EST. 1597
GRESHAM
COLLEGE

After 150 years: The topicality of Roget's Thesaurus

Dr L John Old

16 March 2009

. Introduction

Roget’s Thesaurus, like the bible and the works of Shakespeare, is iconic for native English speakers – it is a cultural artefact. School children are taught how to use it and it is found on educated English writers’ and speakers’ bookshelves. It may be used for solving crossword puzzles, for finding synonyms to avoid repetition in written work, or to find out what a word means by viewing the company it keeps in the Thesaurus. Whatever its use, it is acknowledged to be a rich source of “meaning.”

American professors Sally Yeates Sedelow and Walter A. Sedelow Jr. studied the structure and semantics of Roget’s Thesaurus for more than 30 years, and concluded that Roget’s “might be accurately regarded as the skeleton for English-speaking society’s collective associative memory” (S. Y. Sedelow, 1991, p.108). Insights into this semantic store can have implications for psychology and cognitive science, linguistics, and even anthropology. In research, Roget’s Thesaurus has been used for the automatic classification of text, automatic indexing, natural language processing, word sense disambiguation, semantic classification, computer-based reasoning, content analysis, discourse analysis, automatic translation, and a range of other applications.

Roget’s Thesaurus was also used as the basis for WordNet (Miller, G., Beckwith, Fellbaum, Gross, Miller, K., & Tengi, 1993), the electronic model of the mental lexicon proposed by George Miller, father of Cognitive Science[1]. WordNet and Roget’s differ in their organisation. Roget’s is a topical thesaurus – words are grouped by meaning, and the groups are organised into topics (the Categories, or Headwords). WordNet, like Roget’s, also organises words by meaning, but not by topics. Neither does it organise words alphabetically (by form), as do the alphabetic synonym dictionaries frequently (and erroneously) referred to as “thesaurus” by their publishers.

> The most ambitious feature of WordNet, however, is its attempt to organize lexical information in terms of word meanings, rather than word forms. In that respect, WordNet resembles a thesaurus more than a dictionary, … The problem with an alphabetical thesaurus is redundant entries: if word *Wx* and word *Wy* are synonyms, the pair should be entered twice, once alphabetized under *Wx* and again alphabetized under *Wy*. The problem with a topical thesaurus is that two look-ups are required, first on an alphabetical list and again in the thesaurus proper, thus doubling a user’s search time. (Miller et al., p. 3)

Miller’s final comment is relevant. However, it was Roget’s stated goal, not to produce a synonym[2] dictionary (which could have been organised alphabetically, and, consequently, would have required only a single look-up), but to classify words *according to the ideas they represented*. In fact he wrote in his Introduction “it is hardly possible to find two words having in all respect the same meaning, and being

[1] Author of “The Magical Number Seven, Plus or Minus Two” – see Wikipedia article on George Armitage Miller and *The Making of Cognitive Science, Essays in Honor of George Armitage Miller* Edited by William Hirst, Cambridge University Press, for more.

[2] Roget purposely did not use the term *synonym*. For words within a Category he used the term *Analogous*; and for words in the equivalent, but opposed (antonymic), Category he used *Correlative*.

therefore interchangeable". The "problem" of the double look-up is a consequence of the fact that a Roget's Thesaurus is perhaps less similar to a dictionary and more similar to a library, where, analogously to the way that words are classified within topics, books are classified on library shelves. As in a true thesaurus, library books are not arranged alphabetically and an index look-up (using a Card catalogue or computer search) is required to find the location of a book or topic. Once the user arrives at the location (section of a bookshelf – the second "look-up") they are free to choose the precise book found at that location, or to browse nearby for something that may be even closer to the desired goal.

One of the founders of modern library science, S.R. Ranganathan, observed that the immediate neighbourhood of a book on a library bookshelf contained books on similar or identical topics. Nearby were books of more-generally related topics – both to the left and right, and above and below. Beyond these, at some point, were topics alien to the original topic. He viewed this phenomenon as analogous to a penumbra (a halo-like light effect) around the moon, or a street light on a foggy night.

Those who express an affection for Roget's will have noticed the same effect – when one looks up a word in the Thesaurus one finds the word in close proximity to other words that are uncannily similar. Of course, these are commonly called synonyms, but these are in turn surrounded by, not synonyms, but connotative words and ideas still closely associated in one's mind (see Figure 1). These words and ideas, in turn, drift off into other words and ideas that eventually turn the corner into a different semantic street. This occurs in any direction, just as in Ranganathan's topical penumbra of library books. What's more, it happens to any word or any sense one chooses in the Thesaurus (to varying degrees). This phenomenon somehow gives Roget's Thesaurus its credence – and its magic. It also discriminates it from the alphabetic synonym dictionaries.

The subject of this paper is *the topicality*[3]*, contemporaneousness, contemporaneity, currency, currentness, modernity, nowness, presentness, up-to-datedness (UK), up-to-dateness (US)* of Roget's Thesaurus. It is this author's opinion that, provided the publisher/editor continues to add modern terms, Roget's is always current because it reflects the way our minds work, not just the beliefs and word-associations of a retired Victorian doctor. Unfortunately Roget's Thesaurus is disappearing from book shop shelves, replaced by alphabetic synonym dictionaries.

> ***Adv.*** from top to bottom &c. (completely) [See Completeness]. on high, high up, high -- aloof, aloft -- upward, upwards, up -- **o'er, over, overhead, above** -- over head and ears, above one's head -- skyward, in the clouds, airward, in the air -- upstairs, abovestairs -- tiptoe, on tiptoe -- on stilts -- on the shoulders of

Figure 1 Roget entry illustrating the "semantic context" or penumbra using one sense (of 22 senses) of the word over *(highlighted).*

Alphabetic synonym dictionaries require just one look-up and do not contain a potentially confusing classification system. The *Synopsis of Categories* or hierarchy of concepts was Roget's equivalent of the Dewey Decimal System, or Library of Congress Catalogue for classifying library books. He developed and used this to classify his words according to the ideas they expressed. This may account for the halo (penumbral) effects within his work, but it has never been a friendly avenue for naïve users to find specific senses of particular words. None-the-less the arrangement is necessary in this writer's opinion. Roget's is a system, like an engine. If the parts are separated and placed in alphabetic order neither an engine, nor Roget's Thesaurus, quite work the same. Even some editors of competing volumes recognise this fact. "Other revisers than those in the Roget's family have consistently misinterpreted this volume as a book of synonyms and antonyms and have rearranged it or alphabetised it in the hope of making this [the fact that it is a synonymy] clear. (Webster's Dictionary of Synonyms, Introduction, 1942, xvii)

Some editors of alphabetic editions, perhaps recognising what they have lost, have worked to recapture some of that connotative environment.

[3] *Topical* was Prof. Hüllen's term to describe onomasiological/thematic dictionaries, such as Roget's Thesaurus. For example, he defines "topical, the opposite of alphabetical" (Hüllen, 2004, p. 278). This could imply that he meant *topic-ality*, rather than *topical-ity* as the subject of this talk. The latter was assumed.

> In earlier Merriam-Webster^{TM} publications the pattern of supplementing synonym lists with lists of related and contrasted words, words that were relevant to the group under study yet not quite synonyms or antonyms respectively, was extensively tested. This favorably received feature not only allowed more precise delineation of synonyms and antonyms but provided the user with much additional significant and pertinent assistance. The same plan of supplementing synonyms and antonyms with genuinely germane collateral material has been made a feature of this new thesaurus. (*Webster's New Collegiate Thesaurus*, 1989)

outwit *v.t.*

oval, *adj.*

over, *adv. & prep. –adv*. past, across, by; again; beyond; extra, above, more, remaining, left. *–prep*. On, above. See END, OPPOSITION, REPETITION, SUPERIORITY.

overawe …

Figure 2 Example of an alphabetic synonym dictionary entry for the word over *(all senses) illustrating the loss of semantic context.*

Figure 2 shows an alphabetic synonym dictionary entry for the word *over* (all senses) illustrating the loss of semantic context. The capitalised entries are Roget Headwords (Categories), ordered elsewhere alphabetically, and now devoid of their particular semantic context.

. Roget's biography

Roget's most recent biography (Kendall, 2008, *The Man Who Made Lists*) identifies Roget as an obsessive list maker from an early age, and who used this habit to stave off madness and depression. Wallraff concludes from Kendall's biography:

> We owe a greater debt to mental illness than is commonly recognized. An inmate in an asylum for the criminally insane made important contributions to the Oxford English Dictionary. The eminent lexicographer Samuel Johnson exhibited "odd compulsions, such as pausing to touch every lamppost as he walked down Fleet Street," [Kendall] … Peter Mark Roget, exhibited obsessive-compulsive behavior more than a century before his diagnosis was coined. Evidently, people with mental illness are gravely at risk for compiling language-reference books (Barbara Wallraff, *The Wilson Quarterly*, Spring 2008).

It is true that Dr Johnson probably had Tourette's syndrome, and Roget was dedicated, driven, obsessed … but they were hardly mad. Kendall's most recent, previous work was *Psychological Trauma and the Developing Brain: Neurologically Based Interventions for Troubled Children* (Stein & Kendall, 2004), which appears to have predisposed him to a particular view of Roget – one that undervalues his genius[4], and I hope, is countermanded by the following brief biography.

………………………………………………

Peter Mark Roget was born on January 18, 1779, on Threadneedle Street, London, to Catherine, a Belgian immigrant of Swiss Huguenot extraction, and Jean, a citizen of Geneva who had oversight of the local French Protestant Church. Peter's father died young. When he was fourteen, his mother moved the family to Edinburgh where Roget attended medical school and, at age nineteen, completed his training as a medical doctor. 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, 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

[4] Still, Kendall's biography gives a sympathetic face to Roget, the man, and is well worth reading.

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 (for example the two-volume Bridgewater Treatise on Animal and Vegetable Physiology, 1834), but also some on electricity, galvanism, magnetism and electromagnetism; wrote in English, French (in which he conducted most of his family correspondence (Hüllen, 2004, p.13), German, and Latin; and was a founder member of the Society for the Diffusion of Knowledge.

In 1815, Peter Mark Roget invented the log-log slide rule, which included a scale displaying the logarithm of the logarithm. This allowed the direct calculation of roots and exponents. It was especially useful for fractional powers (Wikipedia) and was the main method of calculation for engineers until the calculator and computer came to predominate. He also developed 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 totalling 300,000 words to the seventh edition of the Encyclopaedia Britannica (on the subjects of ants, bees, apiary, education of the deaf and dumb, kaleidoscope, physiology, phrenology, and on various physicians and scientists (Davidson, G., personal communication); 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); and developed a new laboratory test for arsenic poisoning (Wallraff, 2008).

He was also well connected: (among others) he ate at least one meal with Samuel Johnson, fell out with Charles Babbage, disliked Darwin’s grandfather, at one time worked with Jeremy Bentham, and was the favourite nephew of parliamentarian and reformer, Sir Samuel Romilly; this, in addition to his Royal Society associations.

He did not marry until 1824, when he was 45. His wife, Mary (née Hobson) was 16 years his junior. They had two children, Catherine and John Lewis. Mary died just ten years after their marriage.

Only in 1852, at the age of 73, did he first publish 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 rest of his life (17 years) revising and adding to it. He died in West Malvern, on September 12, 1869, at the age of ninety. His son, John Lewis Roget, took over from him as editor of the thesaurus, and Roget’s grandson, Samuel Romilly Roget, from him in turn.

………………………………………………

. Structure of the Thesaurus

Roget was an admirer of the naturalist Carl Linnaeus, whose division of animals into six classes may have inspired Roget to do the same for his groups of words (totalling one thousand topics, or Categories).

I. Abstract Relations (causation, number, quantity, time etc.)
II. Space (including form and motion)
III. Matter (organic and inorganic, and including the senses)
IV. Intellect (including communication)
V. Volition (different types of actions)
VI. Affections (emotions, and including religion)

The Synopsis and Opposed categories

Below the six Classes Roget subdivided the one thousand topics, not into genus, species, order, or phylum, but sub-classes of Sections, Divisions and differentiating subheadings. Together he called this top level of his hierarchical classification system the *Synopsis of Categories*. The *Synopsis* forms the first part of his book, like a Table of Contents. At the lowest level he arranged any antonymic, or opposed Categories as pairs. For example, *visible* vs. *invisible*; *heat* vs. *cold*; and *attack* vs. *defence*.

The body of the thesaurus is composed of those same Categories, ordered as they appear in the Synopsis. Roget insisted on the opposed Categories being two-to-a-page, side-by-side. He was sometimes thwarted by the printers. Even after his death there has been a struggle. For example, 1962 Roget's editor, Robert Dutch removed the opposed-Category structure from the Synopsis of the British edition, but later, 1982 editor, Sue Lloyd restored it. Lester Berrey, editor of the 1962 American edition, also removed the opposed categories from the Synopsis, while long-time editor (four editions) Robert Chapman recently reorganised the whole Synopsis to remove the "Platonic-Aristotelian" flavour.

Roget was often in conflict with the American publishers who "mutilated" his master plan:

> In the course of last summer, an imperfect edition of this work was published at Boston, in the United States of America, in which the editor, among other mutilations, has altogether omitted the Phrases, which constitute an important part of the original; and has removed from the body of the Work all words and expressions borrowed from a foreign language, throwing them into an Appendix, where, being placed in alphabetical order, they are completely lost to the inquirer who is in search of terms expressive of his ideas, and for whose use the work is specially designed. P.M.R January 18th,1855.

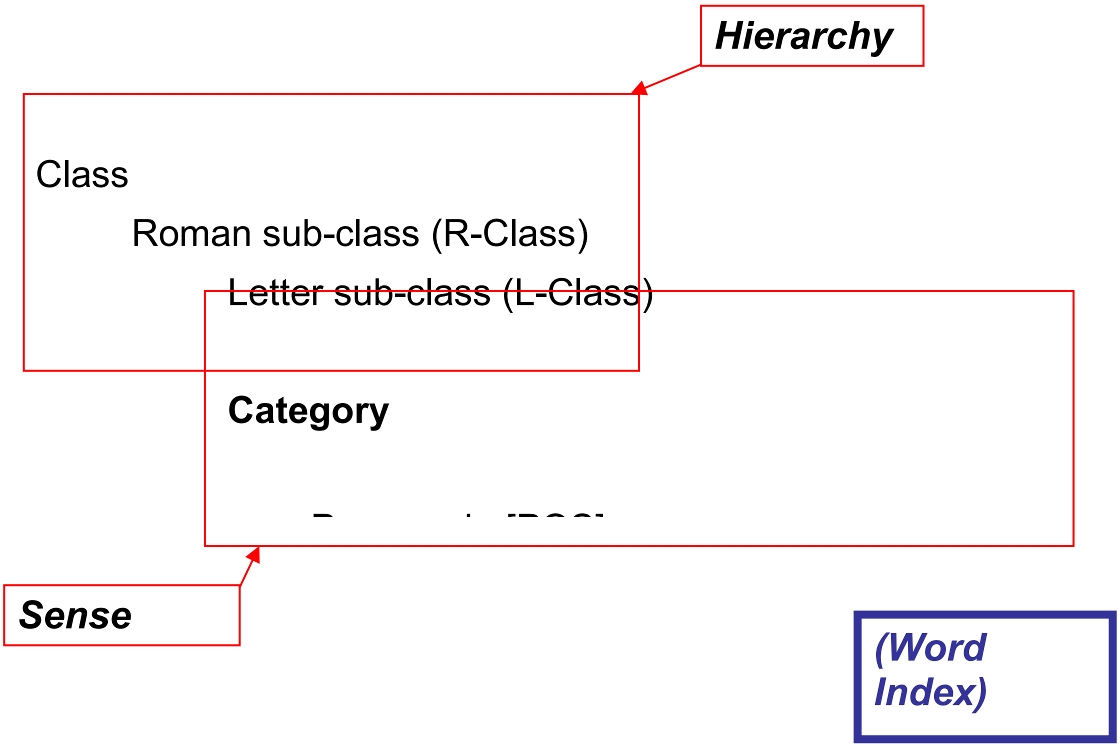

Figure 3 Roget's Conceptual hierarchy showing the relationship between the Synopsis of Categories and the body of the book (Sense Index). The Word Index is an alphabetic listing at the back of the book, which references Categories and parts-of-speech within the Sense Index (based on the American edition).

Eventually the American publishers broke the link with the Rogets and the British edition, but continued to publish their version of the thesaurus.

Whether British or American, all Categories include a Headword which indicates the broad concept; sets of senses arranged by part-of-speech (POS) headings (these are usually, and in this order: Noun, Verb, Adjective, Adverb (optionally, Preposition), Phrases.); subsets of senses organised into paragraphs within the POS groupings, reflecting general concepts; and actual senses i.e., Synsets (Miller et al., 1993) or lists of words describing a shared idea.

. What constitutes a Roget's Thesaurus?

Roget's philosophy is summarised in the quote found on the title page of all of his early editions, from the philologist, John Horne Tooke's book on language, *Epea Pteroenta* (Winged Words)[5]. Tooke believed that we cannot "thoroughly understand the nature of the signs, unless we first properly consider and arrange the things signified" (Tooke, 1786; as cited by Roget, 1852, p. i). There is no doubt that Roget interpreted the "things signified" as ideas or concepts, rather than actual objects of the world.

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. It was, however, 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 occupied him three or four years before he had expanded it from a "small scale" (ibid) to its published form.

His 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 the look-up of a word retrieves 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* (and a harsh critic of Roget's Thesaurus), the first book to achieve this was not Roget's Thesaurus, but John Trusler's *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, as Roget intended.

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*. Werner Hüllen identifies this as the model for the British synonymies that followed. It was Girard's opinion that there are near-synonyms, in the sense of semantically similar lexemes, but that there are no absolutely identical synonyms – at least in the French language. Also, that it is the nature of these "semantically similar" lexemes which make the lexis of a language rich.

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.k.a. Mrs Thrale) a friend of Samuel Johnson. It was another attempt at helping people to choose the right word for the occasion, rather than the right word for the idea. And again this was achieved by *distinctions* or *discriminations*. An example of her method for achieving this is given below, using the Headword *fondness*.

> Amintor and Aspasia are models of true **love**: 'tis now 7 years since their mutual **passion** was sanctified by marriage; and so little is the lady's **affection** diminished that she sat up nine nights by her husband's bedside [while he had a fever]. Nor can anyone allege that her **tenderness** is ill repaid while we see him gaze upon her features…For the rest 'tis my opinion that men love with warmer **passion** than women … and with more transitory **fondness** mingled with that passion.

There were many other "synonymies" 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. Roget was not concerned about usage – that was left up to the reader.

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 help 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 (through the Vikings), and French.

[5] Roget also quotes Tooke several times in the Introduction to the First Edition.

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” and “simple” both come from Latin, via 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. An example of two common synonyms is “over” and “above.” No native English speaker would question whether they are synonyms. Yet “*over* the road” (as in “I live there”) is not the same as “*above* the road.” *Over* and *above* occur together in a Roget’s sense (or Synset) that includes the word *overhead,* but not in any Synset that includes the temporal or spatial senses of *past* and *beyond*.

Note that there do exist “equivalent words” which are, in all contexts, interchangeable. Word equivalents can be defined as *always occurring together in any of their respective Synsets, and never occurring independently*. In other words, they may describe several senses, and always the same senses, so are a type of “perfect synonym.” These may be classified by type (see Old, 1996, for more). For example, *abbreviation* {amidst, amid, midst, ‘mid; F.B.I., Federal Bureau of Investigation}; *spelling variants* {airplane, aeroplane; Führer, Fuehrer; Odin, Woden}; or even “regular” synonyms (words which are not tied by any obvious mechanism, such as form), for example, {absurdly, ridiculously; accountant, bookkeeper}. But, even for *equivalents*, a case can easily be made that context could demand one member before the other.

The case in which Roget’s classificatory system produced Synsets that involved apparently no synonyms at all, is lists – of such things as occupations, parts (such as parts of sailing ships), hosiery, footwear, tackle, weights, armour, and other such objects. 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.

In conclusion, what makes a Roget’s Thesaurus is its classificatory, or topical, structure. Alphabetic organization does not satisfy this criterion. As stated in the introduction, the magic of Roget’s, and usefulness of an engine, is lost if the parts are arranged into alphabetical order. Furthermore, Roget offers his classified words, then expects the user to discriminate their usage – his Thesaurus is not an aid for users to raise their social standing, nor to vary their diction, nor to express more “eloquent” speech; neither is it meant to regulate (or make *apt*) users’ choices.

Publication History of the Thesaurus

In *A History of Roget's Thesaurus: Origins, Development, and Design*, Werner Hüllen (Oxford University Press, 2004) “describes the development in the theory and practice of synonymy from Plato to the seventeenth century, when the first English synonym dictionaries began to appear. Roget's Thesaurus, the first synonym dictionary arranged in topical order, represents an enormously significant peak in this development” (OUP, *product description*). Hüllen gives only Bishop Wilkins as a comparable predecessor to Roget’s Thesaurus. Roget published his new phenomenon with the UK publishers, Longmans who were bought out by Penguin, the current publishers. The British line, runs from 1852-2002. A second, US American line, branched from the British line in 1911, published by Thos. Y. Crowell, and referred to by the publishers as the *International* version (RIT). Crowell merged with Collier, which was bought out by Harper-Row, now Harper-Row Collins, the current American publishers.

Both thesauri contain the hierarchical classification system, the Synopsis of Categories based on Roget’s 1000 categories, with six to eight classes at the highest level. By default, the US thesaurus uses American spellings for common words such as *color* and *jail*, while the British version lists *colour* and *gaol* in the index. It is notable that the current International version does not list *gaol*, even as a Briticism, while the British version does list *jail*. This could reflect the influence of American spelling on British English.

With each new edition of RT, categories may be renamed, added, and sometimes combined or split (very, very rarely deleted). Among the reasons to change the name of a category, modernisation is the most usual. For example, Cicuration (RT, 1852-1953; RIT 1911-1922), meaning the act of taming animals. This was modernised to Animal Culture (RIT, 1946), then Animal Husbandry (RIT, 1962; RT, 1972). A further example is Preterition (RT 1852-1972; RIT 1911-1946), meaning passing, or passed. This was modernised to The Past (RIT, 1962) and Past Time (RT, 1972). Categories may be split where distinct ideas were obviously combined under one head (John L. Roget, 1933, Editors Preface); or expanded, for example, from “Earliness” (RIT, 1922) to “Earliness; Punctuality” (RIT, 1946).

Both versions of RT have in the past borrowed entries from the other, and occasionally, category names: Non-addition; Subduction (RT, 1852-1953; RIT, 1911-1922), Deduction (RIT, 1946); Subduction (RT, 1962), Subtraction (RT, 1982; RIT 1992). Or perhaps these changes simply reflect the changing face of modern English (a term which is always current, no matter in which decade it is used).

The pattern of category changes is different from the British version to the American version. For example the updating of equivalent categories occurs earlier for the American editions. Also, the British version is more tolerant of obsolete (from daily usage) Latin terms. In recent years the addition of new words has mainly been in the areas of science and technology. The addition of these terms also differs between the versions. The British version continues to add new scientific and technical terms to existing Categories, while the American version has added new categories. For example, electronics is found under 160 Power in the British version (RT, 2002), along with strength, force and energy; while the American version has its own category for 342 Electronics (RIT, 1962)

The addition of categories can reflect cultural and political attitudes, just as Roget’s original categories reflected the attitudes, and prejudices, of his day. For example, Category 986 *Pseudo-revelation* (original 1852 edition) contrasts the *heathen* Koran, Buddhism, the Upanishads, and others, with orthodox Judeo-Christian beliefs of the time (under Category 985 “Revelation”).

Military Uses

In the 1950’s and 60’s Margaret Masterman, a pioneer in natural language processing and AI, conducted research using a thesaurus for machine translation at the Cambridge Language Research Unit (CLRU). Her paper, “Potentialities of a Mechanical Thesaurus” is based on experiments with the 1953 of RT, at a time when the paranoia of McCarthyism was rampant in the USA. Not coincidentally, much of their machine translation research was supported by the US military (National Science Foundation, U.S. Air Force Office of Scientific Research, and the Office of Naval Research (Washington, D.C), amongst others (Wilks, 2005; Priss & Old, 2009).

This had no apparent direct effect on RT, but had a massive influence on the ensuing US edition (RIT, 1962). That edition was used directly in the machine translation efforts of the US military to translate Russian military strategy. The following categories were some of those which were either added, or radically expanded: 277 Aeronautics, 280 Rockets and Flying Missiles, 281 Astronautics, 326 Radiation and Radioactivity, 342 Electronics, 345 Radar and Radiolocators, and 348 Automation. A sample of the type of words added is:

> 277 Aeronautics: aircraft hydraulics, jet engineering, kinetics, micrometry, rocket engineering, supersonics, supersonic aerodynamics; aviation medicine, Air Force School of Aviation Medicine, Air Force Department of Space Medicine; aerial navigation, celestial navigation, electronic navigation, automatic electronic navigation, navar, teleran, loran, shoran
>
> 348 Automation: robotic control, cybernetics, automatic electronic navigation, automatic guidance, missile guidance; guided missile, thinking machine, chess-playing machine; ENIAC UNIVAC, IBM 702.

Some of the more cryptic words are acronyms, probably completely unknown to normal English speakers; for example, *loran* (Long Range Aid to Navigation) and *teleran* (Television Air Radar Navigation).

As this was also the beginning of the space race, many related words were added to categories such as 374 Universe (to do with astronomy, star systems, constellations, along with some navigational terminology). 279 Aircraft, in the 1946 (post-war) edition, had extensive lists of Second World War US, British, German, Italian and Japanese military aircraft. In 1962 these were replaced by terms such as “air-sea rescue amphibian, anti-submarine patrol; constant-chord rotor helicopter, intermeshing rotor helicopter; high-altitude reconnaissance plane, long-range patrol bomber, photoreconnaissance plane”.

. Other Roget’s and other Thesauri

Part of the reason why Roget’s Thesaurus seems to be disappearing is the ease of access to all, or part of it through the Internet. There do exist unrelated thesauri, such as the various library, or controlled vocabulary / structured vocabulary / special topic thesauri (for example, the Art & Architecture Thesaurus[6]), but the rest are derived directly from Roget’s original thesaurus.

Evidence of Roget’s Thesaurus is found in word processors, such as Microsoft Word (see Figures 4 and 5). Out-of-copyright thesauri are available for download from such web sites as the Gutenberg Project's free E-text of Roget's. Project Gutenberg’s out-of-print 1911 American edition was submitted as a text file (body/headwords-only) by Micra Corp, and is the basis of most online versions discussed below. It was marked up over the next ten years by this writer, and an index added, and is now available in hyperlinked HTML, MSWord, RTF and other formats from http://www.gutenberg.org/etext/10681.

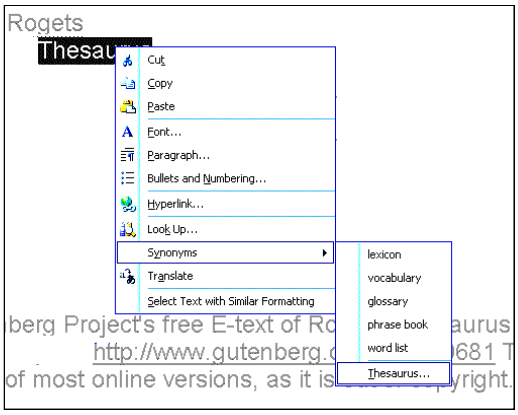

Figure 4 A local search for synonyms from within a MS Word document

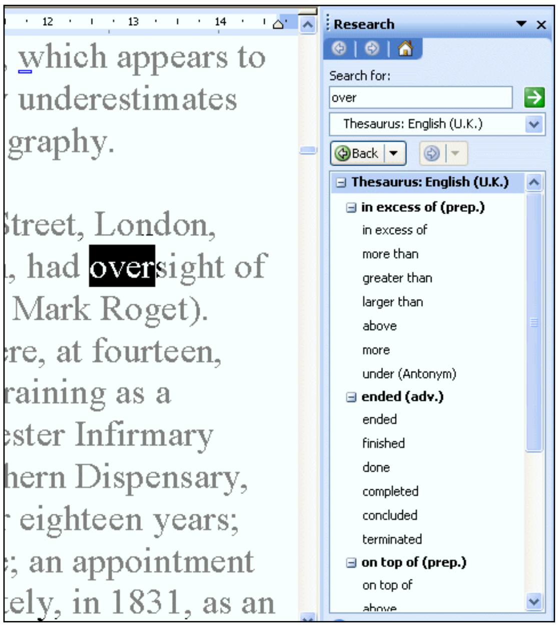

Figure 5 A thesaurus search across the web, direct from a MS Word document.

[6] Available at http://www.getty.edu/research/conducting_research/vocabularies/aat/

Roget’s Thesaurus has been translated into most European languages, but none have been as successful as the English version. The reason may be that English conceptual structures, metaphors, idioms, and other nuances, simply do not translate well into other languages. It has, however, recently spread across Europe in the guise of various “Euro-WordNets” (including rarer languages such as Catalan and Basque) – each with its own unique features, but following Millers WordNet “model of the mental lexicon”.

As mentioned earlier, there are various electronic, or online thesauri available over the Internet. Asadz[7] has a very nice interface to the 1911 edition. Dictionary.com / Thesaurus.com allows searches of *Roget II* (2003) from Houghton Mifflin, an alphabetic thesaurus (though the results are overwhelmed by the advertisements in which they are imbedded). Bartleby provides the entire 1922 American edition (Roget’s International Thesaurus) and *Roget II* (1995) online, but unfortunately like many of the online interfaces (including the interface to Micra Inc’s ARTFL project), their search interface uses a loose pattern matching method, without word boundaries. The result for the search parameter “over” is hundreds of irrelevant hits, because **over** partially matches “p**over**ty”, “g**over**nment”, “**over**t”, “c**over**” and so on.

An interface that invites browsing or exploring via associative links is the hyperlinked Roget’s Thesaurus available at *Roget.org* (without advertisements). It also has a graphical network interface.

“It works like your brain”

Thinkmap Inc’s Visual Thesaurus uses a visual representation of Roget’s Thesaurus that links words based on their synonymy relationships, and connects them via dynamic links to glosses and dictionary definitions.

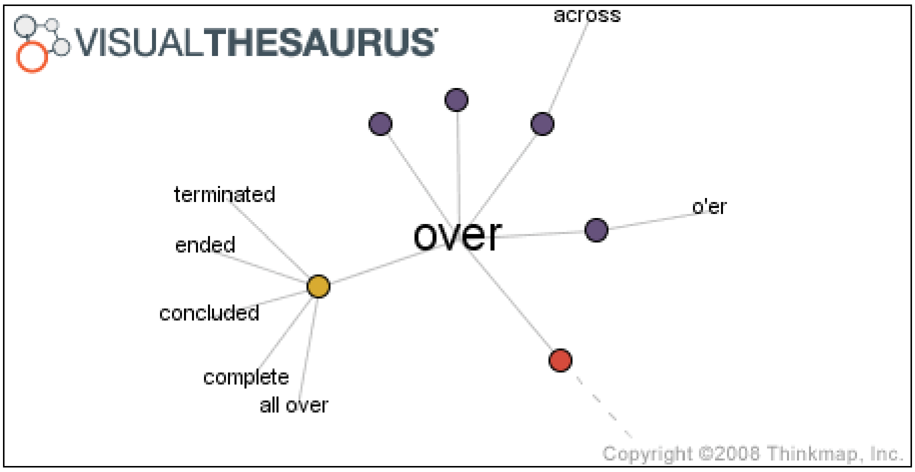

Figure 6 Thinkmap’s Visual Thesaurus

> It works like your brain. Unlike a paper-bound book, the VT [Visual Thesaurus] is fluid and dynamic, like the way you think. Word maps blossom with meaning, helping you find just the right word. You'll write more descriptively and precisely, discover new ideas, and expand your vocabulary. (http://www.visualthesaurus.com/)

Finally, Roget’s classification scheme has been used to create special-topic Roget’s Thesauruses, such as *Roget’s Thesaurus of the Bible*, *Roget’s Thesaurus of Phrases* and *Roget’s Thesaurus of Shakespeare*. For example, *Shakespeare Now* is a project which aims to make all of Shakespeare's plays accessible on the World Wide Web
in language which can be readily understood by English speakers the world over. A group of volunteers led by lexicographer Paul Proctor are paraphrasing and marking up Shakespeare’s plays – including tagging words for classification into Roget’s Thesaurus. So far it contains more than two pages of synonymous

[7] Asadz http://asadz.com/thesaurus/

Shakespearian euphemisms for “have sex” (with).

Comparison among thesaurus sense counts

The Oxford English Dictionary (OED) has about 96 senses for the word “over.” Roget’s original 1852 edition had 4 senses (1 adjectival and 3 adverbial senses) [8].

The 1911 American edition of Roget’s has 9 senses (3 Adj 6 Adv). Roget’s International Thesaurus (1962) has 22 senses (2 Vb 6 Adj 8 Adv 6 Prep). The 150th anniversary British Longman’s (2002) has 17 senses (6 Adj 5 Adv 6 Prep).

Among a sample of the alphabetic “thesauri”, *The New American Roget's College Thesaurus* by Penguin (now called *Roget's College Thesaurus in Dictionary Form*) has 5 senses (4 Adj 1 Prep, plus 4 cross-references). The Oxford Thesaurus has 8 senses (3 Adv 5 Prep) and *Webster’s Dictionary of Synonyms* has 1 sense and one cross-reference.

In general, alphabetic thesauri have fewer senses listed with the Headword, but more cross-references to other Headwords. The difference between the two types of lexicon, it appears, is more qualitative than quantitative.

7. Current research

In *Networks and Knowledge in Roget's Thesaurus* (Oxford University Press, 2008) Werner Hüllen discusses how “Roget's Thesaurus prepared the way for the more recent idea of network semantics. By analyzing retrieval techniques one can show… how the words of languages were (and are) stored in the minds of those who speak them” (OUP, *product description*). Hüllen believed that Roget’s Thesaurus is the first to encompass the semantic *network* of an entire language.

A network is necessary because “A word in total isolation would be non-explainable and, thus, communicatively speaking, empty” (Hüllen, 2004, p.39).

Small Worlds

The Smallworld model can be utilized to account for much of the implicit structure of Roget’s Thesaurus. It 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. The global human social network is only as wide (in terms of the average number of nodes needed to connect any two people) as six 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.

[8] Note that no Roget editions, at least up to 1922, and WordNet, have prepositions. They seem to be subsumed under adverbial senses.

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 uses this principle. 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) proposed that the Roget's network clusters were instead the results of the seeding of concepts in early child development, as children learned speech.

The small-world model suggests the (common-sense, perhaps) 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, John[9] 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 upon as indicating in some degree the natural points of connection between the categories" (ibid, 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

It is agreed by those who have studied Roget's Thesaurus (Miller, Old, Sedelow & Sedelow, Hüllen, and others) that the Thesaurus shares many features with the human mind, and the brain. The arrangement of words offers something other lexicons, so far, do not. The associations that arise from Peter Mark Roget's classification system follow the same patterns as neural networks and other nature-mimicking human constructions.

[9] 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.

Though its popularity has waned in favour of alphabetic synonym dictionaries, it may yet revive if foresighted publishers take the opportunity to do as the OED has done – to convert it into an electronic form that allows users to take advantage of its amazing complexity and intuitive associations.

9. References

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

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

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

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

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*).

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

Hüllen, Werner (2004). *A History of Roget's Thesaurus: Origins, Development, and Design*. Oxford University Press.

Hüllen, Werner (2008). *Networks and Knowledge in Roget's Thesaurus*. Oxford University Press.

Kendall, J. (2008). *The Man Who Made Lists: Love, Death, Madness and the Creation of Roget's Thesaurus*. New York: Penguin.

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

Lloyd, S. M. (Ed.). (1982). *Roget's thesaurus of English words and phrases*. London: Longman.

Masterman, Margaret (1956). Potentialities of a Mechanical Thesaurus. MIT Conference on Mechanical Translation, CLRU Typescript. [Abstract]. In: *Report on research: Cambridge Language Research Unit. Mechanical Translation* 3, 2, p. 36. Full paper in: Masterman (2005).

Masterman, M. (2005). *Language, Cohesion and Form*. Edited by Yorick Wilks. Cambridge University Press, 2005.

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.

Merriam-Webster's Dictionary of Synonyms: A Dictionary of Discriminated Synonyms with Antonyms and Analogous and Contrasted Words. Ed. B. Gove, Merriam-Webster, Inc, 1984.

Miller G.A. (1956). The Magical Number Seven, Plus or Minus Two. *The Psychological Review*, 1956, vol. 63, Issue 2, pp. 81-97.

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

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.

Old, L. John, (1996). Synonymy and Word Equivalence. *Online Proceedings, Midwest Artificial Intelligence and Cognitive Science Society Conference* (MAICS96), Bloomington, IN, April 1996*. http://www.cs.indiana.edu/event/maics96/Proceedings/old.html*

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

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., Old, L. John (2009), *Revisiting the Potentialities of a Mechanical Thesaurus*. Proceedings of the International Conference on Formal Concept Analysis (ICFCA09), Darmstadt, Germany**,** 21-24 May 2009 (in press)**.**

Roget's II: The New Thesaurus, Third Edition by the Editors of the American Heritage® Dictionary. Published by Houghton Mifflin Harcourt Publishing Company, © 2003, 1995.

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. Fellow of the Royal College of Physicians; Member of the Senate of the University of London; of the Literary and Philosophical Societies, &c. of Manchester, Liverpool, Bristol, Quebec, New York, Haarlem, Turin and Stockholm ; Author of the "Bridgewater Treatise on Animal and Vegetable Physiology," &c. Enlarged and Improved, partly from the Author's Notes, and with a full Index by John Lewis Roget*. New York, NY: United States Book Company, Successors to John W. Lovell Company.

Roget, P. M. (1834). *Animal and Vegetable Physiology Considered with Reference to Natural Theology*. Bridgewater Treatise V, 2 vols. London: William Pickering.

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.). (1936/1953). *Thesaurus of English words and phrases, classified and arranged so as to facilitate the expression of ideas and assist in literary composition.* London: Longmans, Green & Co.

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

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.

Shakespeare Now. http://www.shakespearenow.com/

Stein, P.T., Kendall, J. (2004). *Psychological Trauma and the Developing Brain: Neurologically Based Interventions for Troubled Children*. Binghamton, NY: Haworth Press Inc.

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.

Thinkmap Inc. *Visual Thesaurus*. http://www.visualthesaurus.com/

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*).

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 Dictionary of Synonyms. A Dictionary of Discriminated Synonyms with Antonyms and Analogous and Contrasted Words (1942). 1st edition. Menasha, Wisconsin: G & C Merriam.

Webster’s New Collegiate Thesaurus, (1989). Maire Weir Kay, Editor. Springfield, Massachusetts: G & C Merriam.

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

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora. *Proceedings of COLING-92*, 454-460.