟩	🟦	🟩
Has Teeth	🟧	🟩	🟩	🟧	🟦	🟩	🟩
Flies	🟩	🟩	🟩	🟦	🟩	🟦	🟦
Swims	🟩	🟩	🟦	🟩	🟦	🟩	🟦
Walks	🟩	🟦	🟩	🟩	🟦	🟦	🟩

KEY	Number of Shared Objects
🟥	3
🟧	2
🟩	1
🟦	0

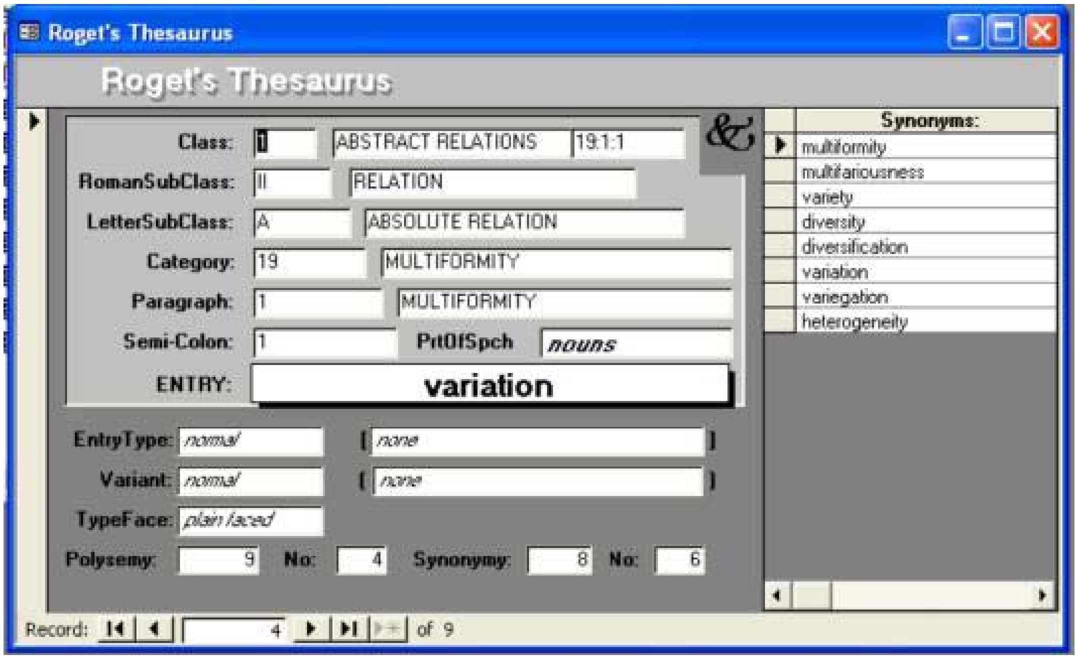
Figure B3. **Information map based on a two-way contingency table of the relation between attributes, color-indexed to represent number of shared objects.**

Interfaces for Data Browsing

In addition to the visualization methods of analysis, the data will be browsed via three sets of interfaces:

- Structured Query language (SQL) -based database interface
- GUI forms-based database interface
- Web-based interface.

An example from the forms interface, displaying one entry for the word "variation," is given in Figure B4.



Roget's Thesaurus

Roget's Thesaurus

Class: 1 ABSTRACT RELATIONS 19:1:1

RomanSubClass: II RELATION

LetterSubClass: A ABSOLUTE RELATION

Category: 19 MULTIFORMITY

Paragraph: 1 MULTIFORMITY

Semi-Colon: 1 PrtOfSpch *nouns*

ENTRY: **variation**

EntryType: *normal* [*none*]

Variant: *normal* [*none*]

TypeFace: *plain faced*

Polysemy: 9 No: 4 Synonymy: 8 No: 6

Synonyms:
multiformity
multifariousness
variety
diversity
diversification
variation
variegation
heterogeneity

Record: 4 of 9

Figure B4. **Form interface to Roget's Thesaurus database**

Attribute information such as polysemy (value: 9), sequence number (4th of 9 entries in RIT), synonymy of the Synset (value: 8), and the actual eight synonyms contained in the current Synset, can be read from the display. The Internet interface contains multiple standard queries and graphical interfaces. Figure B5 shows the explicit structure of the Abstract Relations Class (1911 Edition) viewed through a hyperbolic tree interface. Navigation follows the direction of the red arrows.

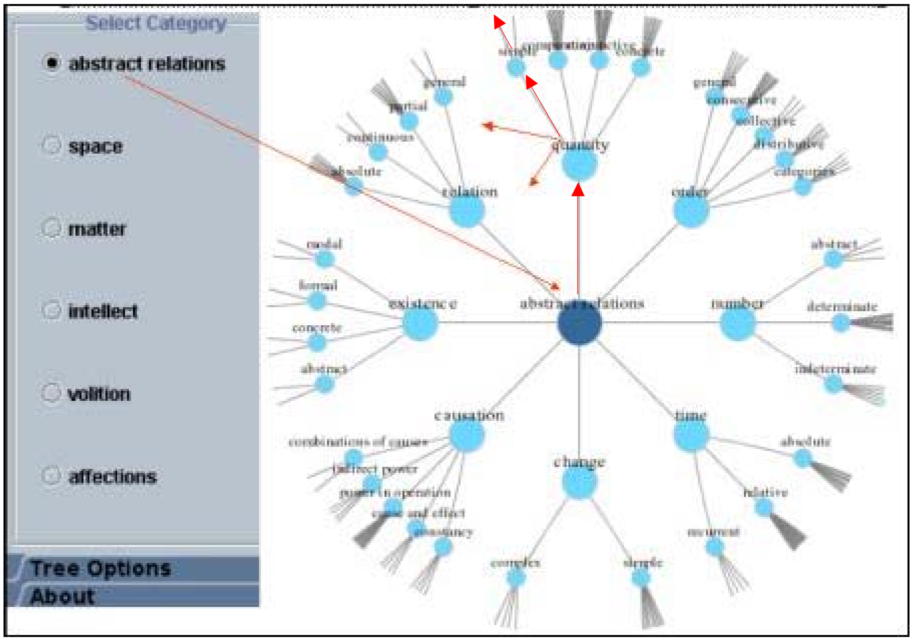
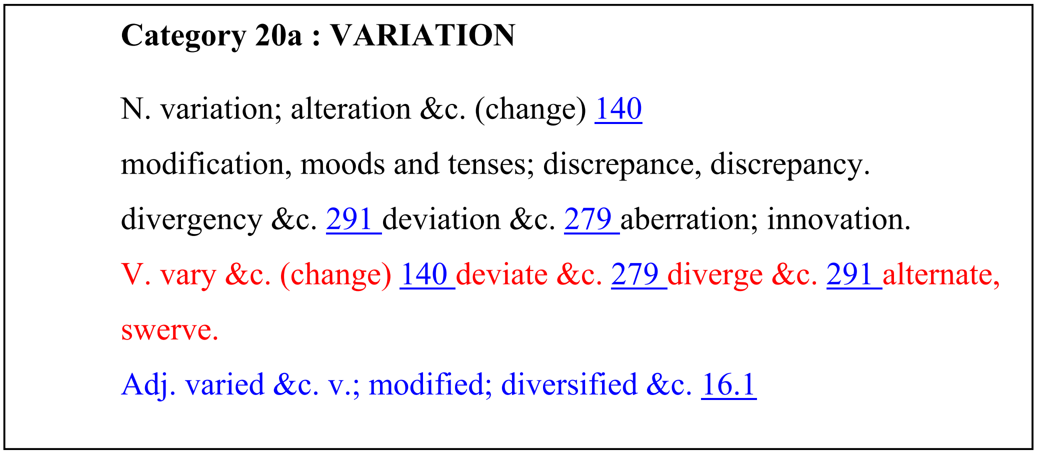


Figure B5. **Hyperbolic tree of Class 1: Abstract Relations**

The leaf nodes of the tree are Categories and a mouse click invokes the on-line database, which then displays the Category information in HTML format as in Figure B6.



Category 20a : VARIATION

N. variation; alteration &c. (change) 140

modification, moods and tenses; discrepance, discrepancy.

divergency &c. 291 deviation &c. 279 aberration; innovation.

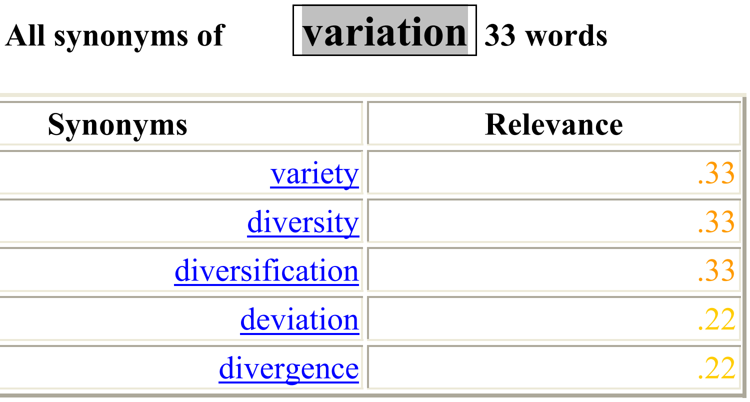
V. vary &c. (change) 140 deviate &c. 279 diverge &c. 291 alternate, swerve.

Adj. varied &c. v.; modified; diversified &c. 16.1

Figure B6. **Category information resulting from a hyperbolic tree query**

A similar interface exists for the 1962 Edition of the thesaurus.

Text queries may also be made against the on-line database. For example, all of the synonyms of a word for all of its senses in the thesaurus can be displayed, ranked by a relevance metric (calculated from the number of shared senses). Figure B7 displays 9 of the 33 results returned for the word “variation.” The synonyms are hypertext links which, if mouse-clicked, will retrieve equivalent results for that word. The relevance is calculated using the number of senses shared with the input word.[42]



All synonyms of **variation** **33 words**

Synonyms	Relevance
variety	.33
diversity	.33
diversification	.33
deviation	.22
divergence	.22

Figure B7. **Ranked output of synonyms resulting from a web-based text interface query**

[42] The equation for the relevance metric, an alternative equation, and a discussion of the related issues can be found in Old, 1996.

Appendix C: Core RIT Prepositions and their Indo-European Roots

Core Prepositions

(prefixed—**unprefixed**)

about	before	***down***	**through**
above	behind	**for**	**to**
across	below	**from**	**under**
after	beneath	**in**	**up**
against	beside/besides	***near***	**with**
along	between	**of**	
amid	beyond	**off**	
among/ amongst	**by**	**on**	
around		**out**	
at		**over**	

Indo-European Roots of Prepositions[43]

Root	**Meaning**	**Descendants**
(A)MBHI-	around	***by****, be-, abaft*
(A)PO-	away, off	***of, off, after****, ab-, opposite*
AD-	at, to, near	***at****, ad-, a-*
AL-1	beyond, other	(*ultra-, alter-, ultimate, utter, alias, alien, other, else …*)
AN-1	on	***on***
ANT-	front, forehead (> before, opposite)	(*along, ante, anti-, un-*)
DE-	this, that, that way, to	***to****, into (*and *too)*
DEL-1	long	*long,* ***along***
DHUNO- (Gc)	fortified place, enclosure	***down***
DWO-	two	***between****, betwixt*
EGHS-	out	(*ex-, ecto-, extra-, exotic, except, extreme, strange, stranger …*)
EN-	in	***in****, and* (also Latin *inter-, intra-*)
ETI-	above, beyond	(*eddy, eider (duck);* also Latin *et,* "and")
GAGINA- (Gc)	in a direct line with	*again,* ***against***
GHEDH-	join, fit, gather, unite	*together (gather)*

[43] Format Key: Roots:- REGULAR; *GERMANIC;* definition. Prepositions:- ***core****; (other)*

GWRES-	fat, thick	*cross,* ***across***
I-	a pronominal stem	*yon, yonder,* ***beyond*** (also *ilk, id, if, yes,* and *yet.*)
KAP-	grasp, hold	*except*
KO-1	this, that	*hind,* ***behind,*** (*hinterland, here;* also, *he, her, him, it*)
KOM-1	beside, near, with, by	*co-, con-, com-, contra-*
KSUN-	with, together	*syn-, sym-*
KWO-	who, what, where, when…	*… which, how, either, quality, quantity*
LEGH-	lie, lay	*low,* ***below*** (and *lair*)
MAG-	knead, fit, fashion	***among***
MEDHYO-	middle	***amid***
NDHER-	under	***under***, *infra-, inferior*
NEHW-IZ- (Gc)	near	***near***, *next*
NER-1	under, on the left, left, left of eastward	*north, northern, Norman, nordic, Norse, Norway, Norwegian*
NEWO-	new	*new*
NI-	down	***beneath***
NU-	now	*now*
PER-1	forward, through	***for***, ***before, from***, *far, farther, further…*
PERh-2	grant, allot	*part, apart, apart from*
PETh-	spread, stretch out,	*past*
RET-	run, roll	*round,* ***around***
SE-2	long, late	*side,* ***beside***, *…*
TERh-2	cross over, overcome, pass	***through***, *throughout, trans-*
UD-	up, out	***out, about***
UPER-	over	***over***, *super-, superior, hyper-*
UPO-	under, up from under	***up, above***, *hypo-, sub-*
WEGH-	go, transport in a vehicle	*way, away, away from*
WER-3	turn, bend	*toward, forward, seaward*
WI-	in half, apart	***with***

VITA

L. John Old

Date of Birth	October 5, 1950
Permanent Address	38 Craiglockhart Terrace, Edinburgh EH14 1AJ, Scotland
Email	j.old@napier.ac.uk

EDUCATION

Ph.D. Program: Information Science; Cognitive Science (minor) 1997 – present
Specialization: Information Systems and Databases
Indiana University, Bloomington

Master of Science, Computer and Information Science 1993
Specialization: Artificial Intelligence
University of Arkansas at Little Rock, Arkansas

Bachelor of Science, Computer Science; Minor in Psychology 1987
Waikato University, New Zealand

Associate of Science, Computer Science 1985
University of Arkansas at Little Rock, Arkansas

PROFESSIONAL

Lecturer, School of Computing, 2002 – present
Faculty of Engineering and Computer Science,
Napier University, Edinburgh, Scotland
