The Semantic Structure of Roget’s Thesaurus Cross-References

L. John Old
Edinburgh Napier University
Scotland, United Kingdom

Abstract This study analyzed a database version of Roget’s Thesaurus (Roget’s International Thesaurus, 3rd Edition, 1962) for connectivity patterns among cross-references in order to identify the implicit conceptual structure. Semantic patterns implicit in the data, at both the local and global levels of the Thesaurus structure, are identified.

1. *Introduction*

This research follows conceptually from the work of W.A. Sedelow, Jr. and S. Yeates Sedelow (1979, 1986, 1990-1993), Priss (1996) and Old (2003), on Roget’s International Thesaurus (RIT: Berrey 1962). Patterns among local views of RIT, such as for example, semantic neighbourhood lattices (Priss, 2005); patterns emerging from global views of RIT such as word-overlap (with implied semantic overlap) between Categories (Old, 2002); and conceptual and semantic *hubs and authorities (semantic switching centres)* among senses and words (Steyvers & Tenenbaum, 2005) have previously been identified and readily represented. Roget’s Thesaurus cross-references, however, which form a kind of shadow, or skeletal network structure of the implicit structure of the Thesaurus as a whole, have not been studied in the same way.

2. *The Explicit Structure of Roget’s Thesaurus*

The explicit structure of Roget’s Thesaurus is a hierarchy, or tree, implemented in the book in three main parts. Following the front matter is the top level of the *hierarchy* represented by what Roget called the tabular *Synopsis of Categories.* The Synopsis lists the structure down to the level of the 1,000 or so Categories (also called headwords, or lemmas, by some researchers). Most of the categories are arranged in opposed pairs, where the meanings of the pairs are antonymous. For example, 27 Equality versus 28 Inequality, and 648 Goodness versus 649 Badness.

The Synopsis is followed by the *body*, or *Sense Index* of the book, which continues the hierarchy down to the lowest levels. The Sense Index lists the Categories representing the notions found at the bottom level of the Synopsis. Each Category contains the actual entries—instances of words, ordered by part-of-speech and grouped by sense, or *synset* (Miller et al., 1993). Synsets are grouped into broader notions, as paragraphs. The entries are commonly referred to as *synonyms*, though frequently there are other semantic relations at work. For example, the part-whole relation of *meronymy*, as illustrated by “parts of a ship” or “historical eras”. At the back of the book is the *Word Index*, listing the words in alphabetic order, along with references to their senses in the Sense Index, ordered by part-of-speech.

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 sense index numbers, not members of the set of Words, and represent a relationship between their own synset and semantically related, but remote, synsets.

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 cross-references can be reciprocated between categories). A further idiosyncrasy is that a cross-reference points not to a single sense, but to a set of senses—always at either the Paragraph level or Category level.

The example in Figure 1 shows a cross-reference from Category 1: *Existence*, Paragraph 2: *Reality*, (found in the third synset of the paragraph); this references Category 515: *Truth*, Paragraph 5: *Genuineness*.

1 EXISTENCE

NOUNS **1**. existence, subsistence, being; entity, essence; presence, occurrence; life 406.

2. reality, actuality, factuality; truth 515; authenticity **515.5**; sober reality, grim reality, no joke, not a dream; thing-in-itself, ultimate reality. …

Figure 1. An example of a cross-reference in situ [1:2:3 – 515:5].

The referencing (or source), and destination, or target, synsets here share a word in common. The source synset includes {**authenticity 515.5**}. The word *authenticity* is an "anchor" word, in the same way that an Internet hyperlink contains both the link (an HTTP reference to a remote location) and anchor text, which usually describes the target of the link. The first synset of the referenced, destination, or target Paragraph, **515:5**, {genuineness, **authenticity**…realness, reality…} also contains *authenticity*. This mechanism occurs frequently among regular (called hereafter, *normal*) Roget's Thesaurus cross-references. However, cross-references do not necessarily indicate shared strings (words in common) between the source and destination locations. The main purpose of a cross-reference is to indicate shared meaning, not shared words.

3. *Types of Cross-References*

Instead of a *normal cross-reference*, a synset may contain several cross-references pointing to a sequence, or range of senses. For example, a cross-reference found in Category 299: *Arrival* (see Figure 2), 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

Figure 2. A range cross-reference: 291:4:1 – 923:2-4

This type of cross-reference (belonging to a set of two or more sequential cross-references) we call a *range cross-reference.* A cross-reference to a whole Category, a *Category-only cross-reference* appears in the text without a *Paragraph index*. An example is seen in Figure 1 as "… ; life **406**." which 'points' to Category 406: *Life*.

Usually there is an *anchor word* in the source location that is the actual name of the destination Category, as the previous examples referencing Category 515 *Truth* and Category 406 *Life*. However, about twenty-percent of cross-references do not anchor on the name of the destination category. Examples are:

Anchor	Source	Destination
explanation **550**	543:3:1 Meaning	550 Interpretation
sameness **14**	30:2:2 Equality	14 Identity

When a set of cross-references from one location refers to multiple locations in a remote Category there is clearly a very strong semantic relationship between the two Categories. In the following example the cross-references serve as links between similar concepts listed under the equivalent parts-of-speech in each Category. This type of cross-reference (belonging to a set of two or more concurrent cross-references) can be seen as *concurrent cross-references*. Note that, in this case, no character strings, or words, are shared between the source and destination.

Source	Paragraph	Destination	Paragraph	POS
763:6:3 Submission	Submit	764:2 Obedience	Obey	Verbs
763:12:2 Submission	Submissive	764:3 Obedience	Obedient	Adjectives
763:17:2 Submission	Submissively	764:6 Obedience	Obediently	Adverbs

.

A final cross-reference type 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 for the 1911 Edition), but is now quite rare. The goal was to link a concept in one part-of-speech section to a more-specific set of words relating to the same concept. An example is found within Category 123: Oldness, in synset 123:4:3 containing the entry "archaeology." This references synset 123:22, which lists branches of archaeology such as "paleoanthropology, paleohydrography, paleolatry, paleolithy, paleometeorology," and "Egyptology."

Antonymous notions are classified, via the Synopsis, in adjacent Categories, so the Editors may have considered such references to be redundant. There are, however, twelve cross-references between adjacent Categories linking such complementary concepts as *bequest* and *inheritance.*

4. Descriptive Statistics

There are approximately 3,772 cross-references in RIT. Of these, 3,171 point to Paragraphs and 601 point to whole Categories. Of the Paragraph-referencing cross-references there are 637 of the range cross-reference type; 1,313 of the concurrent cross-reference type; and 1,164 of the normal cross-reference type.

102 categories are not involved in any cross-references. Of the categories involved in cross-references, there are three types:

- 114 categories contain cross-references, but are not ever referenced
- 136 are referenced by cross-references but contain no cross-references
- 691 categories both reference, and are referenced by, cross-references

Ten percent (335) of the cross-references are of the first type, while ninety percent (3,437) are of the third type. Of course the second type has no cross-references, but they do involve ten percent of cross-references destinations.

5. The Implicit Structure of Roget's Thesaurus

This section describes and illustrates the results of analysis, and patterns, found among the RIT cross-references that imply structures other than those described in Section 2*,* above. This section begins with a brief discussion of the implicit structures discovered in previous research through the analysis of patterns of words and senses in Roget's Thesaurus.

5.1.The Non-Cross-reference Implicit Structure

A *Small-world* model can be utilized to account for much of the implicit structure of Roget's Thesaurus. The model derives (Travers & Milgram, 1969) 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.

Small-worlds may be characterized by particular measures. Word association data has about (on average) 3.0 degrees of separation. Old (2000) shows that Roget's Thesaurus satisfies the criteria of being a small-world network, and Young (1993) shows that the neural network of the brain also fits the criteria. Other researchers (Steyvers & Tenenbaum, 2005; Motter, de Moura, Lai, & Dasgupta, 2002) find 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) shows that the World Wide Web is also a small-world. Because URLs are directed (links go in only one direction) Kleinberg classifies the highly connected nodes (Web sites) into those that link to many Web pages and those that are linked to by many Web pages. The Google search engine also uses this principle. The small-world model suggests the probability that the underlying meanings 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. Roget 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 of Roget's Thesaurus, John knew this was a problem (see Section 7, *Fan-in and Fan-out; Semantic Hubs and Authorities,* below).

5.2. Cross-Reference Patterns

As described in the discussion of the RIT cross-references in Section 3, there are five types of cross-references, termed here *, normal-, category-, range-, concurrent-*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). A reference corresponds to a regular RIT cross-reference or Internet hyperlink (a URL on a Web page).

It is easy, looking at a particular published document, to see what other papers that document references, but impossible to see what citations it has (who has referenced it). 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 cross-reference back-links are not a semantic relation. In support of this view, the thesaurus editors supply return-, or reciprocal-cross-references in about only one third of the cases.

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 or more specific notion, not just to the same or similar 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.

Figure 3 shows an implicit structure formed by reciprocated cross-references chains (Length 5), many of which are between Categories of different classes. This illustrates the type of coherence that exists across the thesaurus, but which is not available through any explicit structure or organization of the thesaurus.

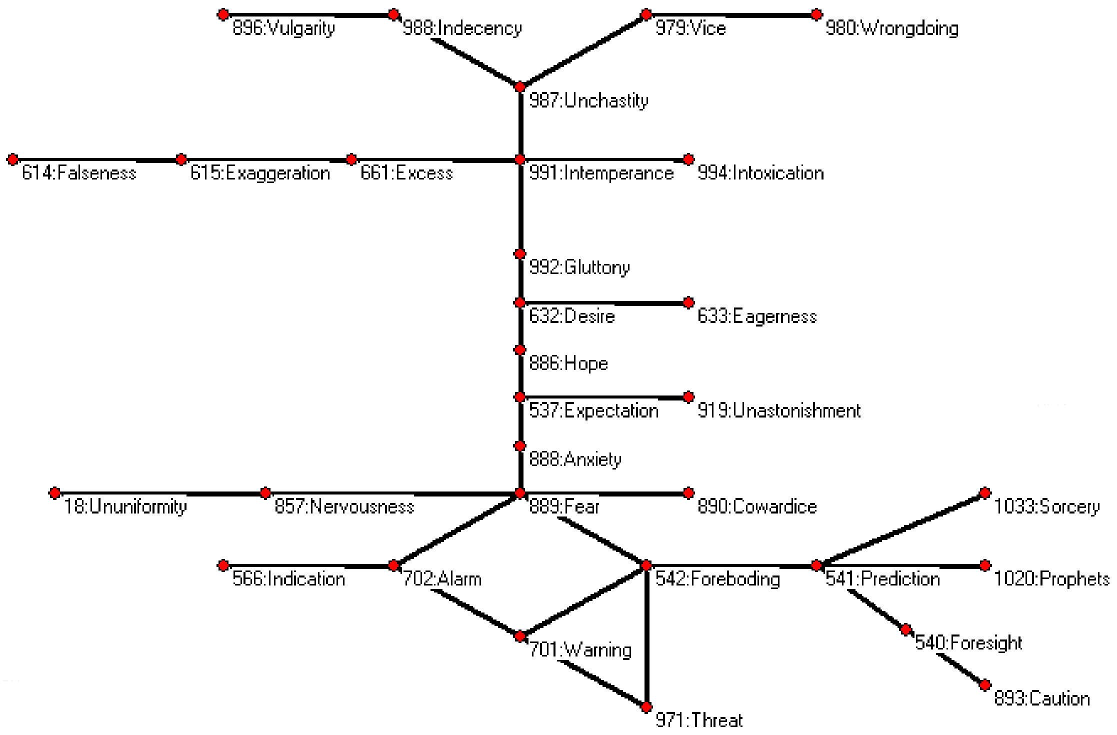


Figure 3 Graph combining all chains of reciprocated cross-references of at least five nodes length

6. Implications of Cross-references among Upper Levels of the Hierarchy

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.

Class-crossing, reciprocated, whole-Category links are represented by the links in the graph in Figure 4, labelled 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 is shown in Figure 5. The example suggests that, despite the fact that their semantics and words overlap (as evidenced by the strong, reciprocated, whole-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 6 to support this observation. Whole-Category links between the Affections Class and Intellect Class, 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.

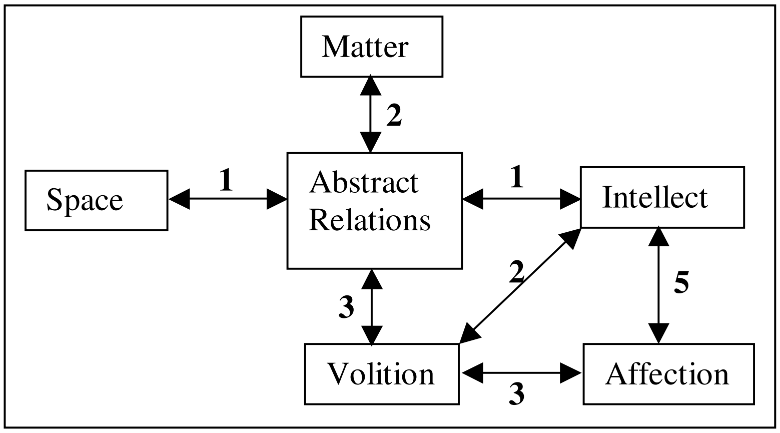


Figure 4. Whole-Category reciprocated links crossing Class boundaries

Class level:	**7: Affections**		**C6: Intellect**
Category Level:	**867: Discontent**	⇔	**539: Disappointment**

Figure 5. A reciprocated, whole-Category link at the Class level

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."

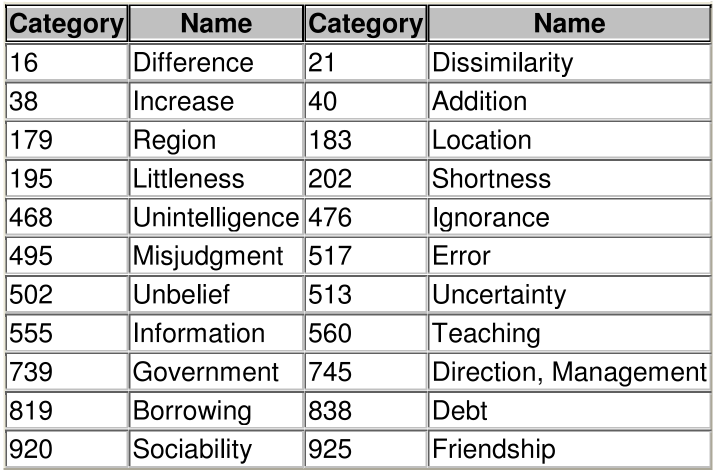
Similar analysis can be made of the second level, or Roman-level Classes, of the hierarchy; and the third-level, or Letter Classes. Examples of strong, reciprocated, whole-Category links that cross only Letter Class boundaries (both derive from the same top level Class and Roman-level Classes) are given in Figure 7.

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 6. Reciprocated, whole-Category link shown at all Class levels

At this lower level of the hierarchy Categories are related only by the fact that they share the very broad notion of their Roman-level 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.

The relationship unearthed by reciprocated, whole-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 warrant 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 have only 10 words in common.

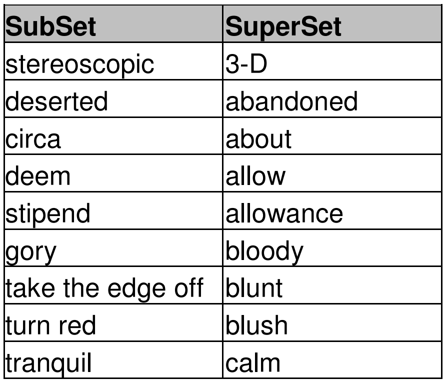


Category	Name	Category	Name
16	Difference	21	Dissimilarity
38	Increase	40	Addition
179	Region	183	Location
195	Littleness	202	Shortness
468	Unintelligence	476	Ignorance
495	Misjudgment	517	Error
502	Unbelief	513	Uncertainty
555	Information	560	Teaching
739	Government	745	Direction, Management
819	Borrowing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

Figure 7. *Reciprocated, whole-Category* links that cross Letter Class boundaries only.

6.1.Implications

The idea of "set implication" suggests that subsets imply their supersets. In general, a word in RIT that has a set of senses that is a subset of senses of a second word, implies the second word. In Figure 8, the words on the left have fewer senses than the words on the right, and those senses are a subset of the senses where the words on the right are found, in RIT. So the words on the left imply the words on the right.



SubSet	SuperSet
stereoscopic	3-D
deserted	abandoned
circa	about
deem	allow
stipend	allowance
gory	bloody
take the edge off	blunt
turn red	blush
tranquil	calm

Figure 8. Set implication between RIT words.

The words on the left are rarer (have fewer senses) and are more specific, while the words on the right are more polysemous (have more senses), and are more general. For native English speakers the words and phrases on the left tend to be less familiar. Consequently those on the right tend to be explanatory.

Implications can form chains: *poodle* and *terrier* both imply *dog*; *dog* 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. Implication associated with cross-reference is illustrated schematically in Figure 9.

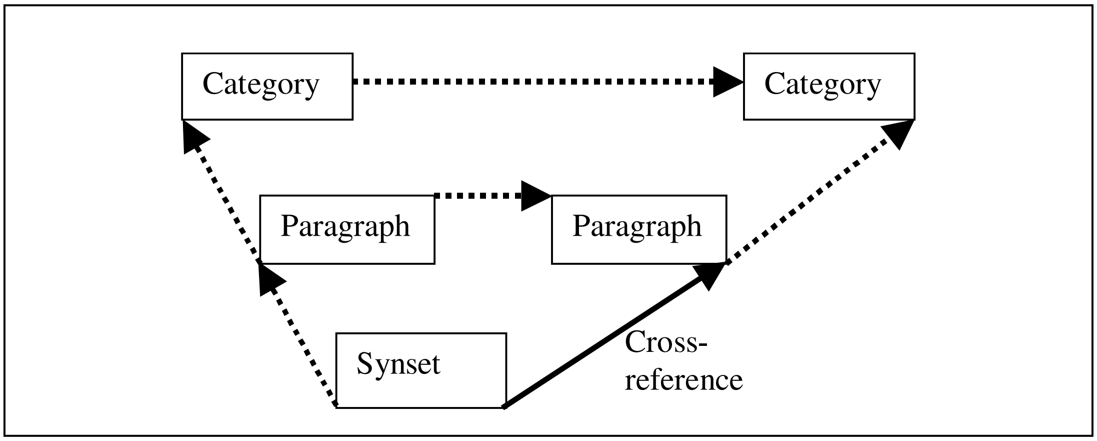
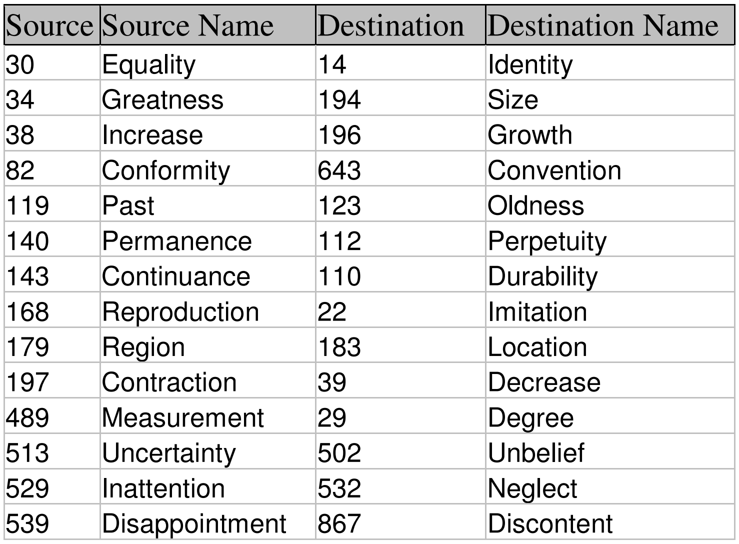


Figure 9. Implications (dotted lines) in cross-reference (solid line)

The source synset of a cross-reference is almost always a smaller set of concepts (96%) than the destination of the cross-reference. It does not, however, always contain 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 implication in cross-references—the source Category concept implies the destination Category concept (there is an inference from the source to the destination). Figure 10 illustrates this (the source Category concepts are on the left).

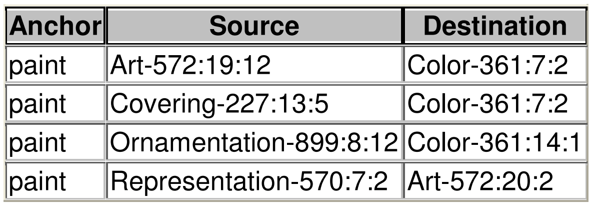


Source	Source Name	Destination	Destination Name
30	Equality	14	Identity
34	Greatness	194	Size
38	Increase	196	Growth
82	Conformity	643	Convention
119	Past	123	Oldness
140	Permanence	112	Perpetuity
143	Continuance	110	Durability
168	Reproduction	22	Imitation
179	Region	183	Location
197	Contraction	39	Decrease
489	Measurement	29	Degree
513	Uncertainty	502	Unbelief
529	Inattention	532	Neglect
539	Disappointment	867	Discontent

Figure 10. Implication in cross-references

As mentioned earlier, many Categories participating as sources and destinations of cross-references share the anchor term.

Figure 11 shows the cross-references anchored on the entry *paint*.



Anchor	Source	Destination
paint	Art-572:19:12	Color-361:7:2
paint	Covering-227:13:5	Color-361:7:2
paint	Ornamentation-899:8:12	Color-361:14:1
paint	Representation-570:7:2	Art-572:20:2

Figure 11. Cross-references anchored on the term "paint"

Examples of cross-references that do not share the anchor term are Category 137: Regular Recurrence, anchored on *holy days* referencing a Paragraph in Category1038: Religious Rites, that contains a list of holy days (but not the actual term "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 clearly method to these cross-references.

It is noteworthy that the average polysemy of cross-reference anchors is 6.33 senses, while the average polysemy of Thesaurus entries, in general, is about 2.3 senses. This is probably a consequence of the fact that normal cross-references indicate further senses of a polysemous word, and exclude words with only one sense (about 40% of all words).

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

The examples of cross-reference types given earlier have all been semantically strong cross-references. The majority of cross-references, however, are of the weaker types—pointing 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 counts of cross-reference are referred to as *fan-in* and *fan-out*[1] (*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.

Figure 12 and Figure 13 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,*

[1] Also known by the terms: in-degree and out-degree.

uncertainty, displeasure (existing in both sets), and other notions with negative connotations such as *disease* and *weakness*. The most common themes for *authority* Categories in RIT are (considering all Categories, and totaling cross-references at the Roman Class level):

- 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.

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. Almost the same Roman Classes appear for the hub Categories, except that V*olition in general* disappears.

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

Figure 12. 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

Figure 13. 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-level Class. This suggests strong coherence within Roman Classes. Figure 14 lists examples of the links running from hubs to authorities that cross Roman Class boundaries. This illustrates the type of coherence that exists between the Roman-level Classes.

RC1	Cat1	Category Name1	ParaName1 <	Shared Word	> ParaName2	Category Name2	Cat2	RC2
2.IV	270	Transference	carrier	letter carrier	postman	Messenger	559	**6.III**
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	495	Misjudgment	misjudge	misconstrue	misinterpret	Misinterpretation	551	**6.III**
7.I	629	Avoidance	dodge	shrink	pull back	Reaction	283	**2.IV**
7.I	655	Way	passage-way	inlet	place for entering	Ingress, Entrance	301	**2.IV**
8.I	860	Impatience	impatient	impetuous	impulsive	Impulsiveness	628	**7.I**
8.I	881	Dullness	triteness	cliché	platitude	Maxim	516	**6.I**

Figure 14. Links running from hubs to authorities crossing Roman-level Class boundaries.

The semantic connections between remote 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-level 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-level Classes. This sample probably illustrates John Lewis Roget's (Peter Roget's son) 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 1) implications, 2) cause and effect, or 3) 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 different vocabulary, and justifies the apparent redundancy of some Categories in different parts of the classification hierarchy.

The alternative to cross-references is for all related senses (Synsets of words) to be repeated, separately under their relevant Categories. But that also has drawbacks --either the categories become so interconnected that they are indistinguishable, or they become so big that their core ideas cannot be discriminated. This reflection of the small-world phenomenon (in modern terminology) 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 upon as indicating in some degree the natural points of connection between the categories" (Roget, J. L, 1879,, p. xi). They solve the essential 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).

8. Conclusion

Cross-references form an elaborate network of links throughout the thesaurus. Latent semantic information can be extracted from the cross-references by 1) classifying them then selecting relationships among the different types of cross-references; 2) by examining the density of cross-references at specific levels of the hierarchy; and 3) by studying the semantics shared by disparate locations in the thesaurus linked by cross-references. The links range from semantically strong, reciprocated, whole-Category cross-references located in different Classes, to weak self-referencing links that reference locations within their own source Categories.

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 of a small-world model (although it has not been tested here mathematically). There is strong coherence within the top, Roman Class, and Letter Class 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 which cross boundaries. Strogatz (2001) points out that a small-world network falls 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 cross 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.

References

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.

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

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

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., & Tengi, R. (1993). Five papers on WordNet. *Technical Report*. Princeton, N.J: Princeton University.

Old, L. John, (2003). *The Semantic Structure of Roget's, A Whole-Language Thesaurus*. (Doctorial dissertation, Indiana University, 2003). Dissertation Abstracts International.

Old, L. John, (2002). Information Cartography Applied to the Semantics of Roget's Thesaurus. Proceedings, *13h Midwest Artificial Intelligence and Cognitive Science Conference* (MAICS'02), Chicago, Illinois.

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. (2005). Linguistic Applications of Formal Concept Analysis. In: Ganter; Stumme; Wille (eds.), *Formal Concept Analysis, Foundations and Applications*. Springer Verlag. LNAI 3626, p. 149-160.

Roget, J. L. (1879). *Thesaurus Of English Words And Phrases Classified And Arranged So As To Facilitate The Expression Of Ideas And Assist In Literary Composition by Peter Mark Roget, M.D., F.R.S.* New York, NY: United States Book Company.

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. (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, 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. (1993). The formal analysis of concepts. *Behavioral Research Methods, Instruments and Computers 25*(2), 314-317.

Sedelow, W.A., Jr., & Sedelow, S.Y. (1979). 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.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*, 327-340.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1).

Strogatz, H. (2001). Exploring complex networks. *Nature*, *410*, 268-276.

Travers, J, & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*(4), 425-443.

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.