INTRODUCTION

This paper reports our preliminary investigation into the classificatory nature of Roget’s International Thesaurus (RIT), 3rd ed. (1962) and the potential applicability of the machine-readable version for facilitating human or system language tasks in a variety of settings. We will discuss the nature and structure of RIT; present background information on Roget, his intellectual milieu, and his view of the purpose of the thesaurus; and enumerate some possible applications of the machine-readable version or RIT including applications in information retrieval, natural language processing, machine translation, and other language-related tasks.

BACKGROUND OF ROGET’S INTERNATIONAL THESAURUS

Roget’s lifelong "devotion to organization ... to order ... [and] fascination with problems of classification" (Emblen, 1970, 258) resulted in the first edition of the Thesaurus of English Words and Phrases (1852). As a member of the Royal Society, he had personal and professional acquaintance with the scientists who made remarkable progress in classification systems in geology, biology, physics, and chemistry during the Victorian era. Thorstein Veblen characterizes "pre-Darwinian" science, which Roget and his colleagues helped define, as moved by the "animus of taxonomy; the consistent end of scientific inquiry was definition and classification" (1919, 36). In his Introduction to the First Edition of the Thesaurus, Roget himself notes that (1982, xlii):

The principle by which I have been guided in framing my verbal classification is the same as that which is employed in the various departments of Natural History. Thus the sectional divisions I have formed correspond to the Natural Families in Botany and Zoology, and the filiation of words presents a network analogous to the natural filiation of plants or animals.

Roget’s primary aims in compiling and publishing the Thesaurus were: (1) philosophical, to aid in developing clarity of thought, (2) stylistic, to aid in spoken expression and literary composition, and (3) scholarly, to support the work of other philologists and other linguists. These aims were based on Roget’s belief that language is essential to thought and that words are symbols of ideas. Thus, the arrangement of the Thesaurus is by Roget’s classification of ideas or concepts, not words themselves (1962, xxvii). The reader is presented with a range of related words through which to browse in order to "open to the mind of the reader a whole vista of collateral ideas" (1962, xxviii) and to show the "relation which these symbols [i.e., words] bear to their corresponding ideas" (1962, xlii).

Later scholars have expressed similar understandings of the suggestiveness of related words grouped according to classificatory principles:
The definition [of a word] is an idea, a solid intellectual center; the emotions which have been felt with it rise in memory with it, and give it an aureole, a halo, a nimbus, a glory, spheres of radiance (March and March, 1913, iii).

The meaning of words is not a fixed point, but an area of variable dimensions (Buck, 1949, vi).

Egan (1942) and Emblen (1970) insist that Roget’s original goals have been consistently misinterpreted and ignored and that his work is now seen as nothing other than a compendium of synonyms and antonyms. This use of the *Thesaurus*, enhanced by the inclusion of an alphabetic index by Roget and his successors, supports this misinterpretation. Roget himself, however, made it quite clear that his aims were both higher and wider. It is as an expression of classification that the *Thesaurus* is most engaging historically, and it is as a classification tool, not just a provider of synonyms, that it may have the most potential applicability today.

**RIT AS A CLASSIFICATION SCHEME**

Roget noted that "the object aimed at in the present undertaking is, ... the idea being given, to find the word, or words, by which that idea may be most fitly and aptly expresses" (1962, xxvii). He felt that users could come to his book with an idea or opinion to express and that the arrangement of the *Thesaurus* would lead them to the appropriate word or phrase. The arrangement as envisioned by Roget, and as kept intact through the 1962 edition, has six major divisions: abstract relations, space, matter, intellect, volition and affections. Each of these is further subdivided into at least three and as many as eight subheadings. The subheadings are further divided into "paragraphs," and the paragraphs by parts of speech. It is within this last division that the entries are found.

For example, someone interested in expressing an idea related to communication might search in the Tabular Synopsis of Categories, find "communication" as the second major subdivision of intellect, and would then examine the further subheadings, i.e., nature of ideas communicated, modes of communication, and means of communicating ideas. Each of these is organized into paragraphs with headers, subdivided into parts of speech. Words related to communication would then be found in those subdivisions.

Such use is quite different from using the *Thesaurus* to find synonyms for particular words by searching the index, and then tracking those "synonyms" in the index entries into their paragraphs. As noted earlier, Roget wanted to encourage the first use, but he also added an index to facilitate the second.

We have been fortunate to acquire *Roget’s International Thesaurus, 3rd edition* (1962) in machine-readable form from Sally and Walter Sedelow at the University of Arkansas. Our intentions are to explore issues of the classificatory nature of *RIT* and the potential applicability of the machine-readable version for facilitating a variety of human or system language tasks. These intentions have given rise to a number of questions.

Before addressing issues of applicability, we must determine whether *RIT* is appropriate for certain settings. First, we must investigate whether *RIT*, as it exists today, embodies Roget’s
basic premise or whether, by the process of continuous revision over the years, the terms included
in it today are only related synonymously as opposed to the broader spectrum of relationships
specified by Roget. Second, there is the question of the extent to which the synonyms listed in
RIT reflect the continuum of synonymy proposed by Cruse (1986). This continuum extends from
absolute synonymy to cognitive synonymy to plesionymy to non-synonymy. A third question is
whether the notation scheme which is attached to the classified organization of RIT can be used to
discriminate which of the possible entries for the terms are appropriate for a particular application.

The answers to these questions affect our approach to considering the appropriateness of the
machine-readable version of RIT for enhancing a particular linguistic task. The answers are,
however, of interest in and of themselves to achieve an appropriate theoretic and conceptual
understanding of the nature of the semantic classification of RIT. Once we have gained a
conceptual understanding of Roget’s International Thesaurus in terms of the nature of the
relationships listed, their classification into categories, and their manifestation in machine-
readable format, we will be in a position to investigate a variety of possible applications.

APPLICATIONS OF RIT

Applications we are considering investigating include use of RIT in information retrieval and
natural language processing as well as applications relating to Roget’s original intended use of
the Thesaurus as a tool for facilitating clarity of thought and fluency in written and spoken
expression.

There are a number of situations where the fact that meaning can be and is conveyed with a
variety of phrasings (i.e., synonymous statements) can affect the performance of an information
retrieval system. One possible area of research might be to investigate empirically which of these
situations, and to what extent, the use of knowledge stored in the machine-readable version of
RIT affects information retrieval. Our basic question is whether queries enhanced by some subset
of terms from RIT will improve retrieval effectiveness.

To investigate this question, we would compare and evaluate retrieval performance using
unenhanced queries, queries enhanced with synonymous terms, and queries enhanced with some
subset of the possible relationships represented in RIT. In our investigation of synonymous
relationships, we would use terms selected along the continuum of synonymy as described by
Cruse (1986). We are interested in determining if retrieval performance drops off at some point
along that continuum. We might also look at how retrieval performance changes as we use
queries enhanced with other types of relationships available in RIT. For example, we might
enhance actual queries with sets of selected terms from RIT representing any or all of the
following lexical-semantic relationships: part and whole, collocation, paradigmatic, taxonomic
and synonymic, and antonymic relationships, and determine which groups of added terms
improve retrieval performance. This work would follow closely on that done by Wang,
Vandendorpe, and Evens (1985), who found that enhancement of queries with all types of
relationships, excepting antonymy, improved retrieval performance.

We recognize that many factors influence a user’s perception of retrieval performance,
including purpose of search, type of query, type of database, etc., and we would consider these if we look at Roget-enhanced queries. Included in our investigation would be an examination of methods for automatically determining the appropriate subsets of terms to be included in a search. Since words appear in several categories in RIT, it will be necessary to determine appropriate senses of words to be included in an enhanced query. Prior work by Sparck Jones (1986) indicated that it may be possible to determine these by creating sets of candidate senses for all meaningful words in a query, forming intersections of the sets, and using the senses which appear in all sets.

A number of other information retrieval applications may be explored. Since people often need to search disparate databases which may employ different terminology (e.g., the use of baby and neonate in different databases), RIT may be able to serve as a bridge between vocabularies. Earlier work by Chamis (1985) describes attempts to develop a switching vocabulary in which a vocabulary interface was interposed between the users' search terms and various technical database vocabularies. We believe that RIT might also be useful in providing the basis for an automatic mapping tool from one vocabulary to another. Other researchers (e.g., Mili and Rada, 1988) have considered automatic methods of improving and supplementing existing thesauri by merging them. RIT might be able to provide the additional lexico-semantic information necessary to determine appropriate augmentations to thesauri.

Another possible line of investigation is the use of RIT in natural language processing (NLP) tasks. Most NLP systems are developed on a subset of the full range of language they will eventually need to accommodate. Such an approach is reasonable and practical, but it limits these systems' later performance. For example, the natural language interface to an expert system (or that part of an expert system which must deal with naturally occurring text) contains a lexicon of terms gleaned from analysis of actual samples. This lexicon and therefore the potential coverage of the system could be enriched by automatic enlargement of the lexicon by the addition of all synonyms of a particular level of similarity to the lexicon entry.

An obvious NLP application is machine translation, using RIT's notation scheme. Based on results of the investigation of the theoretical nature of this scheme, more precise translation may be achieved by facilitating inclusion of appropriate synonyms in the interlingua, the universal canonical representation for text into which the source language text is transformed and from which the target language text is constructed.

Another potential research area is interactive use of this machine-readable RIT to improve language-related tasks such as composition. How do people learn more about their ideas and thoughts through the use of a thesaurus? Is it possible to facilitate that process? In order to answer these questions, we would consider the behavior of people as they use a thesaurus, including browsing strategies and methods for choosing words. This study should include an examination of behaviors employed with the manual version of RIT and an attempt to develop a computer-based tool to facilitate browsing behavior.

Research we are pursuing (Kwasnik, Liddy, and Myaeng, 1989) involves the development of an "explorable vocabulary" based on definitions from The Longman Dictionary of Contemporary English (LDOCE) (1987). The proposed representation resembles a semantic network which will be available to a writer or student for browsing and expansion of their current state of knowledge.
The limitation of LDOCE, however, is that its defining vocabulary is just 2000 words since it is intended for learners of English as a second language. This "explorable vocabulary" could be richly enhanced by the addition of synonymous terms, and its usefulness extended. The classification scheme of RIT could be used to determine the synonyms for words in a definition. Since RIT is based on a broad conceptual classification, the notation scheme which reflects this organization offers a means for automatically determining the correct entry which will provide the appropriate synonyms. For example, the RIT index entries for the nouns and verbs in the rather simple definition of "truck" still suggest a large number of senses of each of the defining terms (Figure 1). RIT's classification scheme provides a means by which the appropriate senses for all terms except "quantity" can be selected, namely the categories 271 or 272. The only false entry is railway car for truck. This type of problem and the previously mentioned problem of too many entries suggest the need to refine these techniques.

truck - a large motor vehicle for carrying goods in large quantities

<table>
<thead>
<tr>
<th>truck</th>
<th>quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>commerce 827.1</td>
<td>abundance 34.3</td>
</tr>
<tr>
<td>communication 554.1</td>
<td>amount 28.2</td>
</tr>
<tr>
<td>groceries 831.7</td>
<td>capacity 195.2</td>
</tr>
<tr>
<td>railway car 272.14</td>
<td>indefinite 28.8</td>
</tr>
<tr>
<td>rubbish 669.5</td>
<td>large number 101.3</td>
</tr>
<tr>
<td>types of 272.26</td>
<td>lump 195.10</td>
</tr>
<tr>
<td>vehicle 272.11</td>
<td>measure 490.2</td>
</tr>
<tr>
<td></td>
<td>meter 609.9</td>
</tr>
<tr>
<td>carry</td>
<td>goods</td>
</tr>
<tr>
<td>adopt 637.15</td>
<td>fabric 378.5</td>
</tr>
<tr>
<td>be pregnant 169.12</td>
<td>quantum 28</td>
</tr>
<tr>
<td>deal in 827.15</td>
<td>sum 86.5</td>
</tr>
<tr>
<td>extend 179.7</td>
<td>vowel quantity 594.12</td>
</tr>
<tr>
<td>give credit 839.6</td>
<td>freight 271.7</td>
</tr>
<tr>
<td>induce 648.22</td>
<td>merchandise 831.1</td>
</tr>
<tr>
<td>keep accounts 845.8</td>
<td>property 810.1</td>
</tr>
<tr>
<td>support 216.21</td>
<td></td>
</tr>
<tr>
<td>transport 271.11</td>
<td></td>
</tr>
<tr>
<td>win 726.4</td>
<td></td>
</tr>
<tr>
<td>vehicle</td>
<td></td>
</tr>
<tr>
<td>conveyance 272</td>
<td></td>
</tr>
<tr>
<td>instrument 658.3</td>
<td></td>
</tr>
<tr>
<td>paint 362.8</td>
<td></td>
</tr>
<tr>
<td>photography 577.10</td>
<td></td>
</tr>
<tr>
<td>stage show 611.4</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. RIT Index Entries for Nouns and Verbs in a Definition.

By enriching the definition by the synonyms of the defining terms in their appropriate index entries (271 or 272), a fuller semantic representation could be made available for machine inferencing or human browsing. For example, the entry freight 271.7 for goods adds the synonyms shipment, freightage, consignment, and cargo to the representation.

CONCLUSION

As we have indicated in this brief paper, Roget's International Thesaurus offers rich opportunities for two types of research endeavors, applied and conceptual. We believe that the

TORONTO, NOV. 4, 1990

E.D. LIDDY, C.A. HERT & P. DOTY

ISSN: 2324-9773
machine-readable version of RIT could improve a variety of language tasks, such as composition, natural language processing, machine translation, and query enhancement. Initially, it will be necessary to better understand the basic classificatory structure and notation scheme of RIT. We would then use this understanding to determine how to employ RIT to facilitate language tasks.

BIBLIOGRAPHY


