Can You Relate? A Study of User Perceptions of Thesaural Subject Relationships

Rachel Ivy Clarke
University of Washington Information School
Box 352840
Seattle, WA 98195-2840
raclarke@uw.edu

Jin Ha Lee
University of Washington Information School
Box 352840
Seattle, WA 98195-2840
jinhalee@uw.edu

ABSTRACT
This poster describes a pilot study investigating what, if any, associative thesaural relationships are identifiable and distinguishable to information seekers. Using associative relationships from the *Library of Congress Subject Headings* (LCSH), we asked potential library users via Amazon Mechanical Turk to identify six narrower distinctions of established associative relationship types. Preliminary results from the survey indicate that some associative relationship types, such as *near synonym* and *agent/process*, are indeed identifiable, while others, like *position in time and space*, remain problematic.

Keywords
Associative relationships, LCSH, Thesauri

INTRODUCTION
In a progressively networked environment, relationships between entities, properties, and concepts are increasingly important. Controlled vocabularies are one established way of documenting such relationships. Scholars in information science have identified over 120 unique relationship types appearing in thesauri (Association for Library Collections & Techncial Services, 1997). Yet international standards only stipulate the use of three basic types: equivalence relationships, hierarchical relationships, and associative relationships. This supposedly reduces the risk of overloading thesauri with 'valueless' relationships (Aitchison, Gilchrist, & Bawden, 2004) that overwhelm indexers and searchers while offering little to no value for the investment (Soergel, 1974).

However, few studies have been done to challenge this assumption. What, if any, associative thesaural relationships are distinguishable and potentially useful to information seekers? This pilot study looks at associative relationships in the *Library of Congress Subject Headings* (LCSH) to see if potential library users can identify narrower distinctions of established associative relationship types. In addition to empirically confirming or refuting established assumptions about associative relationship types, this study and future work to follow may also help provide insight to vocabulary designers, practitioners, and scholars developing tools in a world increasingly driven by information relationships and links.

ASIST 2012, October 28-31, 2012, Baltimore, MD, USA.

LITERATURE REVIEW
In the field of information science, the definition of associative relationships in use today can be traced back to Soergel (1974):

> "Concept A is related to concept B (has an associative relationship to concept B) if the following holds: an indexer or searcher weighing the use of A should be reminded of the existence of B (and there is no hierarchical relationship between A and B)."

While such an open-ended definition of association certainly allows for contextual applications and indexer judgment, such a definition is potentially too broad to be helpful. Soergel himself describes two kinds of associative relationships: "concepts similar in meaning" and "concepts connected empirically" (p. 107-9). Such descriptions are presumably offered as guidance for vocabulary authors as a means of determining, creating, and labeling related terms (RTs) during thesaurus construction, rather than tools for end users of the thesauri. However, the presence of these descriptions does indicate the existence of relationship types within the overall umbrella of associative relationships.

Inspired by Soregel, many scholars took up the task of identifying, labeling, and creating taxonomies of relationship types. Willets (1975) analyzed 10 contemporary thesauri with the specific aim of examining associative relationships and identifying rules governing their use. She found that despite providing some benefits, associative relationships were poorly defined and understood in the sample thesauri. A matrix based on conceptual categories combined with explicit relations

offered a possible 45 associative relationship types. Her analysis of extant term pairs in the sample thesauri seemed to indicate that the type of relation is more important than the conceptual category, and that the most commonly occurring relations were appurtenance ("parts of" relationships, which may be considered hierarchical); concurrence ("mere mental juxtaposition of two concepts"); dimensional (such as spatial and temporal relations) and a general association. This general association was one of the most used relationships, and Willetts notes that this is a "catch-all" category for any "unspecified" relationships, reflecting a lack of definitions, guidelines, and consistency in their creation and application.

Neelameghan et. al. (1978) described a typology of 39 non-hierarchical relationship types, using a facet analysis which attempted to describe relationships between facets based on frequently occurring facet juxtaposition. Nutter (1989) surveyed 15 resources and identified over 100 types of lexical relationships; however, their sources consisted of dictionaries and other seminal works on semantic relationships rather than actual thesauri. The identified relationships were formed into their own taxonomy of lexical relationship types. No claims to comprehensiveness were made; in fact, the opposite stance was emphasized, as specialty domains are bound to have unique relationships not appearing elsewhere.

Well-known for attempting to tackle these issues, Green (1995a) argues that relationships are possibly more important than conceptual entities, and that no current indexing language harnesses the theoretical power available in expressing relationships. In reviewing attempts to create an exhaustive typography of relationship types, Green (1995b) maintains that there is no limit to the number and variety of relationships that might exist, so it is more beneficial to look at what relationships actually do exist. Therefore, Green & Bean (1995) conducted an empirical study attempting to determine what relationship types actually account for topical relevance. However, the only characteristic they found useful for retrieving relevant documents was that of contextual function rather than any inherent properties of the relationships. While certainly the case for information retrieval, this work does not address further possible value in relationships for browsing, navigation, and understanding a domain.

Attempting to improve subject access via controlled vocabularies, the Association for Library Collections & Technical Services' Subcommittee on Subject Relationships/Reference Structures spent nearly ten years exploring subject access structures, focusing on related term references in LCSH as a major area of interest (Miller, Olson, & Layne, 2005). In its first iteration, the Subcommittee identified 122 unique associative relationship types, although 41 of those were debated as hierarchical (Association for Library Collections & Techncial Services, 1997, Appendix B). Like other previous authors, the Subcommittee deemed it potentially unnecessary to determine, encode, and present such detailed relationships to users. However, the committee also admitted uncertainty regarding such a decision, especially in the context of users with varying needs and skills. Despite this potential, further iterations of the committee shifted to discussion and display of pre-existing subject relationships, and never returned to the investigation of more specific semantic associative relationships.

METHOD

Libraries, as purveyors of vocabulary-based access to materials, are major creators of thesauri. LCSH is designed to describe a wide variety of library materials and cover the broad scope necessary for the Library of Congress as well as American public and academic libraries (Stone, 2000). This scope and coverage, in addition to the previous work done by the Subcommittee, makes LCSH an ideal target for investigation.

In order to determine if potential library users can identify narrower distinctions of established associative relationship types, we first needed to establish the ground truth of relationship types. For this we chose the six extant associative relationships recommended by the Subcommittee to be considered for systematic addition (Association for Library Collections & Techncial Services, 1997, Appendix C).

Relationship Type	Term Pairs
Field of study/ object of study	Neurosciences AND Nervous system Nuclear energy AND Nuclear engineering Soil microbiology AND Soilborne plant diseases Veterinary oncology AND Tumors in animals Military psychiatry AND War neuroses
Field of study/ practitioner	Taxonomists AND Biology—Classification Criminal Profilers AND Criminal behavior, Prediction of Preventative medicine physicians AND Preventative medicine Midwives AND Midwifery Plastic surgeons AND Plastic surgery
Agent/process	Ear AND Hearing Fermentation AND Leavening agents Miracle workers AND Miracles Eye AND Vision Perspiration AND Sweat glands
Causal relationships	Plant diseases AND Crop losses Distress in infants AND Crying in infants Library overdues AND Library fines Bacterial diseases AND Pathogenic bacteria Inventory shortages AND Shoplifting
Position in time and space	American bison AND Buffalo meat Grasses AND Hay Spare parts AND Machine parts Snow AND Meltwater Lava AND Volcanic soils
Frequently interchangeable/ near synonyms	Kindness AND Benevolence Window shades AND Blinds Pillows AND Cushions Engraving AND Etching Pedicabs AND Rickshaws

Table 1. Relationship types & selected example RT pairs

To find example term pairs, we downloaded the LCSH RDF data from the Library of Congress ("Authorities & Vocabularies," n. d.). Using Google Refine, we distilled the data to display only terms with associative relationships (called "related terms" or RTs in LCSH). A cursory analysis showed a high number of language and proper family name headings with RTs, which we eliminated from the sample. From there we sourced pairs that represented the above six relationships. We strove to choose examples that were non-jargon and easily understandable to a layperson without further clarification or scope notes. Overly scientific and medical terms were eliminated. For each term pair, two expert researchers in knowledge organization had to agree on the relationship being displayed. Five examples of each of the 6 relationship types were selected (see Table 1).

Once example term pairs were chosen, we deployed an online survey. The questionnaire included 30 multiple choice questions asking respondents to identify the relationship they perceived between the two concepts in the term pair. The answers included the relationship types listed above, as well as a category for "other," in the event that they saw an alternative relationship. We also collected basic demographic data, data about level of library usage, level of familiarity with LCSH, and participants' opinions about whether they thought this level of granularity in relationships would be useful. We surveyed 100 people via Amazon Mechanical Turk (AMT), an established crowdsourcing platform frequently used to collect survey data. AMT has been successfully used for various tasks that require human intelligence including a number of Natural Language Proceedings tasks (Snow et al., 2008), quality rating of Wikipedia articles (Kittur et al., 2008), music mood and similarity judgment (Lee, 2010; Lee & Hu, 2012), and so on.

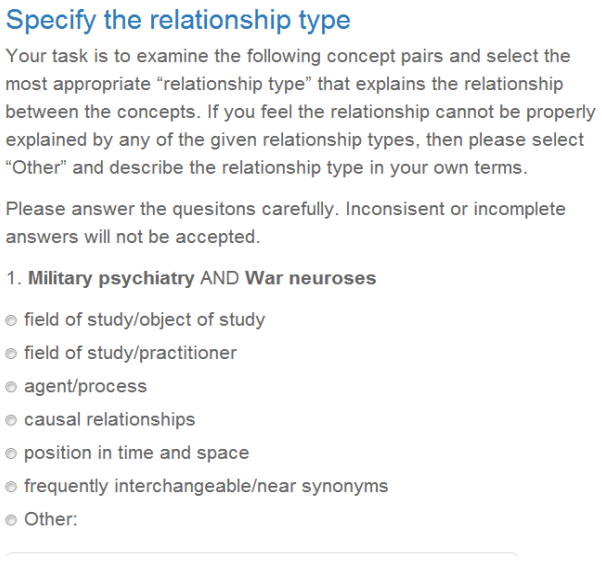

Figure 1. Screenshot of AMT HIT

On AMT, the task requester sets up a "HIT" (Human Intelligence Task, how task is referred to on AMT) and human workers (called "Turkers") recruited by Amazon complete the HIT for a monetary reward. When the HIT is completed and submitted, the task requester reviews the HITs and approves or rejects them. Previous studies using AMT suggest that it is essential to have some sort of filtering mechanism incorporated in the HIT in order to filter out bad responses. In our task, we randomly selected two concept pairs and repeated them in the HIT in order to check the consistency of user responses. We rejected submissions where Turkers responded to the same question with different responses. Of the 134 HITs submitted, 34 were rejected for inconsistent answers. We were able to collect 100 responses in approximately 48 hours. We paid $0.60 for completing each HIT. The average time Turkers spent on the HIT was 11 minutes and 28 seconds.

RESULTS

We examined the level of agreement for each term pair, characterizing agreement of 80% and above as high agreement; agreement of 50%-80% as medium, and agreement of less than 50% as low. We then tallied how many of each relationship type as represented by the term pairs showed high, medium, and low agreement (Table 2).

	>80% agreement	50-80% agreement	<50% agreement
Near synonym	5	0	0
Field of study/ object of study	4	1	0
Causal	0	3	2
Field of study/ practitioner	0	3	2
Agent/process	1	3	1
Position in time and space	0	2	3

Table 2. Tally of term pairs at each level of agreement for the six relationship types.

The *near synonym* relationship type showed the highest level of agreement across the board, with the lowest agreement at 85% to a high of 93%. The *field of study/ object of study* relationship also showed high agreement overall, with only one term pair, "Nuclear energy and Nuclear engineering" seeing 68% agreement. *Causal* and *field of study/practitioner* relationships both showed medium to low agreement overall, and the practitioner relationship was commonly conflated with the *agent/process* relationship. The agreement about *agent/process* itself was widespread. *Position in time and space* showed the lowest agreement.

For the 10 pairs indicating high agreement, the agreed upon response was correctly identified. Of the remaining 20 pairs, there were 7 pairs demonstrating medium and low agreement for which the agreed-upon relationship was not correct: 4 of these were for *position in time and space* (2 medium, 2 low) and 3 for *causal* (1 medium, 2 low).

DISCUSSION

Preliminary results indicate patterns of agreement—some stronger than others—indicating that potential library users can identify certain types of associative relationships. The high agreement and correct identification of the *near synonym* term pairs indicates ease of identification of this relationship type. The high agreement and correct interpretation of 4 out of 5 term pairs representing the *field of study/object of study* relationship also shows promise. The most common conflation for the fifth term pair, "Nuclear energy and Nuclear engineering," was the *object of study/practitioner* relationship (14% agreement). Perhaps the speed of survey completion attributed to misreading "engineer" for "engineering," or perhaps respondents were simply confused about the true definition of "practitioner," since, surprisingly, the medium to low agreement level overall for *object of study/practitioner* was much lower than anticipated. It should be noted that we offered no definitions or explanations of either the terms within the pairs or the descriptions of the relationships types. No definitions were provided by the Subcommittee, although each relationship type did show a sample illustrative term pair. This left the relationship types open to interpretation by the respondents, which may be problematic but is more likely to resemble how these terms would be encountered in a library catalog.

The *agent/process* relationship was unique in range of agreement. 81% of respondents agreed that "Ear and Hearing" represented the *agent/process* relationship, yet only 76% for the analogous "Eye and Vision"; 60% for "Perspiration and Sweat glands"; 51% for "Fermentation and Leavening agents" (despite the word "agent" in the term pair); and 48% for "Miracle workers and Miracles" (commonly confused with *field of study/practitioner*, 40%).

Position in time and space appears to be a highly difficult relationship to identify; agreement rates were low even when they indicated the intended relationship. In fact, *position in time and space* was only selected as a response (correct or not) 80 times in the entire survey. This may indicate a lack of ability to identify such a relationship or a lack of understanding about what the relationship type was intended to mean. This was also the relationship type that the researchers struggled to identify in LCSH and come to agreement on examples.

CONCLUSIONS AND FUTURE WORK

Some associative relationship types are indeed identifiable by potential library users, while others remain problematic. Future data analysis of the qualitative portions of this survey will help shed light on whether the ability to identify these relationships and their explicit inclusion in library catalogs would be beneficial and desirable to potential library users. A planned parallel survey of existing and active library users will offer additional data as well as a comparison between real library users and respondents from Mechanical Turk.

ACKNOWLEDGMENTS

The authors thank David Talley for his generous assistance in sourcing and parsing the LCSH RDF data.

REFERENCES

Aitchison, J., Gilchrist, A., & Bawden, D. (2004). *Thesaurus construction and use : a practical manual*. New York; London: Europa Publications.

Association for Library Collections & Techncial Services, S. o. S. R. R. S. (1997). Final Report. Chicago: ALA.

Authorities & Vocabularies. (2012). from http://id.loc.gov/

Green, R. (1995a). Syntagmatic relationships in index languages: a reassessment. *Library Qtrly, 65*, 365-385.

Green, R. (1995b). Topical Relevance Relationships. I. Why Topic Matching Fails. *Journal of the American Society for Information Science, 46*(9), 646-653.

Green, R., & Bean, C. A. (1995). Topical Relevance Relationships. II. An Exploratory Study and Preliminary Typology. *Journal of the American Society for Information Science, 46*(9), 654-662.

Kittur, A., Chi, E. H., Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. CHI 2008 Proceedings, 453-456.

Lee, J. H. (2010). Crowdsourcing Music Similarity Judgments using Mechanical Turk. Proceedings of the 11th ISMIR, 183-188.

Lee, J. H. & Hu, X. (2012) Generating Ground Truth for Music Mood Classification using Mechanical Turk. Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries.

Miller, D., Olson, T., & Layne, S. S. (2005). Promoting Research and Best Practices in Subject Reference Structures: A Decade of Work by the Subject Analysis Committee. *Library Resources & Technical Services, 49*(3), 154-166.

Neelameghan, A., Maitra, R., & International Federation for Documentation. (1978). *Non-hierarchical associative relationships among concepts : identification and typology*. Bangalore: Documentation Research and Training Centre, Indian Statistical Institute.

Nutter, J. T. (1989). *A lexical relation hierarchy*. [Blacksburg]: Dept. of Computer Science, Virginia Polytechnic Institute and State University.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A.Y. (2008). Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 254-263.

Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance*. Los Angeles: Melville.

Stone, A. T. (2000). The LCSH century: a brief history of the Library of Congress subject headings, and introduction to the centennial essays. *Cataloging & Classification Quarterly, 29*(1/2), 1-15.

Willetts, M. (1975). An Investigation of the Nature of the Relation between Terms in Thesauri. *Journal of Documentation, 31*(3), 158-184.