

Word Association Testing and Thesaurus Construction

Abstract: This paper examines the suitability of word association tests to generate user-derived descriptors, descriptor hierarchies, and categories of inter-term of relationships. Thirty Library and Information Science practitioners were asked to provide as many response words they could for 15 stimulus terms and to describe how the response and stimulus terms are inter-related.

1. Introduction

Word association testing is a technique developed by Carl Jung to explore the complexes in the personal unconscious. Jung came to recognize the existence of groups of thoughts, feelings, memories, and perceptions, organized around a central theme, that he termed psychological complexes. This discovery was related to his research into word association, a technique whereby words presented to patients elicit other word responses that reflect related concepts in the patients' psyche and, in turn, gives clues to their unique psychological make-up (Schultz and Schultz, 2000).

Word association testing has been used extensively in psychology to assess the personality of the test subjects (Galton, 1880; Kent and Rosanoff, 1910; Russell, 1970). Projective techniques, of which word association is a type, typically present respondents with an ambiguous stimulus and ask them to disambiguate this stimulus. The underlying principle behind most projective techniques is that respondents project aspects of their own personalities in the process of disambiguating test stimuli. The interpreter of the projective technique can thus examine answers to these stimuli for insights regarding the respondents' personality dispositions. In a typical word association test, subjects are asked to respond to a stimulus word with the first word that comes to their mind. These associative responses have been explained by the principle of learning by contiguity: "objects once experienced together tend to become associated in the imagination, so that

when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before” (Wettler and Rapp, 1996).

Word association tests present a potentially useful tool in the construction of information retrieval (IR) thesauri, and especially for involving end users in this process. The design of thesauri normally employs a deductive approach: broad categories of terms are selected and are then sub-divided into narrower sets based upon the application of a series of pre-ordained inter-term relationships. In a previous paper (Spiteri, 2002), the author proposed a theoretical framework by which word association tests could be used to generate user-derived descriptors and term hierarchies for IR thesauri. The focus of this present paper is to explore the results of a pilot study, based upon this theoretical framework, which examines the extent to which word association tests can be used to:

- (a) Generate user-derived descriptors, i.e., terms that are most commonly associated with a given concept by the majority of respondents. End-users are provided with a list of domain-specific stimulus terms and are then asked to provide response terms;
- (b) Generate user-derived descriptor hierarchies, i.e., the most commonly-associated attributes, properties, characteristics, parts, etc., of a given concept as identified by the majority of respondents. End-users are asked to not only provide response terms, but to specify how they think these terms are related to the stimulus terms; and
- (c) Generate user-derived categories of inter-term relationships, i.e., the most commonly-associated types of relationships identified by the majority of respondents.

2. Rationale

Word association tests have been used in the construction of a variety of lexical tools such as ontologies, taxonomies, and thesauri to elicit the most typical terms that people associate with a given stimulus term in order to understand how end users categorize vocabulary around a central concept (Spiteri, 2002). The assumption underlying a number of these uses of word association tests is that the response terms function as either synonyms or antonyms; the interpretation of these relationships is made by the researchers, rather than the participants (Deese, 1965; Nielsen, 1997; Miller et

al.1993). Word association tests have been used, to a limited extent, to ask participants to provide attributes and activities associated with the stimulus terms (Battig and Montagu, 1959; Smith and Mark, 1999; Tversky and Hemenway, 1983); once again, however, the researchers categorized how the response terms were related specifically to the stimulus terms, rather than the participants themselves.

IR thesauri contain more than mere listings of antonyms and synonyms, however; they contain also terms that are bound in a variety of hierarchical and associative relationships (e.g., whole-part; an object and the tools used to produce it, etc.). Given this, the assumption that response terms are necessarily synonyms or antonyms of stimulus terms restricts unnecessarily the potential of word association tests. When presented with the word *dogs*, for example, many people respond with the word *cats*. A cat is clearly not a synonym for a dog, neither is it an antonym, yet in the minds of many people, these two terms are closely connected to each other. Rather than assume how people inter-relate these two terms, it may be more useful to ask the participants to explain why they think these two terms are related (e.g., they are both types of domestic animal).

Word association tests often restrict participants to providing only one response term per stimulus term, which could also be overly restrictive. Is *cat*, for example, the only term that people associate commonly with *dogs*? Since IR thesauri act as tools to assist in indexing and searching, it would be useful to use word association tests to elicit as large a set as possible of inter-related terms that reflects the variety of ways in which end users approach a given concept.

IR thesauri rely typically upon the use of symbols such as USE/UF, BT, NT, and RT to demonstrate inter-term relationships. The exact nature of the inter-term relationship expressed by any one of these symbols is not necessarily obvious, however;

for example, is the BT/NT relationship based upon a whole-part, instance, or a genus-species division? ISO and NISO guidelines suggest that the symbols BTG/NTG, BTP/NTP, and BTI/NTI be used to distinguish respectively amongst the genus-species, whole-part, and instance hierarchical relationships but, for the most part, the more generic BT/NT symbols are used (ISO, 1986; NISO, 1993). The equivalence relationship can include synonyms, quasi-synonyms, and even antonyms; the use of USE/UF indicates only that some type of equivalence relationship exists, but not the exact nature of this relationship. The situation becomes ever murkier with the associative relationship, where the generic RT is used to express up to 11 different types of inter-term relationships (ISO, 1986; NISO, 1993).

Word association testing could be thus used also to generate sets of relationship labels (or facet indicators), based upon the terminology participants use to describe how their response terms are related to the respective stimulus terms. Some ontologies, for example, specify the exact nature of inter-term relationships through the use of labels such as “IS A”, “IS A TOOL OF”, “IS A DOMAIN OF”, and so forth (*Theory-Frame Ontology*, 1997; *OpenCyc Selected Vocabulary and Upper Ontology*, 2002). By using end-user generated relationship labels, IR thesauri could follow the model set by such lexical tools to design hierarchies that display more clearly and intuitively the nature of inter-term relationships.

3. *Methodology*

Since most thesauri are domain specific, it is essential that the stimulus terms chosen for the word association test be drawn from the domain at hand. For this pilot project, the subject domain of Library and Information Studies (LIS) was chosen, although this methodology could be applied to a variety of domains, as needed. A test bed of stimulus words for LIS was drawn from the following sources:

- (a) *Open directory project*¹

- (b) *ASIS Thesaurus*²
- (c) *Government of Canada Core Subject Thesaurus*³
- (d) *ERIC Thesaurus*⁴
- (e) *Legislative Indexing Vocabulary*⁵

Stimulus terms were chosen if they were common to at least two-thirds of the sources consulted; in this way, some degree of term familiarity amongst the participants could be anticipated. The total number of stimulus terms chosen was 15. Participants were drawn from the library practitioner population in Atlantic Canada. Calls for participation were communicated via the listservs of the Atlantic Provinces Library Association (APLA) and the Nova Scotia Library Association (NSLA). The total number of participants was 30. For each stimulus term, the participants were given a maximum of two minutes to write down as many response terms that they thought were related to the stimulus term. Participants were asked also to explain in written form how they thought each of their response terms related to the respective stimulus term. The stimulus terms used were:

Authority Control	Information Services
Cataloguing	Librarians
Censorship	Library Science
Digital Libraries	Reference Materials
Information Literacy	Special Collections
Information Retrieval	Technical Services
Intellectual Freedom	Thesauri
Intellectual Property	

3.1 *User-derived response terms*

For each stimulus term, all the response terms provided by each participant were noted. These terms were divided into two categories: (a) terms that occurred uniquely (i.e., that were cited by only one participant); and (b) terms that were cited by two or more participants. It should be noted that the singular and plural forms, and variant spellings of the same response term, were considered to constitute one term (e.g., librarian/librarians, cataloguing/cataloging). The average number of response terms

assigned by the participants per stimulus term was calculated. Stimulus terms were ranked in order of: (a) the total number of unique response terms assigned to them; and (b) the average number of unique response terms assigned to them per participant. A list of the most commonly-occurring stimulus term/response term word pairs was generated. Since one of the foci of word association tests is to examine consensus in the way that participants react to a stimulus term, a response term had to be cited by at least 50% of the participants to make it a candidate for a word pair.

3.2 Inter-term relationships

For each stimulus terms, a list of participant-defined inter-term relationships was derived. The inter-term relationships were divided into two categories: (a) those that occurred uniquely (i.e., were cited by only one participant); and (b) those that were cited by two or more participants. The matching of inter-term relationships was rather more complicated than the matching of response terms, since in the latter case, the possible overlaps in the types of relationships expressed needed to be determined. In other words, if one participant says that Term A is a *type of* Term B, and another participant says that Term A is a *form of* Term B, is this, in fact, the same type of relationship? The relationship labels cited by the participants were thus examined independently by the principal researcher and a research assistant. The two evaluators determined independently which of these labels constituted unique types of relationships, and which constituted overlapping types of relationships, and then compared their results. >From this exercise, a single list of user-derived types of relationships was established, and the frequency with which these types were cited by the participants was noted.

4. Findings

4.1 Incidence of response terms

Figure 1 shows the stimulus terms ranked in order of the total number of unique response terms assigned by the participants, with an average of 70 unique response terms per stimulus term.

<i>Stimulus term</i>	<i>Number of unique response terms</i>
Digital libraries	97
Cataloguing	93
Censorship	88
Librarians	84
Information literacy	77
Information services	74
Authority control	72
Library science	71
Thesauri	66
Intellectual freedom	59
Reference materials	59
Special collections	59
Intellectual property	55
Information retrieval	53
Technical services	48

Figure 1: Stimulus terms ranked in order of total number of unique response terms

Figure 2 shows the stimulus terms ranked in order of the average number of response terms assigned by each participant, with an average of 4.1 response terms per stimulus term.

<i>Stimulus term</i>	<i>Average no. of response terms per participant</i>
Cataloguing	5.3
Reference materials	5.0
Information retrieval	4.6
Censorship	4.5
Information services	4.3
Digital libraries	4.1
Intellectual freedom	4.1
Thesauri	4.0
Information literacy	3.9
Authority control	3.7
Library science	3.7
Intellectual property	3.6
Librarians	3.6
Special collections	3.6
Technical services	3.0

Figure 2: Stimulus terms ranked in order of average number of response terms per participant

Figure 3 shows the stimulus term/response term pairings that were cited by at least 50% of the participants; the complete list of word pairs is found in Appendix 1.

<i>Stimulus Term</i>	<i>Response Term</i>	<i>Frequency</i>
Intellectual property	Censorship	82%
Information services	Reference services	79%
Technical services	Cataloguing	74%
Reference materials	Encyclopedias	71%
Intellectual freedom	Censorship	58%
Technical services	Acquisitions	56%
Reference materials	Dictionaries	51%

Figure 3: Word pairs cited by $\geq 50\%$ of the participants

The large number of response terms (Figure 1), compared to the average number of response terms per participant (Figure 2) suggests that there is not always a high degree of overlap amongst response terms; in fact, each stimulus term contains response terms that are mentioned only once. On the other hand, the fact that, on average, participants cited 4.1 response terms per stimulus term means that restricting responses to only one term can, in fact, place a limit on the full potential of word association tests. The word pair *Reference materials/Encyclopedias* is a case in point: 71% of the participants cited *Encyclopedias* as a response term to *Reference materials*, yet *Encyclopedias* was not always the first term cited by the participants, as is the case with all the word pairs that appear in Figure 3. Figure 3 indicates, also, that a stimulus term may be associated frequently with more than one response term, as is the case with *Reference materials/Encyclopedias*, *Reference materials/Dictionaries*, *Technical services/Cataloguing*, and *Technical services/Acquisitions*.

Another factor to be noted is that in Figure 3, a number of the word pairs do not, in fact, constitute incidences of synonyms or antonyms; in fact, perhaps only *Information services/Reference services* could be considered as synonyms. The only seemingly-

obvious antonyms are *Intellectual freedom/Censorship*, which may serve to support the suggestion that restricting word association tests to the derivation of only synonyms and antonyms is too restrictive and fails to make full use of the potential of these tests.

4.2 Incidence of inter-term relationships

The two evaluators agreed that the following labels constituted the same type of relationship, to which they assigned the label that had been cited the most frequently by the participants:

- Type of/Form of = Type of
- Participant/Member/Advocate = Participant
- Component of/Part of = Part of
- Goal of/Aim of/Purpose = Purpose
- Action/Activity = Activity
- Equivalent term/Synonym = Synonym
- Place/Location = Location

Figure 4 shows the stimulus terms ranked in order of the total number of unique relationships assigned to their response terms by the participants, with an average of 11.7 relationships per stimulus term.

<i>Stimulus term</i>	<i>Total no. of unique types of inter-term relationships</i>
Authority Control	16
Cataloguing	16
Censorship	14
Intellectual Freedom	14
Information Retrieval	14
Digital Libraries	13
Information Literacy	13
Information Services	13
Librarians	13
Library Science	13
Intellectual Property	11
Reference Materials	11
Thesauri	11
Technical Services	08
Special collections	08

Figure 4: Stimulus terms ranked in order of total number of unique inter-term relationships

Figure 5 shows the participant-defined inter-term relationships ranked in order of

frequency, with a total number of 20 unique types of relationships.

<i>Inter-term relationship</i>	<i>Total no. of occurrences</i>
Type	216
Part	209
Synonym	166
Activity	144
Tool	87
RT	75
Attribute	65
Product	59
Participant	58
NT	45
Is	31
BT	30
Antonym	26
Location	22
Purpose	17
Format	11
Skill	05
Use	07
Requirement	03
Source	03

Figure 5: Inter-term relationships ranked in order of frequency of occurrence

Although the incidence of synonymous relationships is high, in keeping with the more traditional uses of word association testing, Figure 5 indicates that the *Type* and *Part* relationships are cited the most frequently by the participants, which suggests that word association tests may not necessarily produce only synonyms and antonyms. The fact that the participants identified a total of 20 types of relationships suggests also that word association tests can serve as a valuable tool in examining the different ways in which users group terms and the types of inter-term relationships that end users most commonly associate with any given concept and its response terms. More importantly, perhaps, is the importance of asking participants to explain how their response terms are related to the stimulus terms. True synonyms, e.g., *elevators/lifts* and antonyms, e.g., *life/death* may be relatively easy to identify by the researcher, but without the

explanations provided by the participants, it would be difficult for the researcher to interpret the inter-term relationships of most of the word pairs found in Figure 3.

The application of this word association test resulted in a total of 192 incidences of equivalence relationship (synonyms and antonyms), 531 incidences of hierarchical relationship (part of, type of, BT, NT, Is), and 556 incidences of associative relationships (all remaining relationships). The total number of inter-term relationships is 1279: 15% equivalent, 42% hierarchical, and 43% associative. As can be seen, the equivalence relationship constitutes the minority of inter-term relationship identified by the participants, which is not in keeping with traditional assumptions about the results of word association tests. The hierarchical and associative relationships constitute almost identical proportions of the participants' relationships. What is clear also is that the participants do distinguish amongst different types of hierarchical relationships, which suggests that they go beyond the simple BT/NT distinctions that one finds in most thesauri.

5. Conclusions

The word association test applied in this study was successful in generating a set of user-derived descriptors. Although the response terms provided by the participants did vary quite significantly at times, areas of consensus did emerge, where at least 50% of the participants provided the same terms in response to a given stimulus term. Participants provided an average of 4.1 response terms per stimulus term, which suggests that the traditional restriction upon one response term per stimulus term can serve to limit the contribution of word association testing to the creation of a collection of descriptors for a thesaurus.

The findings suggest that word association tests could be used to generate user-derived term hierarchies. Results indicate that synonyms/antonyms (i.e., the equivalence relationship, according to ISO and NISO) are not, in fact, the only type of inter-term

relationship reflected in the response terms, which has often been the underlying assumption of previous applications of word association tests. Participants, in fact, provided, in practically equal measure, instances of equivalence, hierarchical, and associative relationships. The importance of asking the participants to explain how their response terms are related to the stimulus tests cannot be overlooked. Without this explanation, any interpretation of the relationship between, say, *Librarians* and *Information professionals* would reflect the mental model of the researcher, rather than that of the participants. This application of word association therefore lends itself to the use of inductive reasoning in the construction of thesauri. Rather than start with a general concept and assume an existing relationship between or among terms associated with that concept, which is the typical procedure used in the construction of many thesauri, word association allows thesaurus designers to study patterns of inter-term relationships that emerge. Word association tests could be used also to test existing term hierarchies: do the end-users (or even the thesaurus designers themselves) inter-relate terms in the same way as these hierarchies?

If word association tests are to be used as aids to thesaurus construction, it would be very useful to examine also the degree of consensus amongst the type of relationship proposed between word pairs. All participants cited *Information professionals*, for example, as a synonym of *Librarians*; *Copyright* was always cited as an antonym of *Intellectual freedom*; and *Dictionaries* as a type of *Reference materials*.

Save for the occasional use of the generic BT, NT, and RT labels, the participants had no difficulty making clear distinctions between how different response terms were related to the same stimulus term; in other words, they did not say that a response term was merely broader or narrower than a stimulus term, or that it was simply related to the stimulus term. It may therefore be helpful if thesauri could show an equal degree of

clarity in the way they display inter-term relationships. The relationship labels suggested by the participants could be used as follows, for example:

Librarians

Synonym(s) Information Professionals

Types Academic Librarians
Public Librarians
Reference Librarians
Special Librarians

Activity Acquisitions
Cataloguing
Collections Development
Information Services

The typical thesaurus display for the hierarchy above would be:

Librarians

UF Information Professionals

NT Academic Librarians
Public Librarians
Reference Librarians
Special Librarians

RT Acquisitions
Cataloguing
Collections Development
Information Services

The labels used would vary amongst the displays, since not all descriptors may have *parts* or *tools*, but then again, not all descriptors have BTs, NTs, or RTs. The thesaurus could be designed to allow users to sort displays according to type of relationship, e.g., all *activities* associated with the term *Librarian*.

Reaching true consensus in the design of thesaurus displays is a near-impossible task, given the potential variety within the population it serves. The admittedly limited application of the word association test in this study has provided a degree of consensus amongst the participants, however, which suggests that there would be merit in

conducting further studies with larger numbers of participants, and with more varied populations. This study has not attempted to measure the potential impact that the social and ethnographic composition of the participants could have on the latter's selection of response terms and of their determination of inter-term relationships. Further studies in this area could provide interesting and valuable insight into the degree to which term hierarchies may be affected by such cultural, educational, and social factors.

References

Battig, William F., and William E. Montague. 1969. Category norms for verbal items in 56 categories: a replication and extension of the Connecticut Category Norms. *Journal of Experimental Psychology Monograph* 80, No. 3, Part 2: 1-46.

Deese, James. 1965. *The structure of associations in language and thought*. Baltimore, MD: The Johns Hopkins Press.

Galton, F. 1880. Psychometric experiments. *Brain* 2: 149-162.

ISO. 1986. *Documentation-guidelines for the establishment and development of monolingual thesauri. ISO 2788:1986*. International Organization for Standardization.

Kent, G. H., and A. J. Rosanoff. 1910. A study of association in insanity. *American Journal of Insanity* 67: 37-96, 317-390.

Miller, George, Richard Beckwith, Christine Fellbaum, Derek Gross, and Katherine Miller. 1993. *Introduction to WordNet: an on-line lexical database*. Available from World Wide Web: (<http://www.cogsci.princeton.edu/~wn/obtain/5papers.pdf>)

Nielsen, Marianne Lykke. 1997. *The word association test in the methodology of thesaurus construction*. In *Proceedings of the 8th ASIS SIG/CR Classification Research Workshop*, 43-58. Washington, DC: American Society for Information Science.

NISO. 1994. *Guidelines for the construction, format, and management of monolingual thesauri. ANSI/NISO Z39.19-1993*. Bethesda, MD: National Information Standards Organization.

OpenCyc selected vocabulary and upper ontology. 2002. Available from World Wide Web: (<http://www.cyc.com/cycdoc/vocab/vocab-toc.html>).

Russell, W.A. 1970. The complete German language norms for responses to 100 words from the Kent-Rosanoff Word Association Test. In *Norms of word association*, eds. L. Postman and G. Keppel, 53-94. New York: Academic Press.

Schultz, D. P., and S. E. Schultz. 2000. *The history of modern psychology*. Seventh edition. Harcourt College Publishers.

Smith, Barry, and David Mark. 1999. Ontology with human subjects testing: an empirical investigation of geographic categories. *American Journal of Economics and Sociology*. 58(2): 245-272.

Spiteri, Louise F. 2002. Word association testing and thesaurus construction: defining inter-term relationships. In *Advancing knowledge: expanding horizons for information science. Proceedings of the 30th Annual Conference of the Canadian Association for Information Science, 30 May-01 June 2002*, eds., Lynne C. Howarth, Christopher Cronin, and Anna Slawek. Toronto, ON: Faculty of Information Studies, University of Toronto

Theory-frameontologia. 1997. Available from World Wide Web: (<http://www.ksl.stanford.edu/people/brauch/demo/frame-ontology>).

Tversky, Barbara, and Kathleen Hemenway. 1983. Categories of environmental scenes. *Cognitive Psychology*. 15(1): 121-149.

Wettler, Manfred, and Reinhard Rapp. 1996. *Computation of word associations based on the co-occurrence of words in large corporations*. Available from World Wide Web: (<http://www.fask.uni-mainz.de/user/rapp/papers/wvlc93/latex2html/wvlc93.html>).

WordNet: a lexical database for the English language. Available from World Wide Web: (<http://www.cogsci.princeton.edu/~wn/>)

¹ <http://www.dmoz.org>

² <http://www.asis.org/Publications/Thesaurus/isframe.htm>

³ http://en.thesaurus.gc.ca/these/thes_e.html

⁴ <http://www.ericfacility.net/extra/pub/theseearch.cfm>

⁵ <http://thomas.loc.gov/liv/livtoc.html>

APPENDIX 1

Incidence of overlap in unique response terms per stimulus term

Stimulus Term	Response Term	Frequency
Authority control	Cataloguing	30%
	Consistency	20%
	Thesauri	20%
	Standardization	20%
Cataloguing	Organizing	50%
	AACR2	40%
	Arranging	20%
	Filing	20%

	MARC	20%
	Metadata	20%
Censorship	Filtering	50%
	Book burning	40%
	Banned books	20%
Digital libraries	Virtual libraries	40%
	Online resources	30%
	Electronic libraries	20%
	Websites	20%
	Digital collections	20%
Information literacy	Library instruction	30%
	Information use	20%
	Search skills	20%
Information retrieval	Search engines	20%
	Computers	20%
	Databases	20%
Information services	Reference services	60%
	Reference	30%
	Libraries	20%
Intellectual freedom	Censorship	60%
	Freedom of thought	30%
	Freedom of speech	20%
	Freedom of expression	20%
Intellectual property	Copyright	80%
	Patents	60%
	Trademarks	40%
Librarians	Information professionals	70%
	Reference librarians	30%
	Information specialists	20%
	Professionals	20%
	Libraries	20%
Library science	Information science	60%
	Library studies	30%
	Information studies	20%
	Library and information studies	20%
	Library school	20%
Reference materials	Encyclopedias	60%
	Dictionaries	50%
	Books	20%
	Atlases	20%

Special collections	Rare books	50%
	Archives	40%
	Music collections	30%
	Limited access	30%
	Manuscript collections	30%
Technical services	Cataloguing	50%
	Acquisitions	40%
	Computer services	30%
Thesauri	Synonyms	40%
	Controlled vocabulary	40%
	Hierarchy	30%
	Authority control	30%