Analysis of Polysemy and Homographs of the Word "lead" in Roget's International Thesaurus, 3rd Edition

John Old, Graduate Research Assistant
Department of Computer and Information Science
University of Arkansas at Little Rock

Abstract: This paper follows from previous research on the relatedness or un-relatedness, in Roget's International Thesaurus (RIT), of entries which have identical spellings. A comparison is made between the output from research on a VAX 11/780 computer-accessible version of RIT, the RIT text, and a micro-computer database of the hierarchical and cross-reference information of RIT.

1. Introduction:

The author believes that all meaning is defined by associations and that the context of a word (the words associated with it) define its meaning in that context. Roget's Thesaurus is a culturally validated instantiation of an abstract thesaurus in which words are organized according to their relatedness.

The goal of this analysis was to explore the problem of polysemy (where one word may have several senses) by studying the various entries of the word "lead" found in the 3rd edition of Roget's International Thesaurus (RIT). This analysis was made at several levels of abstraction using the organizational hierarchy of RIT, the cross-reference information, and a combination of both. The analysis focussed on the categories (a middle level of the RIT hierarchy) in which "lead" entries are located.

Data for the analysis was derived from the output of research done by Dr. J. Talburt and D. Mooney, Graduate Assistant; from the body and index of the text of RIT; from the computer-accessible version of RIT; and from a relational database containing the higher levels of the hierarchy and the cross-reference information.

The occurrences of "lead" in RIT were identified from the alphabetic listing output from the Talburt-Mooney research. This data includes the entry ("lead"), the location in RIT of the information (Category: Paragraph: Semicolon-group) and "component" number. The component number is the key to those semicolon groups which are partitioned into groups of entries of like meaning (or senses) according to the type-10 chain definition of Robert Bryan. A type-10 chain is a method of linking groups of entries in one part of a thesaurus with groups of entries in other parts of the thesaurus. Type-10 chains require the two groups to have at least two words in common.

Homographs (semantically unrelated entries) occur in separate components. According to the Bryan Model, two entries in a thesaurus that have the same spelling are homographs if and only if they cannot be the end points of a type-10 chain (5 p.1).

The 34 entries of "lead" in RIT fall into 3 components. Component 0 (zero) contains approximately half the "lead" entries (16), Component 1 contains approximately half the entries

(16) and Component 1256 contains 2 entries. Component 0 entries (approximately 133,000) are identified as being single entries having distinct meanings according to the Talburt/Mooney research. The Component 1 entries are part of one, large (approximately 23,000) amorphous group of entries which have related meanings (or at least, indiscriminable meanings using type-10 chains). Component 1256 is one of a further (approximately) 6000 components which contain 2 or more semantically related RIT entries.

The multi-word phrases containing "lead", such as "lead by the nose", were not included in this analysis in order to keep the data tractable (though obviously there is a great deal of information in these entries which overlaps with the single-word entries).

2. Observations

2.1 Initial Observations

At the highest level of the RIT hierarchy, most of the lead (/follow) entries occurred in Class One, Abstract Relations while all of the lead/metal entries fell in Classes 2,3 and 4 -- Space, Physics and Matter. The several meanings of "lead", the metal, were generally unambiguous. However, I would like to briefly note the meanings of the other "lead", such as "precede" and "guide" before continuing.

The most pervasive sense of the several meanings of "lead" (/follow) which I could derive from introspection and dictionaries such as Webster's New World Dictionary (6), is that having to do with directed process or passively guided movement (in the way that one follows a path rather than the way that a leader actively directs a subordinate). Etymologically "lead" is directly related to "go", "travel", "way", "load", "lode", "laden" and "leave", rather than, for example, the stronger "induce" or "go before". "lead" appears to have generalized from this original sense to the now more familiar "induce", "pull", "direct" etc., types of meanings.

2.2 Categories With More Than One Entry

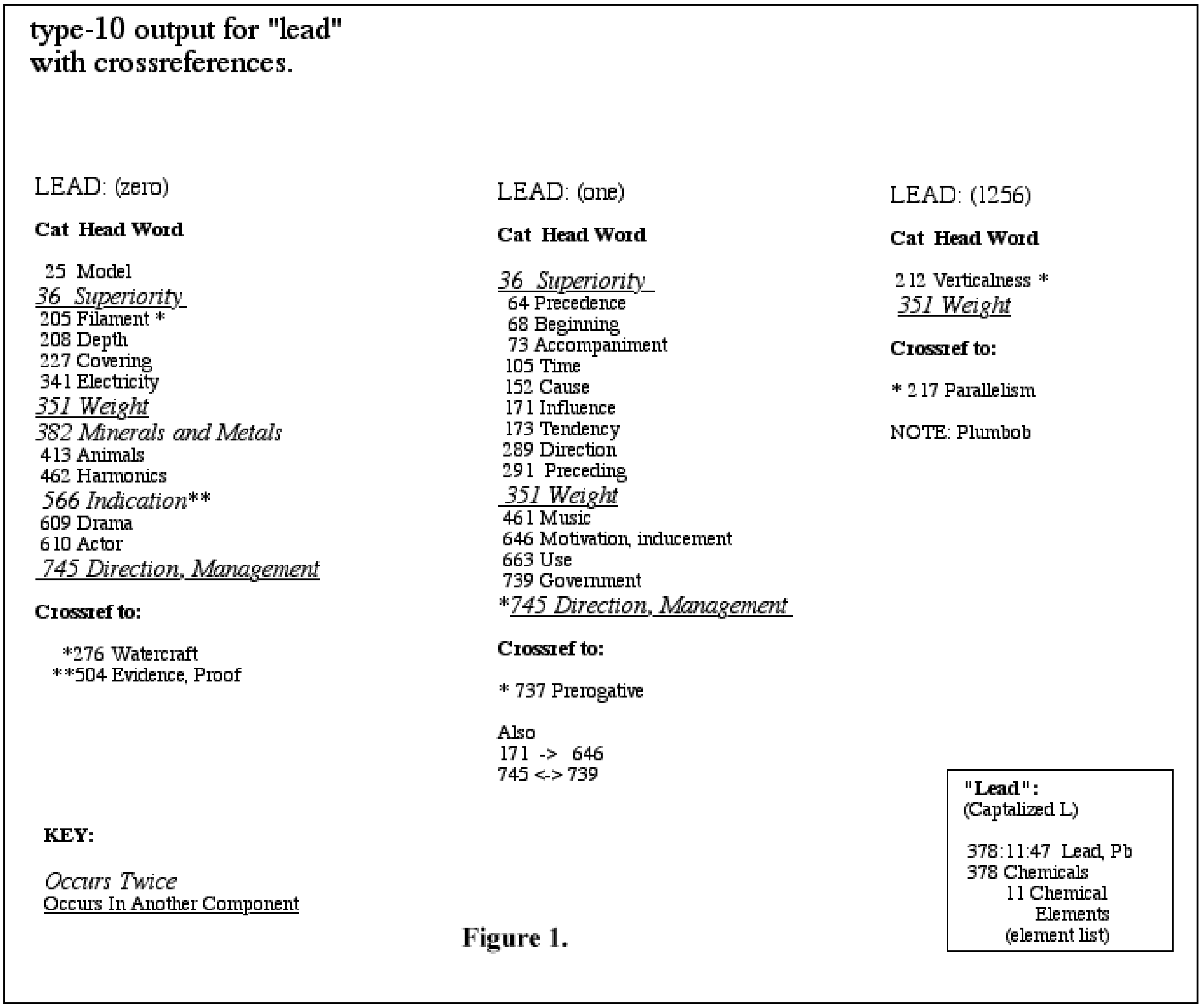
The locations of the entries were sorted by category number and grouped by component number (refer to Figure 1.). There are a total of 28 categories containing at least one occurence of a "lead" entry. Categories 36 (Superiority) and 745 (Direction, Management) appear in both Component 0 and Component 1, and Category 351 (Weight) appears in all three components indicating that there are several "lead" entries within these categories which have discriminable meanings. Categories 382 (Minerals and Metals) and 566 (Indication) appear twice in Component 0, also indicating that they have more than one "lead" entry with discriminable meanings. In other words, Categories, 36, 382, 566 and 745 each have two entries and Category 351 has three entries. All other categories have one entry only. Categories which occur more than once (have more than one "lead" entry) and occur in more than one component, are underlined in Figure 1.

2.3 Cross-references

The cross-references found in paragraphs where "lead" entries are located are included in Figure 1. The actual location (semicolon word group) where the cross-reference is located is shown (asterisked) to the right of the category name. The second number in the cross-reference location identifier (e.g., 351:**12**:3) is the paragraph number, and discriminates among the entries within a category. The foot of each category list shows the location and category name of the referenced location (paragraph word group). There are other cross-references between these categories. However, the additional cross-references are found in paragraphs which do not contain "lead" entries. These will be discussed further below. The cross-references at this level suggest a strong relationship between the meanings of Categories 739 and 745.

The cross-references are only ad hoc approximations to the connectedness of word meanings. Without denser sets of connections, many meanings fall through the cracks. However, they provide a crude abstraction or map of the underlying associations.

For completeness the word "Lead" (capitalized L), which is a/the chemical element and has one-only entry in RIT, is shown in the inset at the bottom right hand corner.



type-10 output for "lead"
with crossreferences.

LEAD: (zero)

Cat Head Word

25 Model
36 Superiority
205 Filament *
208 Depth
227 Covering
341 Electricity
351 Weight
382 Minerals and Metals
413 Animals
462 Harmonics
*566 Indication***
609 Drama
610 Actor
745 Direction, Management

Crossref to:

*276 Watercraft
**504 Evidence, Proof

LEAD: (one)

Cat Head Word

36 Superiority
64 Precedence
68 Beginning
73 Accompaniment
105 Time
152 Cause
171 Influence
173 Tendency
289 Direction
291 Preceding
351 Weight
461 Music
646 Motivation, inducement
663 Use
739 Government
**745 Direction, Management*

Crossref to:

* 737 Prerogative

Also
171 -> 646
745 <-> 739

LEAD: (1256)

Cat Head Word

212 Verticalness *
351 Weight

Crossref to:

* 217 Parallelism

NOTE: Plumbob

"Lead":
(Captalized L)

378:11:47 Lead, Pb
378 Chemicals
11 Chemical
Elements
(element list)

KEY:

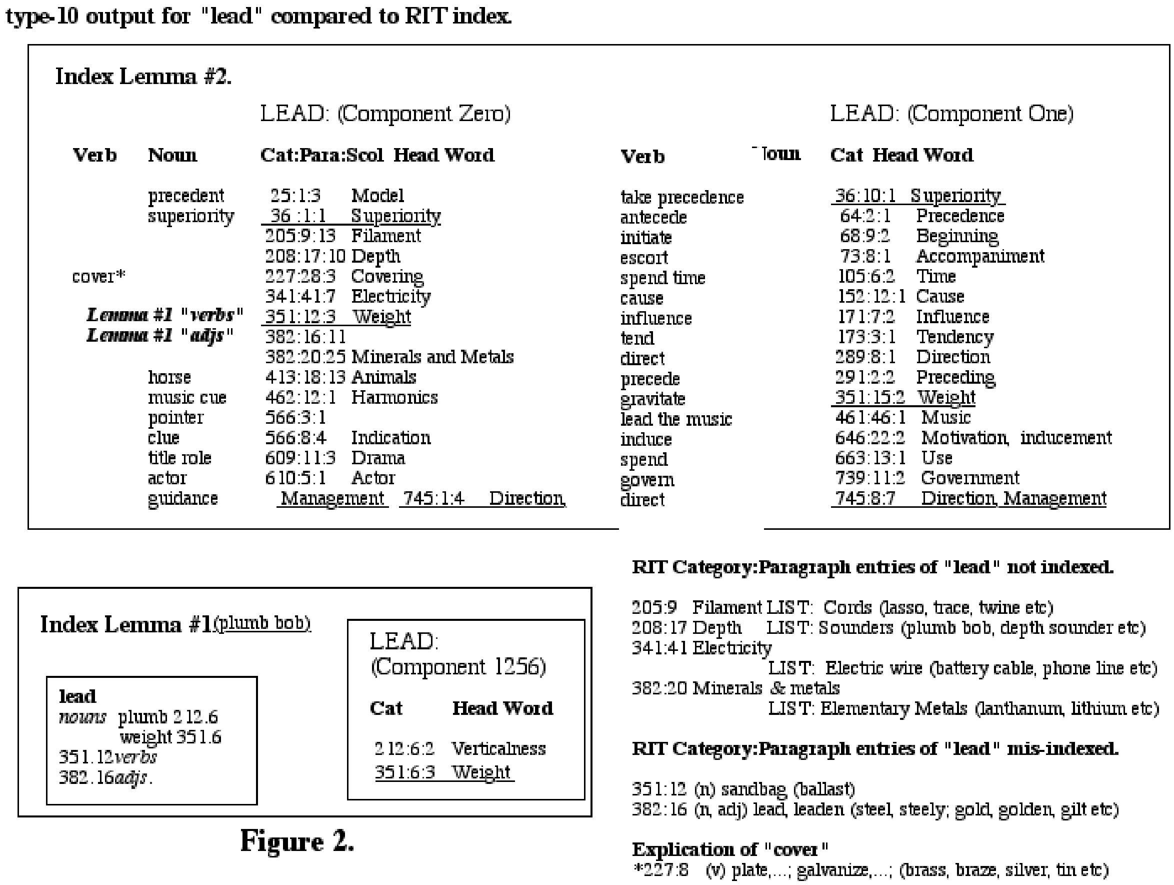
Occurs Twice
Occurs In Another Component

Figure 1.

2.4 Components Vs. Index Entries of RIT

2.4.1 Parts of Speech

By comparing the RIT index entries to the component entries it is obvious (referring to Figure 2.) that the separated meanings (represented by category:paragraph:semicolon group locations) of Component 0 (whether polysemous or homographic) contain entries which are almost all nouns, while the categories of Component 1 are all verbs. All of these index entries fell under the second of two head words or "lemmas" of the index. Component 1256 entries corresponded to the first lemma (lemma #1) and are reproduced in the lower left hand box along with the Component 1256 information.



type-10 output for "lead" compared to RIT index.

Index Lemma #2.

LEAD: (Component Zero)

Verb	**Noun**	**Cat:Para:Scol**	**Head Word**
	precedent	25:1:3	Model
	superiority	36 :1:1	Superiority
		205:9:13	Filament
		208:17:10	Depth
cover*		227:28:3	Covering
		341:41:7	Electricity
Lemma #1 "verbs"		351:12:3	Weight
Lemma #1 "adjs"		382:16:11	
		382:20:25	Minerals and Metals
	horse	413:18:13	Animals
	music cue	462:12:1	Harmonics
	pointer	566:3:1	
	clue	566:8:4	Indication
	title role	609:11:3	Drama
	actor	610:5:1	Actor
	guidance	Management 745:1:4	Direction,

LEAD: (Component One)

Verb	**Noun**	**Cat**	**Head Word**
take precedence		36:10:1	Superiority
antecede		64:2:1	Precedence
initiate		68:9:2	Beginning
escort		73:8:1	Accompaniment
spend time		105:6:2	Time
cause		152:12:1	Cause
influence		171:7:2	Influence
tend		173:3:1	Tendency
direct		289:8:1	Direction
precede		291:2:2	Preceding
gravitate		351:15:2	Weight
lead the music		461:46:1	Music
induce		646:22:2	Motivation, inducement
spend		663:13:1	Use
govern		739:11:2	Government
direct		745:8:7	Direction, Management

Index Lemma #1(plumb bob)

lead
nouns plumb 212.6
weight 351.6
351.12*verbs*
382.16*adjs*.

LEAD:
(Component 1256)

Cat	**Head Word**
212:6:2	Verticalness
351:6:3	Weight

RIT Category:Paragraph entries of "lead" not indexed.

205:9 Filament LIST: Cords (lasso, trace, twine etc)
208:17 Depth LIST: Sounders (plumb bob, depth sounder etc)
341:41 Electricity
LIST: Electric wire (battery cable, phone line etc)
382:20 Minerals & metals
LIST: Elementary Metals (lanthanum, lithium etc)

RIT Category:Paragraph entries of "lead" mis-indexed.

351:12 (n) sandbag (ballast)
382:16 (n, adj) lead, leaden (steel, steely; gold, golden, gilt etc)

Explication of "cover"
*227:8 (v) plate,...; galvanize,...; (brass, braze, silver, tin etc)

Figure 2.

Examination of the entries which were identified by the Talburt/ Mooney/Bryan algorithm but which have no entry in the index (as indicated by the fact that they have no part of speech information beside them in Figure 2.) shows that that they are all participants in lists (of nouns). These lists are sampled in the lower right hand corner.

2.4.2 The Metal Lead

Though the locations/meanings/entries of 351:12 (verbs) [sandbag] and 382:16 (adjs.) [lead, leaden] are identified in the RIT index as having the same meaning as Categories 212.6

and 351.6 (in Component 1256 and lemma #1), they clearly belong elsewhere -- they have different senses from that of lemma #1. This is indicated in part by their placement in Component 0 by the Talburt/Mooney algorithm, and in part by the "intuitive fact" that "plummet", "plumb line", "plumb bob" etc. of 212:6 and 351:6 do not mean the same as ballast and leaden (refer to "mis-indexed" entries in the bottom right hand corner).

An observation in the same theme is that the list of 208:17 (refer to mis-indexed entries and Component 0) <u>should</u> have been included in lemma #1 and Component 1256, and was missed by both the <u>RIT</u> indexers and the Talburt/Mooney/Bryan algorithm. Its entries include "plummet," "plumb line," and "bob." It is an apparent anomaly that the Talburt/Mooney algorithm missed this. The explanation appears to be that word lists are entered (in the machine accessible form of <u>RIT</u>) as separate semicolon entries, so cannot be compared as a group (list) to other, "complete" semicolon groups.

382:20 (the elementary metal, lead) and 378:11 (the chemical element Lead, Pb) were missed completely by the indexers. These are listed as separate Component 0 items in the Talburt/Mooney data.

Lemma #1 should probably be the index entry for the metal and all of the "lead" entries in Categories 208.17, 212.6, 227.28, 351.6, 351.12, 382.16 and 382.20, plus the "Lead" element entry, 378.11 of Figure 1., should be grouped together. Then, for example, the "plumb bob" categories could be listed as one <u>sense</u> of the word.

2.4.3 Categories With Lead Meanings But No "lead" Entries

64.1 "precedence" and 291.1 "precession" were given as noun entries for "lead" in the index, but as they do not contain the word "lead", they were not picked up by the alphabetic output. They will of course appear under precedence or precession. This perhaps reflects one of the few advantages of the manual processing of semantic "features".

2.5 Component 1 Categories

Component 1 contains thousands of semi-colon groups which are to some extent related to each other. The "lead" entries in Component 1 are all related to the index lemma# 2 (i.e., not the metal lead). Analysis (not exhaustive) of the entries associated with each "lead" entry shows clearly why they are grouped together by type-10 chains (bold-face category numbers indicate recurrence):

Category 171:7:2 (Influence) contains (amongst other words) "incline", "sway" and "lead" as does Category 646:22:2 (Motivation, Inducement). Category 351:15:2 (Weight) contains "incline", "lean", "tend" and "lead" as does Category 173:2:1 (Tendency). "incline" and "lead" are sufficient to tie the four semi-colon groups together according to Bryan's model.

Categories 64:2:1 (Precedence): "come first", "go before", 68:9:2 (Beginning): "take the lead", "lead off", 173:3:1, 289:8:1 (Direction): "bear", "point", "aim", **351:15:2**, and 739:11:2 (Government): "command", "preside over", also contain the entry "head". This ties or links this group together and, via 351:15:2, to the previous group.

Other examples are Category 73:8:1 (Accompaniment) contains "escort", "guide", "lead" and "conduct" and 461:46:1 (Music) contains "conduct", "direct" and "lead". 36:10:1 (Superiority) contains "take precedence" and "come first" and **64:2:1** contains "antecede" and "come first".

This method did not pick up the remote relationship between Categories 351 and 212 (plumb bob - Component 1256). This remote relationship may be a reflection of the novel connections which man can make between "distant" functions which results in invention or the realization of "mechanical objects". Referring to the lower right corner of the appended graph: a plumb bob is a tool -- a mechanism -- a combination of the advantages of a dense metal, gravity and vertical tension on a restraining cord (382, 351 and 212). And when applied to a hidden sea floor, all of the above plus the advantage of the strength yet flexibility of a cord or filament (205) for the measurement of the distance of that sea floor (208).

3. Conclusion:

It appears that the selection of part of speech information could do as well as type-10 chains for separating word senses and homographs (refer to the discussion on nouns and verbs). Talburt and Mooney (5, p.7) noted that type-10 chains discriminate beyond homographs, correlating more closely with dictionary sense divisions.

The Talburt/Mooney research output showed a greater completeness in extracting word senses than did the RIT index, though it missed the senses where the word group for that sense did not contain an explicit token (refer to the discussion on Category 64.1 and 291.1) and where the word was part of a list.

The RIT index was incomplete in that it did not pick up several meanings of the word "lead".

There is a strong correlation between the index information and the Talburt/Mooney output. Neither method picked up the strong relationships between "govern" and "direct" (739, 745), or between "antecede" and "precede"(64, 291) found in the cross-references.

The large number of senses of the word lead(/follow) is represented as many associations. Associations are the basic medium of meaning. If a word contributes to the contexts or meanings of many other words it constitutes a large part of the medium. This may imply a measure of "primitiveness", especially if the "underlying sense" of the word -- the common denominator -- cannot be reduced. Perhaps awareness of the directed movement aspect of the "lead" concept, in relation to two points, directions or objects helps identify a primitive axis on which lead/follow, direct/guide, inducement/allurement, superiority/inferiority etc. are discernibly separate locations or attributes. Perhaps lead has assumed its stronger generalized meanings ("guide", "director", "clue" etc.) as an extension to this axis by working as a "tangible" indicator of one end of the axis when the other is a known point, location, object or "self". Other polysemous words may share similar attributes

A final observation is that the entries for lead (metal) as opposed to the entries for lead (precede, govern) were, surprisingly the index entries which provided the most confusion. The former is a mass noun while the latter is a primarily used to describe process (used as a verb). It seems that when the meaning of a mass noun (or perhaps all nouns) is extended (through metaphor, "borrowing" etc.) it tends to create new and different meanings, while extending process meanings "extends" that semantic space (the meaning) without partitioning it.

As pointed out earlier in the discussion, most of the lead/follow entries occurred in Class One, Abstract Relations while all of the lead/metal entries fell in the Space, Physics and Matter Classes (2,3 and 4). Further research might focus on whether the distribution of Talburt/Mooney components bears similar types of relationships.

References:

1.
Bryan, Robert, Abstract Thesauri and Graph Theory Applications to Thesaurus Research. In S. Yeates Sedelow, et al., "Automated Language Analysis, 1972-1973", pp. 45-89.

2.
Roget's International Thesaurus, Third Edition, Pub., Thomas Y. Crowell Company, New York,1962.

3.
Talburt, John R. and Mooney, Donna, Determination of Strongly Connected Components in Abstract Thesauri by the Method of Quartets, Proceedings: ACM Workshop on Applied Computing, March 1989, Stillwater, Oklahoma, pp. 205-209.

4.
Talburt, John R. and Mooney, Donna, The Decomposition of Roget's International Thesaurus into Type-10 Semantically Strong Components, Proceedings: ACM South Central Regional Conference, November, 1989, Tulsa Oklahoma.

5.
Talburt, John R. and Mooney, Donna, An Evaluation of Type-10 Homograph Discrimination at the Semi-Colon Level in Roget's International Thesaurus, University of Arkansas at Little Rock, Department of Computer and Information Science Preprint, 1990.

6.
Webster's New World Dictionary of American English, Third College Edition, Pub. Simon and Schuster,1988.