

HUMAN CONCEPT COGNITION AND SEMANTIC RELATIONS IN THE UNIFIED
MEDICAL LANGUAGE SYSTEM: A COHERENCE ANALYSIS

Shimelis G. Assefa, B.S, M.S

Dissertation Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2007

APPROVED:

Brian C. O'Connor, Major Professor
Victor R. Prybutok, Committee Member
Rada Mihalcea, Committee Member
Samantha K. Hastings, Committee
Member
Herman Totten, Dean of the School of
Library and Information Sciences
Sandra Terrell, Dean of the Robert B.
Toulouse School of Graduate Studies

Assefa, Shimelis G., *Human concept cognition and semantic relations in the unified medical language system: A coherence analysis*. Doctor of Philosophy (Information Science), August 2007, 164 pp., 7 tables, 30 figures, references, 133 titles.

There is almost a universal agreement among scholars in information retrieval (IR) research that knowledge representation needs improvement. As core component of an IR system, improvement of the knowledge representation system has so far involved manipulation of this component based on principles such as vector space, probabilistic approach, inference network, and language modeling, yet the required improvement is still far from fruition. One promising approach that is highly touted to offer a potential solution exists in the cognitive paradigm, where knowledge representation practice should involve, or start from, modeling the human conceptual system.

This study based on two related cognitive theories: the theory-based approach to concept representation and the psychological theory of semantic relations, ventured to explore the connection between the human conceptual model and the knowledge representation model (represented by samples of concepts and relations from the unified medical language system, UMLS). Guided by these cognitive theories and based on related and appropriate data-analytic tools, such as nonmetric multidimensional scaling, hierarchical clustering, and content analysis, this study aimed to conduct an exploratory investigation to answer four related questions.

Divided into two groups, a total of 89 research participants took part in two sets of cognitive tasks. The first group (49 participants) sorted 60 food names into categories

followed by simultaneous description of the derived categories to explain the rationale for category judgment. The second group (40 participants) performed sorting 47 semantic relations (the nonhierarchical associative types) into 5 categories known a priori. Three datasets resulted as a result of the cognitive tasks: food-sorting data, relation-sorting data, and free and unstructured text of category descriptions.

Using the data analytic tools mentioned, data analysis was carried out and important results and findings were obtained that offer plausible explanations to the 4 research questions. Major results include the following: (a) through discriminant analysis category members were predicted consistently in 70% of the time; (b) the categorization bases are largely simplified rules, naïve explanations, and feature-based; (c) individuals theoretical explanation remains valid and stays stable across category members; (d) the human conceptual model can be fairly reconstructed in a low-dimensional space where 93% of the variance in the dimensional space is accounted for by the subjects performance; (e) participants consistently classify 29 of the 47 semantic relations; and, (f) individuals perform better in the functional and spatial dimensions of the semantic relations classification task and perform poorly in the conceptual dimension.

Copyright 2007

by

Shimelis G. Assefa

Dedication

This Dissertation is dedicated to my late mother Abonesh Yimer.

ACKNOWLEDGMENTS

Before all, I honor and praise Almighty God for the many blessings and for the success I have now. There are several people whose love and support has made a difference in my study. First my sincere gratitude goes to Dr. Brian O'Connor, major professor and committee chair for his generosity of spirit, incisive and cogent ideas, and willingness to help always and at any time. A very special thank you also goes to my committee members: Dr. Victor Prybutok for clearly pinpointing the heart of the matter in my study and for the immediate feedback; Dr. Rada Mihalcea for her helpful insights that helped me better shape the dissertation; and Dr. Samantha Hastings, the true champion whose feedback, enthusiasm and encouragement is always a source of inspiration.

I would also like to gratefully acknowledge Professor and Dean Herman Totten for continued financial support. I must acknowledge two wonderful professors in the school with whom I have had the privilege to work and for their unwavering support, Dr. Yvonne Chandler for letting me teach in her class and for having the confidence in me and Dr. Carol Simpson for allowing me work at the hours of my choosing. I am very grateful to the research participants for their time, to Richie Kasprzycki from Megaputer Intelligence for providing me the license to use PolyAnalyst™ 5.0, to Dr. Larry Wood from Websort for his support via email, and to Ereny Gawargy, Gloria Remijio and Chuck Tucker, who helped me organize and recruit more volunteers in the research task.

I am equally grateful to my families, my dad, my sisters, and my brothers, and friends back in Ethiopia whose prayer and encouragement are my staple diet. I also thank my wife, Seblework Gobena, for her love and support. My dear friends Daniel Gelaw and

Dr. Abebe Rorissa also deserve my sincere appreciation for their friendship, for the informal brainstorming, discussion, and feedback to my work.

TABLE OF CONTENTS

| | |
|----------------------------|----|
| ACKNOWLEDGMENTS..... | iv |
| LIST OF TABLES..... | ix |
| LIST OF ILLUSTRATIONS..... | x |

Chapter

| | |
|---|----|
| 1. INTRODUCTION..... | 1 |
| General Background | |
| Statement of the Problem | |
| Research Questions | |
| Purpose of the Study | |
| Significance of The Study | |
| Basic Assumptions | |
| Scope and Limitations of the Study | |
| Definition of Terms/Concepts | |
| Summary | |
| 2 Literature Review..... | 20 |
| Introduction | |
| Concepts, Categories, Classes | |
| Cognitive Theories of Concepts, Categories, And Relations | |
| The Classical Theory | |
| The Prototype View | |
| Exemplar Model | |
| The Two-Tiered Model | |
| Theory-Based Approach | |
| Cognitive Approaches in Information Retrieval | |
| Conceptual Coherence | |
| Knowledge Structure | |
| The Umls Knowledge Sources | |
| Metathesaurus | |
| The UMLS Semantic Network | |
| Semantic Relations | |
| Data-Analytic Methods | |

| | |
|---|----|
| Spatial/Dimensional Models | |
| Clustering | |
| Set-Theoretic Models | |
| Graph-Theoretic Models | |
| Summary | |
| 3. MATERIALS AND METHODS..... | 58 |
| Introduction | |
| Stimulus Materials | |
| Research Participants | |
| Procedure | |
| Data Collection | |
| Unconstrained Sorting | |
| Description of Categories | |
| Classification of Semantic Relations | |
| Data Analysis | |
| Sorting Data | |
| Analysis of Sorting Data | |
| Analysis of Category Descriptions | |
| Analysis of Semantic Relations | |
| Summary | |
| 4. ANALYSIS OF DATA, RESEARCH FINDINGS, AND DISCUSSION..... | 76 |
| Introduction | |
| Operational Definitions | |
| Description of Participants | |
| Analysis of Sorting Data | |
| Standardizing Categories for Analysis | |
| Describing Sortings Data | |
| Multidimensional Representation of Sorting Data | |
| Analysis of Category Description Data | |
| Analysis of Classification of Semantic Relations | |
| Research Findings And Discussion | |
| Categorization Basis and Coherence Criteria (RQ1) | |
| Concept as a Decision Rule (RQ2) | |

The Relationship Between Human Cognitive Map and Knowledge Structure (RQ3)
 The Relationship Between Human Relation Classification and Relation Structure in the Semantic Network of the UMLS (RQ4)

Summary

| | |
|---|-----|
| 5. SUMMARY AND CONCLUSION..... | 119 |
| Introduction | |
| Summary of Findings | |
| Limitations of The Study | |
| Concluding Remarks | |
| Implications of Research Findings | |
| Suggestions for Future Research | |
| APPENDIX A: THE UMLS 135 SEMANTIC TYPES..... | 130 |
| APPENDIX B: THE UMLS 54 SEMANTIC RELATIONS..... | 133 |
| APPENDIX C: THE EXPERIMENTAL 60 FOOD NAMES | 135 |
| APPENDIX D: LETTER OF INVITATION TO THE FIRST GROUP..... | 137 |
| APPENDIX E: LETTER OF INVITATION TO THE SECOND GROUP..... | 140 |
| APPENDIX F: FOOD SORTING TASK | 143 |
| APPENDIX G: SEMANTIC RELATION CLASSIFICATION TASK | 145 |
| APPENDIX H: DENDROGRAM TREE ACCORDING TO PARTICIPANTS’ RELATION CLASSIFICATION | 147 |
| APPENDIX I: DENDROGRAM TREE ACCORDING TO UMLS RELATION HIERARCHY | 149 |
| APPENDIX J: THE ALSICAL PROCEDURE IN SPSS | 151 |
| APPENDIX K: IRB APPROVAL LETTER..... | 153 |
| REFERENCES..... | 155 |

LIST OF TABLES

| Table | Page |
|--|------|
| 1. Kruskal's Goodness of Fit Measures | 72 |
| 2. Summary of Research Participants Profile | 80 |
| 3. Summary of the 21 Category Names Identified | 83 |
| 4. Summary of Individual Sorting Data | 85 |
| 5. Statistical Summary of Text Analysis Report for the 21 Categories | 95 |
| 6. Summary of Descriptive Statistics for the Semantic Relations Classification | 106 |
| 7. Decision Tree Report for each Category | 109 |

LIST OF ILLUSTRATIONS

| Figure | Page |
|---|------|
| 1. Palmer's five conditions of a representational system | 25 |
| 2. Spatial/dimensional model | 52 |
| 3. Partition model | 53 |
| 4. Hierarchical clustering model | 54 |
| 5. Overlapping nonhierarchical model | 54 |
| 6. Graph-theoretic models | 55 |
| 7. Categories of the associative relation types | 61 |
| 8. Text analysis report of category names | 82 |
| 9. Co-occurrence matrix of sorting data with frequency | 87 |
| 10. Decision tree diagram for "Fruit" category | 88 |
| 11. Two dimensional representations of 60 food names | 89 |
| 12. Scatter plot of dissimilarities (sorting data) against distance | 89 |
| 13. Transformed scatter plot of 60 food names | 90 |
| 14. Predicted vs. real graph for "Fruit" category based on liner regression | 91 |
| 15. Dictionary building process | 93 |
| 16. Text analysis report for the "Alcoholic beverage" category..... | 94 |
| 17. Link analysis graph for "Alcoholic beverage" category..... | 96 |
| 18. Link analysis graph for "Candy" category | 97 |
| 19. Link chart for "Alcoholic beverage" & "Nonalcoholic beverage" categories..... | 98 |
| 20. Link chart for "Candy" category with positive and negative correlation..... | 98 |
| 21. Link chart for "Candy" category (only positive links)..... | 99 |

| | |
|--|-----|
| 22. Hierarchical clustering of the physical relation types | 101 |
| 23. Hierarchical clustering of the spatial relation types | 102 |
| 24. Hierarchical clustering of the temporal relation types | 102 |
| 25. Hierarchical clustering of the functional relation types | 103 |
| 26. Hierarchical clustering of the conceptual relation types | 104 |
| 27. Boxplot data summarizing relation types by the 5 classes | 106 |
| 28. Link analysis graph for “Condiment” category..... | 113 |
| 29. Link chart for the “Fish” category..... | 113 |
| 30. Common space points for the 60 food names | 115 |

CHAPTER 1

INTRODUCTION

This study falls under the broader context of knowledge representation systems in information retrieval (IR). Put in other words, this study can be characterized as a cognitive research on concept, categories and relations, or it is a cognitive study on the structure of knowledge representation. The goal of the study is to investigate conceptual coherence in the unified medical language system (UMLS) knowledge structure based on human cognitive performance. This Chapter reviews the general framework under which the study is situated.

General Background

How humans conceptualize the world around them is an age-old question that has been a source of immense interest since Aristotle. The general theory is that concepts, categories, and relations are crucial elements for humans to understand their surroundings. Because the perceived world is so vast, humans cope with the myriad of incoming stimuli by partitioning related objects into categories and forming conceptual representation. Rosch (1978, p.28) describes this notion of partitioning objects into categories as “cognitive economy”. This conceptual representation is believed to have some type of framework or knowledge structure that enables humans to reason, infer, and perform numerous cognitive tasks. These cognitive activities are variously described by different researchers in the field. According to Barsalou (1992, p.153), categorization and conceptualization are the primary cognitive tasks that result from the use of concepts. For Smith and Medin (1981), categorization and

conceptual combination are important cognitive activities which for them are a result of the function of concepts.

The notion of concept, category, and relations play a significant role not only in our day-to-day sense making endeavor, but also as a topic of research interest for several domains such as experimental and cognitive psychology, philosophy (epistemology), linguistics, computer science (machine-learning in artificial intelligence), and information science (information retrieval and knowledge representation). In information retrieval (IR) systems, research has long recognized the need for a better knowledge structure based on the understanding of concepts, categories, and the relation between the two. Likewise, in experimental and cognitive psychology, there is a tremendous effort to characterize the workings of human cognition vis-à-vis concept, category, and relationships between the two. Many (Allen, 1991; Belkin, 1990; De Mey, 1982; Ingwersen, 1999; Quillian, 1968; Robertson, 2000; Wille, 2005; van Rijsbergen, 1983) believe a better understanding of the human cognition process will yield a better model of knowledge structure and representation in IR systems.

This study is an effort to study the nexus between human cognition of concepts and categories and the knowledge structure in the unified medical language system (UMLS). It is an approach largely motivated by the idea that humans' created knowledge structures such as the UMLS, despite the tremendous amount of time, money and research effort expended, do not adequately support the very reason they were created for, i.e., relate queries to documents in the domain of bio-medicine. More succinctly, the underlying structure does not always expose the implicit and

explicit relationship that exists among and between concepts in the domain of interest. Moreover, the relation between concepts in the knowledge structure is governed by strict rules which usually are incongruent with the flexible nature of the human conceptual system. While an exact one-to-one correspondence between human conceptual structure and knowledge structure in systems does not exist, a better approximation of the human conceptual structure by a knowledge structure is desirable. The rationale for this claim is that a powerful representational scheme allows desired reasoning and inference by the system while blocking unwanted inference and reasoning.

It is apparent that the performance of an IR system is largely dependent on the ability of the system to process and understand natural language. This in effect translates to the understanding of the semantic content of documents. In a formal IR system setting, documents are the “phenomena of interest to the field” to borrow the phrase from Thomas Kuhn’s (1970) paradigm of a field. Stated in other words, Wilson’s “bibliographic universe” described as the totality of writings and recorded sayings (Wilson, 1968, p.6), to which I might add images, explains the phenomena of interest to the field of information science in general and information retrieval in particular.

Taking this point one step further, we find interesting discussion that is most relevant to the focus of this dissertation. Karl Popper, a contemporary philosopher, talking about the theory of knowledge makes a distinction between two kinds of knowledge, i.e., objective and subjective knowledge (Popper, 1972, p.73). For Popper, subjective knowledge (which he calls it organismic knowledge because it

consists of the dispositions of organisms) is the world of our conscious experiences. Popper's objective knowledge (knowledge in the objective sense) is defined as the logical content of our theories, conjectures, and guesses. Examples of objective knowledge, according to Popper, include theories published in journals and books and stored in libraries, discussions of such theories, difficulties or problems pointed out in connection with such theories, computer memories, and any information bearing objects.

I believe the logical route from the above discussion is to learn what happens when a human cognitive agent with its world of conscious experience (conceptual representation of the world) comes in contact with the bibliographic universe or the objective knowledge. In this study we consider that the human cognitive agent uses the IR system and one of its key components, the knowledge structure, as a surrogate to navigate through the world of the bibliographic universe. This sums the tenet of this dissertation: to seek human cognitive performance vis-à-vis established knowledge structures in the UMLS so as to use the understanding to better knowledge representation practices in IR systems.

The Unified Medical Language System

The UMLS is dubbed as one of the largest knowledge sources containing millions of bio-medical concepts and relationships between the concepts. In 1986, the national library of medicine (NLM) started a grand project with the aim of developing the UMLS so as to address medical vocabulary problems by "improving the ability of computer programs to *“understand”* (italics mine) biomedical meaning in user inquiries and then using this understanding to retrieve and integrate relevant machine-

readable information" (Houston et al., 2000). The UMLS contains three components, two of which, the metathesaurus (Meta) and the semantic network (SN) are important to this study.

The Meta is the central vocabulary storehouse. The current UMLS release, 2006AC (U.S. National Library of Medicine, 2006) contains more than 1.3 million concepts and 6.4 million unique concept names (terms) from more than 100 distinct vocabulary sources. The Semantic Network (SN) is a tree structure created to unify the millions of concepts in the Meta. The SN can be considered as a high-level categorization having 135 semantic types, which are abstracted from the more than 1.3 millions of concepts. These semantic types are again tied to each other by 54 semantic relations. The current relations in the semantic network are divided into two types namely *isa* and *associative* type relations (U.S. National Library of Medicine, 2006).

Cognitive Theories and Conceptual Coherence

In experimental psychology and cognitive science, we find numerous theoretical views that provide account on the notion of concepts, categories and conceptual coherence of entities that belong to categories. These theories include the classical view (Smith, 1989; Smith & Medin, 1981), prototype/exemplar (Rosch, 1978; Rosch & Mervis, 1975; Smith, 1989; Smith & Medin, 1981), goal-driven (Barsalou, 1992, p. 174), feature-based (Tversky, 1977), theory view (or knowledge-based view) (Mechelen & Michalski, 1993; Medin, 1989; Murphy & Medin, 1985), the frame view (Barsalou, 1992, p.157; Barsalou & Hale, 1993, p.124), and the two tiered concept representation (Michalski, 1993, p.146).

Coherence is explained in terms of what makes members of a category hang together to form a comprehensible class (Murphy & Medin, 1985) or how does the perceptual system is capable of “making sense” out of stimulus variation? (Rodwan, 1964). Many (Arocha, Wang, Patel, 2005; Bodenreider & McCray, 2003; Medin, 1989; Murphy & Medin, 1985; Sanders, 1997) agree on the significance of semantic relations in explaining the coherence and structure of concepts and categories. As a result of investigating several theoretical models for explaining conceptual coherence and structure, researchers are now going beyond explanations that are based on similarity of features to those that are based on the relations, functions, and configurations of features in explaining why certain features are more important than others in determining category membership (Markowitz, 1988; Khoo & Na, 2006).

The idea that concepts are organized around theories (which is sometimes known as knowledge-based categorization) is gaining ground in research on categorization and conceptual structure (Medin, 1989). This new approach is believed to give a better account of conceptual coherence (Murphy & Medin, 1985), where conceptual coherence is described as a mechanism to make meaningful relation between concepts or sets of concepts. We say related ideas cohere together. What allows related concepts to stick together? In this research, we consider the theory-based approaches to categorization as a general framework to undertake an exploratory investigation of how humans’ understand, represent, and use concepts and categories, and their relationship based on a set of concepts and relations taken from the UMLS.

Statement of the Problem

In the field of IR it is generally agreed that systems built on the principles of vector space, probabilistic, inference network, and language modeling, are techniques for matching words from queries with words in documents (Lin & Demner-Fushman, 2006). While empirical validation can be obtained for these methods in TREC evaluations, the fact remains words alone can not capture the semantic content of queries and documents. There is now an increasing recognition that a retrieval system based on concepts and relations is believed to outperform the term-based approaches (Khoo & Na, 2006; Lin & Demner-Fushman, 2006; Wang, 1999). Research has long established the significance of understanding the content of documents as a pretext for a better IR system (Bar-Hillel, 1964). For many researchers in the field of IR, understanding the content of text is a function of understanding natural language by the system. Understanding natural language by the system, in turn, implies linguistic competence, including knowledge of syntax, semantics and pragmatics, and the ability to disambiguate sentences (Koll, 1979).

Natural language processing and understanding is inherently human. For IR systems to model this innate human intelligence, it is plausible first to learn how humans understand text or natural language. This remains true in the IR system as well. Leading scholars in the field of IR research go to the extent of suggesting the importance of understanding human cognitive behavior and the structure of knowledge as a venue for a grand theory of IR (Robertson, 2000). Positioning humans as integral parts of an IR system requires first and foremost incorporating their characteristics in order to enhance IR system performance (O'Connor, 1996).

When IR systems are involved in the understanding of the content of documents the primary challenge will be a representation problem. Everything is a representation of the reality where information needs are represented by queries, and documents are represented by terms, and the matching algorithm in the IR system is in turn a matching of the two representations. Knowledge representation schemes like the UMLS aim to overcome the representation problem by creating a knowledge structure that is based on concepts and semantic relations or by establishing a conceptual connection between users' needs and machine-readable information (Humphreys & Lindberg, 1993). There are several studies that suggest the hierarchical nature of information in human mind and that attempt to recreate those structures in order to improve the efficiency of knowledge based expert systems. Semantic networks or conceptual graphs are believed to have been developed on these grounds (Oroumchian, 1995).

Knowledge representation at the center of this study calls for the need to understand the role of concepts and categories in theories of knowledge representation, and which theory of concepts best account for how people understand the world around them. This, among other things, requires, according to Hampton and Dubois (1993), to know what concepts are and how people understand, represent, and use them. In cognitive psychology research has developed in which concepts and categories are considered central to theories in knowledge representation and long-term memory (Hampton & Dubois, 1993).

It is in view of the above discussion that this research aims to bring a cognitive research on concepts and categories that are the very fabric of knowledge

not only in human conceptual structure but also in knowledge structures like the UMLS. The idea of developing a system that “understands” the content of a document in the UMLS project is a very highly charged mission. This is so because, despite extensive research in natural language processing and understanding, artificial intelligence and machine learning, the handling of complex logical, syntactic or semantic structures by machines especially in retrieval activities is still a far-fetched goal. Humans, however, have the ability to understand meaning, reason, make judgment, infer, and generate new knowledge based on what they know now and based on several other factors. In light of these, this research aims to start with understanding how humans use, organize, and represent concepts so as to better inform knowledge representation practice.

The UMLS as it stands now is a compilation of intractable knowledge sources that is difficult for humans to visualize, audit, and maintain consistency within. Although there are certain attempts to develop automated algorithms that detect inconsistencies among and between concepts in the Meta and the SN, they are not total solutions (Cimino, Min, & Perl, 2003). We would like to ascertain the consistency and conceptual coherence of concepts in the knowledge structure because the ability of retrieval systems to reason from knowledge bases such as the UMLS is largely dependent on the quality of the structure and semantic organization/relationship of concepts. A special issue of the *Journal of Biomedical Informatics* published several research papers on structural issues all of which discuss a wide range of problems and errors regarding inconsistencies in the UMLS (Perl & Geller, 2003).

The UMLS knowledge sources are very complex. They not only aim to unify the separate vocabulary sources by providing a higher-order ontology structure, but they also allow source vocabularies to retain their own structure (Burgun & Bodenreider, 2001). Through this integration process and mapping from the Meta to the SN, various structural issues have remained a source of intensive research (Perl & Geller, 2003). One of the key aspects of this research includes the development both of tools to audit concepts and of better visualization approaches (Bodenreider & McCray, 2003; Cimino, Min, & Perl, 2003; Schulze-Kremer, Smith, & Kumar, 2004). Human intervention in the concept auditing is very arduous and minimal. A recent study conducted to determine inconsistencies between the *isa* relationships in the Meta and the SN identified 17,022 (24.3%) of the *isa* relationships in the Meta could not be explained based on the semantic types of the concepts (Cimino, Min, & Perl, 2003).

One promising direction that offers the potential to overcome the bottlenecks of current IR systems is the semantic IR, where concepts and meaning are the bases for its development. Research in the area of semantic IR is largely focused to create systems that “understand” what a document is about (McCray & Nelson, 1995; Ng, 2000; Raphael, 1968; Schauble, 1987; Yao, 2004). The goal of the UMLS since its inception has been to do exactly the same, i.e., create an IR system that “*understands*” the discourse in bio-medicine (Nelson, Powell, & Humphreys, 2002). It is probably fitting to mention De Mey’s (1982, P.4) thesis about cognitive IR, that posited “any processing of information, whether perceptual or symbolic, is mediated by a system

of categories or concepts which, for the information processing device, are a model of his (its) world".

By way of understanding the human conceptual system, this study aims to investigate the conceptual coherence in the UMLS knowledge structure using theory-based approaches to concept representation as a general framework. The role of theories in cognition was widely acknowledged in that representations of concepts were best thought of as theoretical knowledge or, at least, as embedded in knowledge that embodies a theory about the world, i.e., theoretical knowledge fills many of the gaps in explaining conceptual coherence (Medin, 1989; Murphy & Medin, 1985; Murphy, 1993).

Research Questions

The notion of concept is closely tied to the idea of mental representation and in this regard the discussion of concept itself is situated in the larger discourse of human cognition. In addition, the human conceptual system is regarded as having some kind of conceptual framework or structure that supports several cognitive tasks; including representation, reasoning, categorical inference, conceptual combination, and many more. The various cognitive theories offered to explain concept representation do not sufficiently constrain how humans conceptualize the world around them. Although classical, similarity, feature correlation, exemplar, prototype or probabilistic theories have appealing characteristics, they do not adequately capture the rich, dynamic, and flexible human conceptual system, which is context and background-knowledge dependent. An alternative theoretical view that is believed to account for the flexible, non-modular, and dynamic human conceptual

system is to be found in the theory-based approaches to categorization (Keil, 1989; Medin, 1989; Murphy & Medin, 1985).

Based on the general framework of the theory-based approaches to categorization, this research aims to answer the following research questions:

RQ1: What are the common coherence criteria (categorization bases) humans use in categorization?

RQ2: What is the role of human theory/explanatory principle (intension) in discriminating category members (extension)?

RQ3: To what extent does human concept representation match a sample of concept representations in the unified medical language system (UMLS)?

RQ4: What is the relationship between human relation classification and relation structure in the semantic network (SN) of the unified medical language system (UMLS)?

Purpose of the Study

The primary goal of this study is to investigate the nexus between human concept cognition and concept representation in the UMLS knowledge structure. The field of cognitive psychology provides diverse theoretical bases about the central role played by concepts, categories, and relations as a foundation to basic cognitive activities such as knowledge acquisition, representation, reasoning, inference, etc. By considering appropriate theoretical frameworks that explain the rich and flexible nature of the human conceptual system, this study aims to set a new direction on how better to mesh elements of human concept representation into knowledge

representation and/or knowledge structure in IR systems. Furthermore, this study will provide a step forward in the formalization of the “theory-based” approaches to concept representation.

This research is driven by the problem that current practices in knowledge structure/representation are very rigid and heavily depend on similarity and feature-correlation and do not sufficiently support humans’ understanding, use, and organization of concept. By going beyond theoretical views that dwell on similarity, this research aims to advance a cognitive view that imparts richer common-sense to knowledge representation systems for IR systems to process and understand information in a more constructive and adaptive manner. Using sorting (both constrained and unconstrained) as a data collection method, and together with nonmetric multidimensional scaling (MDS) and content analysis, this study aims to approximate how humans use, organize and represent concept.

In general, this study aims to accomplish the following key objectives:

1. Investigate the theory-based approaches to concept representation.
2. Investigate the relationship between human’s concept representation and knowledge structure in the UMLS for a sample of concepts.
3. Determine the role of semantic relations in conceptual coherence by taking the “associative” class of relations in the UMLS.
4. Determine the proximity relations for selected concepts in the UMLS based on nonmetric multidimensional scaling (MDS).
5. Suggest alternative methods of auditing conceptual consistency and coherence in the UMLS knowledge structure.

6. Determine the common criteria or rules humans use as bases for categorization.
7. Compare the observed proximity relations of concepts and relations with the UMLS knowledge structure.

Significance of the Study

Institutions spend large amounts of money, time, and energy to build knowledge structures with the aim of supporting IR systems (Humphreys & Lindberg, 1993). How much of these knowledge structures consistently support to bridge the gap between the human concept space and the document space is a source of immense research interest. The consistency of the knowledge structure is normally weighed against how well the structure represents concepts in a manner that exposes the implicit and explicit relationship between concepts in the domain of discourse. In natural language processing and understanding, several algorithms have been developed that attempt to understand the content of documents. However, much remains to be done to understand content by IR systems.

In view of the fact that knowledge representation plays a central role in IR, the development of a representation scheme that bears the characteristics of representational adequacy and cognitive validity is a grand objective. This study is, therefore, one that attempts to investigate the efficacy of knowledge structure based on cognitive theories. The study is anchored on the fact that a knowledge representation that is valid in cognitive terms will help to develop a better model for knowledge representation in IR systems. The significance of this study should, therefore, be viewed along these perspectives.

This study is unique in addressing the issue of concepts and categories in knowledge representation schemes based on cognitive theories. It is a new start because it is not based on the widely recognized notion of similarity and feature correlation, but on humans' theory of the domain of knowledge. The result of this study will shed light on how better to bring knowledge structure in accord with human cognition. It further provides a better insight on how to develop and maintain a knowledge structure that is rich to respond to the profoundly flexible, context and background knowledge dependent human conceptual system.

Basic Assumptions

The general assumption in IR that a retrieval system will perform better if and when the underlying knowledge representation scheme allows desired reasoning, while at the same time blocking undesired ones, is the overriding assumption that holds in this study. Other major assumptions include:

1. Concepts are central to human cognition and humans have the conceptual representation upon which they perform cognitive tasks such as categorization, recognition, and inference.
2. In a given domain of knowledge, human conceptual knowledge is represented in their theoretical explanation about the domain.
3. Humans' theory of a domain of knowledge provides coherence to a category structure and concept representation.
4. Humans' conceptual system provides a cognitive structure that can help model a knowledge representation system in IR.

5. Humans' cognitive performance about category structure is stable across individuals.

Scope and Limitations of the Study

This study is a cognitive research on categories and concepts. The concepts are taken from the UMLS, which serves as a testbed for this study. The UMLS is a complex knowledge structure with millions of concepts and relationships between the concepts. Addressing the UMLS knowledge structure in its entirety is beyond the scope of this research. This study limits itself to one semantic type named "food" and considers 60 food names for the cognitive performance task by the research subjects. The result of this study can not be generalized to the UMLS knowledge structure in general. The UMLS is a bio-medical knowledge source and requires domain expertise to conduct the desired cognitive task as specified in this study. The samples of concepts taken from the UMLS for this study, however, are food names that we assume can be understood by any individual.

The selected food names from the semantic category "food" are small in size compared to the total food names that exist in this category. There are about 5000 food names in the semantic type "food" and the selection criteria are based on the theory of indexing (Salton, 1997) where atomic elements are favored to characterize document objects and based on the atomistic theory of concepts (informational atomism) advanced by Fodor (1998). Atomic concepts in this study are regarded as the basic concepts that will not in themselves contain other concepts (Fodor, 1998), for example, cabbage instead of vegetables. The other categories of stimulus materials in this research come from the associative relation types. There are 47

associative relation types in the SN of the UMLS that are organized under five classes.

Definition of Terms/Concepts

Categorization:

The process by which distinct entities are treated as equivalent (Medin & Aguilar, 1999, p. 104).

Cognition:

The processing of information by the brain; specifically, perception, reasoning and memory (Turkington & Harris, 2001, p.54).

Concept:

A related explanation in psychology states a concept as an internal model (that captures the commonalities that exist across a particular collection of stimulus patterns or situations) and as a decision rule (for discriminating members from non-members) (Bower & Clapper, 1989).

A mental representation of a *class* or individual and deals with what is being represented and how that information is typically used during *categorization*,” (Smith, 1989).

Knowledge representation:

Refers to the general topic of how information can be appropriately encoded and utilized in computational models of cognition. It is a broad, rather catholic field with links to logic, computer science, cognitive and perceptual psychology, linguistics, and other parts of cognitive science (Hayes, 1999, p.432).

Knowledge structure:

Concepts and procedures (as elements), and their inter-relationships (Koubek & Mountjoy, 1991).

Semantic relations:

Semantic relations can refer to relations *between concepts in the mind* (called conceptual relations), or relations *between words* (lexical relations) or between text segments. However, concepts and relations are inextricably bound with language and text and it is difficult to analyze the meaning of concepts and relations apart from the language that expresses it (Khoo & Na, 2006).

Thesauri:

Controlled vocabularies that organize concepts for indexing, browsing, and searching. A thesaurus structures concepts by means of a set of standard semantic relationships (NISO, 2005).

Summary

This Chapter presented a broad overview of the issues involved in this study. Most importantly, the theory-based approaches to concept representation is highlighted as a framework that sufficiently accounts for conceptual coherence. The rationale behind this study that knowledge representation systems in IR do not offer a rich structure that is both adequate in representational terms and valid in cognitive terms is clearly outlined. It is also emphasized that understanding how humans use, organize, and represent concepts offers promise on how to improve knowledge representation in IR. Background description is also given about the target domain (the biomedical knowledge structure) on which the investigation was carried out. As a

result of the extensive review of literature to appear next in Chapter two, appropriate research questions were framed and presented. Why this study is important and what significance it will provide if conducted are also discussed in this Chapter. Basic assumptions and definitions of basic concepts are given. The review of literature also offers supporting extant work that highlights the need for a cognitive research on concepts, categories, and relations together with data-analytic tools that aim to formalize such cognitive theories.

CHAPTER 2

LITERATURE REVIEW

Introduction

The literature review is focused to understand human concept cognition and concept relations in the unified medical language system (UMLS) knowledge structure. Noting the goal of this research, i.e., to explicate the nexus between human concept cognition and semantic relations in the UMLS knowledge structure, the review of the literature will place greater emphasis in cognitive psychological theories employed to study concepts and categories. Through the prism of the theoretical views in cognitive and experimental psychology, the literature review aims to provide the context into which this study can be situated. In addition to the theoretical models in cognitive and experimental psychology that account for the study of concepts and categories, the literature review aims to present a wide range of data analytic methods that furnish an adequate formalization of concept cognition and/or representation. A review of the UMLS knowledge sources and knowledge structure together with basic topics such as concepts, categories, and semantic relations will also be reviewed.

Concepts, Categories, Classes

A wide range of literature in cognitive psychology, linguistics, philosophy, computer science, and information science talk about the notions of concepts, categories, and classes and their relation with one another. These notions receive different treatment from the disciplines mentioned and it is of paramount importance

to establish the context firmly from the beginning. In cognitive psychology (Bower & Clapper, 1989; Ingwersen, 1999; Mechelen & Michalski, 1993; Murphy, 1993; Rosch, 1975; Rosch & Lloyd, 1978) concepts are explained in view of a cognitive structure or a cognitive representation of events, entities, and objects. In linguistics (Blair, 1990; Cruse, 2004; Pris, 2005), concepts are treated in their role as word meanings. In philosophy, especially in epistemology (Wille, 2005, 1992; Yao, 2004), concepts are considered as a basic unit of thought. In computer science, most particularly in artificial intelligence (Allen & Frisch, 1982; Griffith, 1982; Sowa, 1984, p.22; Wille, 1992) concepts are explained in connection with knowledge representation. In information science, there is a host of extant literature that largely falls under the “cognitive paradigm” that talks about concepts and categories in ontology, classification, and vocabulary research (Allen, 1991; Belkin, 1990; De Mey, 1982; Ellis, 1992; Ingwersen, 1999, 1993).

The philosophical understanding of a concept constitutes two parts, intension and extension. The extension is the set of all objects and /or entities which belong to the concept, and the intension includes all attributes (properties, meanings) which apply to all objects of the extension (Bower & Clapper, 1989; Wille, 2005). A related explanation in psychology states a concept as an internal model (that captures the commonalities that exist across a particular collection of stimulus patterns or situations) and as a decision rule (for discriminating members from non-members). While the set of entities to which the model applies is the category (intension) the decision rule is its extension (Bower & Clapper, 1989).

In experimental psychology concepts are defined from different theoretical views such as the exemplar theory (Medin & Schaffer, 1978; Medin & Smith, 1984; Smith & Medin, 1981); prototype theory (Rosch & Mervis, 1975) and the classical and probabilistic theory (Smith, 1989; Smith & Medin, 1981). The discussion of concept holds a great deal of emphasis in cognitive science because concepts are regarded as vital to the efficient functioning of human cognition (Cruse, 2004). It is argued that because the world around us is so vast, concepts allow humans to categorize individual entities into classes so the amount of information humans perceive, learn and reason about will be much less. We see the role of concepts allowing us to slice (categorize) the perceived world into classes. This function of concepts, i.e., providing maximum information with the least cognitive effort, also known as “cognitive economy,” is explained as one of the basic principles of categorization (Rosch, 1978) and it appears that both categorization and concepts have the same function. A further function of concepts states that concepts permit inductive inference (Smith, 1989) and this can be restated as concepts as facilitating categorizations.

Despite the fuzzy distinctions between the notions of concepts, categories, and classes, there is no doubt that these ideas are very closely related. The definition of a concept given as “a mental representation of a *class* or individual and deals with what is being represented and how that information is typically used during *categorization*,” (Smith, 1989, p.502) (italics mine) clearly shows how each of these terms are inextricably united. It now appears to be clear that a concept is an internal summary of our experiences around us and one would be led to believe there is a

conceptual framework of some type in our cognitive realm that organizes concepts in some form of a structure. While these organized bundles of stored knowledge are referred to as concepts (Cruse, 2004), the set of objects to which the internal representation refers to can be roughly described as a category (Bower & Clapper, 1989).

One may ask the rationale for discussing concepts in a research like the one addressed in this dissertation. The understanding of concepts and conceptual knowledge is very important in this type of research because there is a greater acknowledgment of grasping knowledge by concepts and their relations (Wille, 1992). A more elaborate and comprehensive theory of concepts is given by Thomas Bernhard Seiler, in his work titled “Conceiving and Understanding: a Book on Concepts and Understanding,” where he describes concepts as cognitive structures whose development in human mind is constructive and adaptive (Wille, 2005). Seiler discusses numerous theories in philosophy and psychology and posits his own theory in which he states 12 aspects of a concept. The first two aspects of his approach are worth mentioning here. They are: (1) concepts are cognitive acts and knowledge units; and (2) concepts are not categories, but subjective theories (Wille, 2005).

Medin (1989, p.1469) expressed the view that “a concept is an idea that includes all that is characteristically associated with it”. He also suggests that “a category is a partitioning or a class to which some assertion or set of assertions might apply”. From the above discussion, it is plausible to conclude that there is a general understanding of a concept as mental representation and its role in facilitating categorization tasks.

Cognitive Theories of Concepts, Categories, and Relations

In cognitive and experimental psychology tradition, there are three known theories that characterize how humans represent and organize conceptual categories. These are the classical, prototype/exemplar, and probabilistic. This section reviews the extant literature about these three theoretical views followed by a discussion of new and alternative theories, i.e., the two-tiered concept representation and the theory-based approaches. The alternative theories are a result of the increasing recognition of the inadequacy of “similarity” as sufficient criteria to constrain conceptual coherence. It is motivated by the quest for a more rigorous account that explains the issue of how to represent concepts in a rich and context-dependent manner.

Although there are numerous other fields of study that investigate concepts and categories, we restrict our discussion to the cognitive theories because this research is a cognitive approach to concepts and categories. The cognitive theories in general treat concepts and categories in terms of the notion of representation and they provide numerous accounts of models and methods. Sometimes the nature of representational assumptions about concepts and categories dictate the underlying principles of these theories. At other times, the types of rules, principles or heuristics used in categorization are the bases for the cognitive theories, i.e., how one employs the concept intension to determine the extensional category of the concept (Hampton & Dubois, 1993, p.19).

Palmer’s (1978, p.262-3) cognitive representation metatheory describes five conditions that satisfy a representational system: (1) the represented (target) domain,

(2) representing (modeling) domain, (3) certain aspects of the represented domain is relevant, (4) certain aspects of the representing world are relevant, and (5) there is a systematic correspondence between the relevant aspects in the represented domain and relevant aspects in the representing domain. Barsalou (1992, p.53) elaborates on Palmer's representational system metatheory and states the representing domain is consulted for answers in the absence of the represented domain. This probably explains perfectly the situation in IR activities where indexing records (representing world) are consulted to find answers about the represented world (the bibliographic universe or documents).

The following figure (see Figure 1) is adopted from Barsalou (1992, p.53) to illustrate representational systems that satisfy the five conditions stated by Palmer.

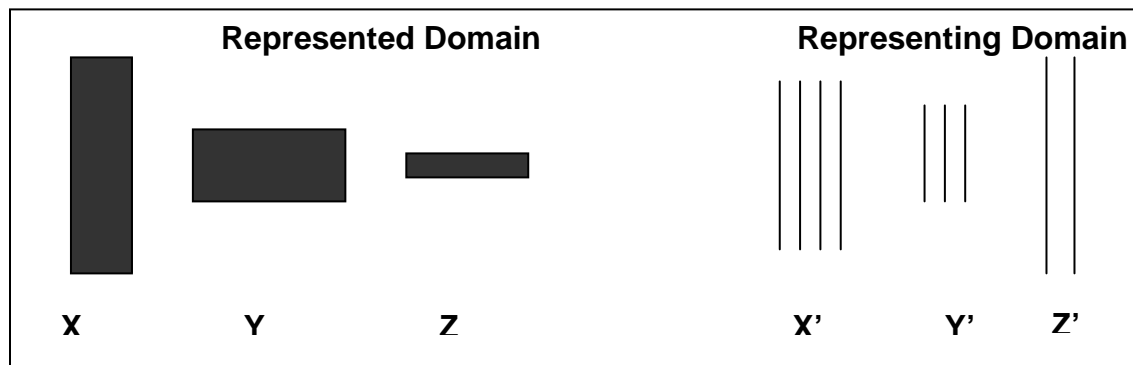


Figure 1. Palmer's five conditions of a representational system.

In the above illustration, the heights of the rectangles in the represented world are relevant, which are captured by the number of lines in the representing domain. As the height of the rectangle decreases, we have fewer numbers of lines in the representing domain. However, the heights of the lines in the representing world and the widths of the rectangles in the represented domain are not relevant.

The Classical Theory

The classical approach goes back to Aristotle. According to the classical view, concepts are defined by strict rules that are singly necessary and sufficient features for category membership (Smith, 1989; Smith & Medin, 1981). This notion leads to a conclusion that an entity becomes a member of a category if it only satisfies the category's rule, which according to Barsalou (1992, p.29), category membership in classical models is "all-or-none." The idea that common features or properties apply to category members has been, however, challenged (Rosch & Mervis, 1975). As in the "games" example, Wittgenstein having wrestled, long in search of a common criteria for category membership, later concluded that there is no such single criteria but a complicated network of similarities (Hampton, 1993).

The major problems with the classical view pertain not only to unique features applying to conceptual categories but also to the situation where this theory does not account for conceptual coherence sufficiently (Medin & Smith, 1984; Mervis & Rosch, 1981; Murphy & Medin, 1985; Smith & Medin, 1981). There is also an increasing understanding among researchers that there is a degree of variation in how well exemplars match their category prototype. Due to emerging phenomena in categorization such as graded structure, typicality, borderline cases, or goodness of exemplar, the viability of the classical theory has been seriously questioned (Barsalou, 1992, p.30).

The Prototype View

After the seminal work of Eleanor Rosch (1978), and that later of by Rosch and Mervis (1975, 1981), the prototype approach aims to provide an account of

conceptual categories based on what is known as best examples or prototypes of the categories. According to this approach, members belong to a category according to whether they sufficiently resemble the prototype or not (Hampton, 1993; Smith & Medin, 1981; Smith, 1989; Cruse, 2004). Concepts as prototypes gained popularity because attempts to find well-defined sets based on common criteria have failed. Due to the typicality effect where some members are more typical than others and because of the borderline cases or vagueness where we have no clear idea whether to assign a member to a category or not, in general gave way to the prototype approach (Hampton, 1993).

The prototype view is not, however, without its own critique. Researchers in the area of conceptual combination claim the inadequacy of the prototype view to account for complex concepts and further allege the theory is too general as a psychological theory of concepts (Osherson & Smith, 1981).

Exemplar Model

Developed by Medin and Schaffer (1978), the exemplar model views that concepts need not have a single prototype representation, but instances should be compared to a set of stored exemplars already classified. Stored exemplars or memories of exemplars are central to the exemplar model because when one wants to categorize an entity the cognitive system tries to find the exemplar memory that is most similar to the entity (Barsalou, 1992, p.27). According to Murphy and Medin (1985), the exemplar view does not offer principled accounts of conceptual structure because it does not constrain which exemplars are category members and which are not.

A more elaborate critique of the exemplar model is given by Barsalou (1992, p.27) where he argues against the assumption of the exemplar view that states the cognitive system stores a tremendous amount of idiosyncratic exemplar information for categories. The exemplar and prototype models heavily rely on the principle of similarity. The alternative theoretical views discussed below aim to go beyond the notion of “similarity” and present a richer framework. The role of similarity as a categorization principle has been questioned in recent cognitive studies (Mechelen & Michalski, 1993; Medin 1989).

The Two-Tiered Model

The two-tiered (TT) concept representation is a departure from the assumptions that concepts cannot be completely defined by necessary or sufficient features, by a prototype, or by a set of representative exemplars (Michalski, 1989, p.122). In contrast, the TT concept representation is based on the assumption that the meaning of a concept is a dynamic structure built each time anew in the course of an interaction between some initial base meaning and the cognitive agent’s background knowledge in the given context of discourse. The TT view aims to extend the classical and other views by recognizing an inherent “duality” of concept representation, which is a result of an interplay between two parts, i.e., its first component ('tier') captures most stable aspects of a concept, and its second component handles the concept's flexibility and context-dependence through a dynamic inference process (Mechelen & Michalski, 1993, p.2).

The two components also known as the base concept representation (BCR) and the inferential concept interpretation (ICI) taken together are believed to account

for an adequate cognitive model of human concept representation. While the BCR is an explicit structure residing in memory, recording both specific facts about the concept and general characteristics of it, the ICI is regarded as a process that assigns meaning to a concept using the BCR and the context. The idea that human concepts are flexible, context-modifiable (as opposed to a well-defined structure), and background-knowledge-dependent units of knowledge is closer to the constructionist view which treats semantic memory as dynamic and contextual (Michalski, 1993, p.146; Barsalou, 2001, p. 186).

According to the TT theory, the process of matching the base representation of a concept to observations is carried out by ICI, i.e., by conducting inference involving the contextual information and background knowledge. As a result the degree of match between a concept representation and the observed entity is a function of four-arguments, i.e., concept representation (CR), observed entity (OE), context (CX), and background knowledge (BK) (Michalski, 1989, p.30). The following formula describes the TT theory formalization

$$\text{Degree of match (CR, OE)} = f(\text{CR, OE, CX, BK})$$

Michalski (1993, p.156-157) illustrates the BCR component of the TT representation for the concept “Chair” as a seat with legs, made of wood or plastic, usually with 4 legs, and that is normally classed under “furniture” and as well having its own subordinates such as straight chair, armchair, and the ICI component of the TT representation of the concept “chair” as any variation that does not fall in the BCR specification and does not serve to seat a person, for example, if the chair is a

museum exhibit, or if it is a children's smaller toy object or if the context is related to an organizational structure as in chairman.

Theory-Based Approach

Another theory that stems from the notion that existing standard views on concept representation do not provide sufficient account of conceptual coherence is the theory-based approaches to categorization. Proposed by Murphy and Medin (1985), the main thesis of the theory-based view is that people's theories of the world embody conceptual knowledge and their conceptual organization is partly represented in their theories. The term "theory" is used in this approach in the sense that it refers to the causal, naïve, and mental explanations, rather than a complete, organized, scientific account (Murphy & Medin, 1985). The notion that people's concepts are tied up with their theories has been discussed long before Murphy and Medin's proposal. The fact that people's theories and knowledge of the world around them plays a significant role in conceptual coherence is equivalent to saying that people impose more structure on concepts than simple similarity would seem to suffice (Miller & Johnson-Laird, 1976; Murphy & Medin, 1985).

In their proposal, Murphy and Medin (1985) heavily emphasize on conceptual or categorization cohesiveness, which they argue will be best served by theoretical knowledge or, at least, as embedded in knowledge that embodies a theory about the world. They further explain that theories are flexible and, as a result, conceptual coherence can also be flexible. By splitting the cognitive theories on concepts and categories into "similarity-based" and "theory-based," Murphy and Medin (1985), compare the two categories on several dimensions such as concept representation,

category definition, unit of analysis, categorization basis, weighting of attributes, inter-conceptual structure, and conceptual development.

The difference in these two approaches is that the “similarity” based theory to categorization is largely dependent on attributes or features while the “theory-based” approaches is knowledge-based and the underlying principle requires “explanatory relationships” (Medin, 1989; Murphy & Medin, 1985). It can, however, be argued that the theory-based approaches is very general and probably the most difficult to operationalize. Humans have countless explanatory relations upon which they relate items that share very few or no features at all in common (Medin, 1989; Murphy & Medin, 1985).

The approach followed in the UMLS knowledge structure resembles this later new approach to categorization, i.e., the organization of concepts is knowledge-based. In the UMLS Metathesaurus, each meaning is represented as a single concept linked to the names for that meaning in any Metathesaurus source. The notion of concept rather than term is central to the purpose of the UMLS Metathesaurus. By linking different terms used to express the same concept, the UMLS Metathesaurus transcends specific vocabularies, conveys meaning, and reduces ambiguity (Schuyler, Hole, Tuttle, & Sherertz, 1993).

Cognitive Approaches in Information Retrieval

In information retrieval (IR) in particular or information science in general, there is an increasing interest in cognitive approaches which primarily focuses on user-centered research on information problems. This category of research generally comes under the banner “cognitive paradigm” and is more focused on a user’s

information seeking behavior. One widely cited theory in this regard has come to be known as anomalous state of knowledge or ASK for short (Belkin, Oddy, & Brooks, 1982). The ASK model is widely accepted to explain the motivations and reasons that drive people to engage in information exploration tasks, which is the anomalous state in the cognitive structure or the knowledge gap. Other information exploration behavioral theories include the theory of sense-making (Dervin, 1999), berrypicking (Bates, 1989), and the information foraging model (Pirolli & Card, 1998; O'Connor, Copeland, & Kearns, 2003).

The significant contribution of user-centered IR research is to place strong emphasis on incorporating cognitive elements to the application of a wide range of issues in information science in general (Allen, 1991; Belkin, 1990; Ellis, 1989; Ingwersen, 1993, 1999). A cognitive approach to the problem of information retrieval (IR) in particular or to information science in general stems from the recognition of the fact that IR is inherently an interactive process with a phenomena pertaining to machine functions and human cognition. One notable mention in this regard can be found in the early works of De Mey (1982, p.4), who after investigating the history of the cognitive theory development concluded that “any processing of information, whether perceptual or symbolic, is mediated by a system of categories or concepts which, for the information processing device, are a model of his (its) world”.

Conceptual Coherence

It is often said that related ideas cohere together. This statement is probably true because concepts and relations are the very fabric of knowledge. Concepts are the building blocks of knowledge while relations act as the glue that links concepts

into knowledge structures (Khoo & Na, 2006). In the discussion of conceptual structure and categorization, the notion of coherence holds an important place. The fact that members of a category cohere together to form a comprehensible class appears to be very easy to understand. What makes them belong to a category is a subject of intensive research. Several theoretical views have offered explanations about conceptual coherence/category cohesiveness and conceptual structure, including theory of classical view (Smith, 1989; Smith & Medin, 1981), prototype/exemplar (Rosch, 1978; Rosch & Mervis, 1975; Smith, 1989; Smith & Medin, 1981), goal-driven (Barsalou, 1992, p. 174), feature-correlation and similarity (Tversky, 1977), theory view (or knowledge-based view) (Mechelen & Michalski, 1993; Medin, 1989; Murphy & Medin, 1985), the frame view (Barsalou, 1992, p.157; Barsalou & Hale, 1993, p.124), and the two tiered concept representation (Michalski, 1993, p.146).

These theories heavily focus on the role of correlated features/property descriptions and similarity for conceptual coherence. Feature correlation and similarity are, however, regarded as inadequate because of the flexible nature of human concept use and varied ways of cross-classifying objects depending on context, goal, and the constraints imposed by human cognitive system (Markman & Makin, 1998).

Semantic relations are now regarded as important elements in explaining the coherence and structure of concepts and categories (Khoo & Na, 2006). The discussion of semantic relations together with conceptual coherence brings forth the notion of conceptual combinations. Conceptual coherence involves combining two or

more concepts to make sense of a situation or a set of situations (Thagard, 1997). This further illustrates the fact that concepts do not exist alone, rather they correlate with other concepts based on logical and/or biological necessity. Animals and artifacts have structural properties in order to fulfill various functions, so that some structural properties tend to occur with others, and certain structures occur with certain functions. In more specific terms, there are two components to conceptual coherence, i.e., the internal structure of a conceptual domain and the position of the concept to the complete knowledge base. Concepts that have their features connected by either functional structure or by causal schemata of one sort or another will be more coherent than those that do not (Murphy & Medin, 1985).

Knowledge Structure

Knowledge structure is defined as “concepts and procedures (as elements), and their inter-relationships” (Koubek & Mountjoy, 1991; Wang, 1999). This section is devoted to the discussion of knowledge structure in the context of knowledge representation schemes. Classification scheme and categorization are special cases of knowledge bases and the discussion in this section is focused to knowledge representation schemes such as thesaurus, semantic network, and ontology.

Thesaurus

The world science information system of the United Nations educational, scientific and cultural organization (UNESCO) known as UNISIST defines thesaurus in terms of its function and structure. A functional thesaurus, according to UNISIST, is a terminological control device used in translating from the natural language of documents, indexers or users into a more constrained "system language"

(documentation language, information language). In terms of structure, a thesaurus is a controlled and dynamic vocabulary of semantically and generically related terms which covers a specific domain of knowledge (Foskett, 1997). Terms are the basic elements of a thesaurus and the basic configuration of a thesaurus includes definitions, explanations, and symbols of relationships.

In IR, the development of a thesaurus as a means of navigating through the domain of knowledge is becoming popular. The primary role of the thesaurus in the IR system is to provide a grouping or classification of terms assigned to a topic area into categories known as thesaurus classes (Ng, 2000). The implementation of the thesaurus in IR systems is believed to play a major role in improving document classification by using different techniques (such as query expansion, relevance feedback, and use of thesauri) that either expand or translate the user query to the collection's indexing vocabulary (Houston, 1998). It is also believed that thesauri exhibit structures similar to human word-association networks.

The concept of a thesaurus class where it allows to group related terms of a topic into a category is important feature in thesaurus implementation in IR systems. The thesaurus class together with index terms are in turn assigned to a document either to refine or broaden the interpretation of the index term. When high frequency index terms are substituted by more specific thesaurus classes (refining), the thesaurus achieves a document to be identified more specifically. When the thesaurus class replaces or is added to low frequency terms (broadening), the chance of the document to be found by more queries increases. This way, by refining or broadening

index terms, the likelihood of documents being retrieved is increased due to a thesaurus implementation (Oroumchian, 1995).

Ontology

The discussion about ontology is very rich in the literature of philosophy. However, there is an increasing trend to treat ontology in information science as a formally structured terminology or controlled vocabulary. Thomas Gruber defines ontology as “an explicit specification of a conceptualization” (Zúñiga, 2001). Despite its heavy usage in the literature of philosophy, the concept of ontology has been embraced by researchers in information science as of late to address the issue of knowledge representation (Smith, 2004). The treatment of ontology in information science/systems is, however, different from the philosophical treatment of ontology. While philosophical ontology is concerned with the representation of universals and particulars in reality, the information science/systems discipline ontology is a formal language designed to represent a particular domain of knowledge (Zúñiga, 2001).

Notable mentions of well developed and maintained information systems ontology include the UMLS, the open biological ontologies (OBO), and LinkBase ®. The UMLS semantic network is highly regarded as a higher order ontology framework where attempt has been made to provide a higher-level category structure of 135 types which are abstracted from the more than 1.3 million concepts in the Meta.

Semantic Network

Wood's "What's in a Link" paper has stirred a growing interest in the development of formalization languages in the study of knowledge representation

(Allen & Frisch, 1982). Semantic network (SN) as a knowledge representation tool is exhibited when information is conveyed in node-edge graphical form. In SN specification, the nodes are referred to as subjects and the edges as relationships, where those nodes and edges are assigned meanings and the topology of the graph is significant to those meanings (Griffith, 1982). There are basically three aspects of relationships between two subjects, i.e., the existence of the relationship, the type of relationship, and the ordinality or semantic sense of direction. There is no definitive claim that suggests semantic networks exist physically in the brain. Instead, they are viewed as graph-theoretic structures of relations and abstractions whose primary aim is to impart “common sense” to computer applications (Lehmann, 1992). The most widely used relationships or links between nodes are *is-a* and *part-whole*.

In order to achieve representational adequacy and cognitive validity for generalized reasoning over realistically complex domains, semantic networks are required to incorporate richer relational links. Markowitz, Nutter, and Evens (1992) describe three classes of links, namely queuing, similarity, and part-whole, that they say are necessary for adequate conceptual representation in SN. Queuing is a link that expresses order or sequence. Similarity is the name that expresses a strong correspondence between two nodes or referents and forms the basis of human categorization and is a component of many cognitive activities. Others, however, view part-whole relation as a family of at least four distinct relations, i.e., functional components (such as a bicycle and wheel), members of sets, set inclusion (or is-a), and pieces cut from otherwise undifferentiated wholes (such as a slice or pie) (Iris, Litowitz, & Evens, 1988; Markowitz, Nutter, & Evens, 1992). It is argued that these

four relations of the part-whole have different logical properties, and semantic network models that provide only one part-whole link will necessarily make errors in reasoning, at least some of the time, either by failing to support correct inferences or by warranting illegitimate ones.

The emphasis in any knowledge representation scheme is to simulate the human memory and reasoning process for use in computer information systems. The adequacy of the representation scheme including semantic networks to model human memory and human reasoning processes is therefore a great concern to all involved in the research. Markowitz, Nutter, and Evens (1992) argue that semantic networks need a wide range of links that are basic to conceptual information processing in people and make use of the logical properties of those links in modeling human reasoning. Looking at what Markowitz, Nutter, and Evens (1992) suggest, it is plausible to support the notion they propose that queuing and similarity should be part of the semantic network models on top of considering the part-whole (part-of) relations as a family of four distinct relations with distinct logical properties.

The application of semantic network in artificial intelligence (AI) is popular. However, the semantical and epistemological foundations of these applications are quite unclear. The main question that is being asked in this regard is whether semantic networks are something like language formalizations or perhaps something like "mental conceptual structures" (Hautamaki, 1992). If semantic networks are to be viewed in terms of the classical doctrine of concepts as mental constructs, which form a conceptual space, semantic networks are representations of such conceptual spaces.

Semantic Memory

Memory is often considered as a record of our experiences. Knowledge of semantic memory is central to the understanding of this dissertation. The whole exercise in the cognitive view approach for information retrieval systems has been to model the workings of the semantic memory such that it can be adequately simulated in computer based knowledge representation systems. Semantic memory has been characterized as our mental storehouse of knowledge about language as well as general knowledge about the world (Khoo & Na, 2006; Smith, 1978). Central questions asked in the literature of semantic memory include how semantic information is stored and organized in human memory or in other words what constitutes a reasonable view of how semantic information is organized within a person's memory?(Quillian, 1968). The answer for these questions are, however, offered by different competing theories in fields of studies such as philosophy, psychology, linguistics, and natural language processing in computer science.

Some of the well known memory organization theories include taxonomic, thematic (frame), hierarchical (Barsalou, 1992, p.125-127); and semantic network with spreading activation (Loftus, 1975). The whole discussion in cognitive psychology about semantic memory in one form or another comes down to the notion that cognitive and memory structures consist of nothing more than an aggregate of associated elements. The memory model, according to Quillian (1968), consists of basically of a mass of nodes interconnected by different kinds of associative links.

It is usually assumed that for any one word meaning there is exactly one and only one “type node” in the memory and as a result the general structure of the

memory model is one consisting of “planes,” each made up of one type node and a number of token nodes. The token nodes have a pointer to the same unique type node for the concept. The challenge to mimic human semantic memory, therefore, remains to be one of developing a model that aims to link nodes together into configurations that are as varied and rich as the ideas expressed in natural language. Most of all, memory is a complex network of attribute-value nodes and labeled associations between them where each association create both within-plane and between-plane ties, with several links emanating out from the typical token node and many links coming into almost every type node (Quillian, 1968).

The UMLS Knowledge Sources

The UMLS is a complex knowledge source with millions of concepts and relationships between the concepts. The concepts are designed to represent the field of bio-medicine in both clinical, research, and administrative respects. Developed by the National Library of Medicine (NLM) in 1986, the current UMLS release (2006AC) contains more than 1.3 million concepts and 6.4 million unique concept names from more than 100 distinct vocabulary sources (US National Library of Medicine, 2006). At the start of this grand project, aside from the aim of seeking a capability to integrate information from disparate sources, including bibliographic databases, patient clinical records, and knowledge bases, the major goal had been to address the complex problem of relating user inquiries to the content of biomedical information sources (Nelson, Powell, & Humphreys, 2002). Stated in other words, the chief goal of the project had been to develop UMLS knowledge sources that address medical vocabulary problems by "improving the ability of computer programs to

“understand” (italics mine) biomedical meaning in user inquiries and then using this understanding to retrieve and integrate relevant machine-readable information" (Houston et al., 2000; Perl & Geller, 2003; Nelson, Powell, & Humphreys, 2002; Schuyler, Hole, Tuttle, & Sheretz, 1993)

The UMLS is basically a suite of three knowledge sources, namely the Metathesaurus (often known as the Meta), the Semantic Network (SN) and the SPECIALIST Lexicon (Nelson, Powell, & Humphreys, 2002). The two major components that are particular interest to this study are discussed below.

Metathesaurus

The central vocabulary component and the most complex of the UMLS knowledge sources is the Metathesaurus (Meta). The word “Meta” is meant to signify "more comprehensive, transcending," because its scope is determined by the combined scope of its source vocabularies, more than one hundred separate vocabulary sources (Nelson, Powell, & Humphreys, 2002). As much as possible, the Meta preserves the meanings, hierarchical connections, and other relationships between terms present in its source vocabularies. In certain instances, it adds certain basic information about each of its concepts and establishes new relationships between concepts and terms from different source vocabularies. The notion of concept instead of term is fundamental to the overall configuration of the Meta. By linking different terms used to express the same concept, the Meta is regarded as having transcended specific vocabularies, conveyed meaning, and reduced ambiguity (Schuyler, Hole, Tuttle, & Sheretz, 1993).

The Meta is organized by the principle of semantic locality, a principle that brings concepts close to one another in meaning in the same area. Each separate concept in turn is linked to other concepts through relationships. It is this web of relations that basically forms the knowledge structure in the Meta and this is an indication that concepts do not exist on their own but with relation to others. It also shows that concepts underlie the bases of the organization in the Meta (Houston, 1998). The concept structure of the Meta is, thus, one which aims first to bring alternative names for the same concept and then to establish a relationship between different concepts (Cimino, Min, & Perl, 2003). While many of the relationships are taken directly from the source vocabularies, altogether, there are nine types of relationships in the Meta, namely broader, narrower, other related, like, arent, child, sibling, AQ (is an allowed qualifier for a concept in Meta source vocabulary), and QB (can be qualified by a concept in a Meta source vocabulary) (Nelson, Powell, & Humphreys, 2002).

Each Meta concept is in turn assigned one or more semantic type from the semantic network, the second major component of the UMLS. The semantic types are a higher-order categorization of the concept and are intended to unify and reduce the complexity of the Meta. The assignment of the semantic type to the Meta concept is such that the most specific semantic type available in the hierarchy is assigned to the concept and the level of granularity may be different (Nelson, Powell, & Humphreys, 2002)

The UMLS Semantic Network

The semantic network (SN) of the UMLS is regarded as a graph-theoretic structure (Bodenreider & McCray, 2003) or as ontology for the biomedical domain (McCray & Nelson, 1995). Despite the naming, the SN is a higher-level general categorization consisting of 135 semantic types that aims to unify the millions of concepts in the Meta (Appendix A presents the names of the 135 semantic types and their hierarchical structure). The nodes in the SN are organized into two high-level, single inheritance hierarchies known as entities and events (Bodenreider & McCray, 2003; McCray & Nelson, 1995). This specification of the biomedical domain into entities and events is equivalent to saying that everything there is to be represented is either an entity or an event. The UMLS SN defines entities as “a broad type for grouping physical and conceptual entities” and events as “a broad type for grouping activities, processes and states” (U.S. National Library of Medicine, 2005). In the ontological literature, entities and events are regarded as “continuants” and “occurants,” respectively (Schulze-Kremer, Smith, & Kumar, 2004).

Entities and events serve as root-node in the SN and the remaining types descend from these top-nodes. The semantic types (or types) are linked to one another by two types of semantic relations, i.e., “*isa*” and “*associative*” type. By way of these relations, a hierarchy of type is established within the semantic network. The “*isa*” relation type allows nodes (semantic types) to inherit properties from higher-level nodes. Moreover, there are five categories of associative relations that link the semantic types (Bodenreider & McCray, 2003). The 54 semantic relation types are discussed in detail in the section below.

Inherent properties (e.g., a mammal is a vertebrate) or attributed features (e.g., a professional group is a set of individuals classified by their vocation) are the criteria used to group items in the type hierarchy (McCray & Nelson, 1995). Within the type hierarchy, however, the place of the types is determined based on their definition, whether the definition is based on inherent or attributed features. In other words, types are arranged according to their intensions (the manner in which they are described) rather than their extensions (what they refer to in the real world) (McCray & Nelson, 1995).

The semantic types are abstracted from the UMLS Meta. However, the semantic relationships between the semantic types do not necessarily filter down to instances of concepts in the Meta that have been assigned to those semantic types. The relation type, "treats", for example is one of several valid relations between the semantic types "pharmacologic substance" and "disease or syndrome." Penicillin is an instance of a concept from the Meta that has been assigned the semantic type "pharmacologic substance," and AIDS is an instance of a concept from the Meta that has been assigned the semantic type "disease or syndrome." Though the relation "treats" holds between the semantic types, it does not hold true between the instances of penicillin and AIDS (Nelson, Powell, & Humphreys, 2002; Ng, 2000; Schuyler, Hole, Tuttle, & Sheretz, 1993).

This does not mean, however, inter-concept relationships in the Meta do not instantiate specific low-level knowledge, such as "aspirin treats fever". This rather has to do with one of the principles used in building the UMLS and that is parsimony. The principle of parsimony is aimed to prevent unneeded categories from being

represented. Three principles of parsimony guide the construction of the SN and the way Meta concepts are categorized. These principles, according to Burgun and Bodenreider (2001), are (a) assign the most specific semantic type available; (b) assign multiple semantic types if necessary; and (c) assign a less specialized semantic type (super-type) if no more specific semantic type (subtype) is available.

Semantic Relations

The UMLS knowledge structure is not merely a list of concepts but of concepts linked to one another through relations. As concepts and relations are the foundation of knowledge and thought in humans they are as well the foundations of the knowledge structure in the UMLS. The view that human perceptual system automatically segments the world into concepts and categories is a widely held notion. Concepts are generally regarded as the building blocks of knowledge and, while relations act as the glue that link concepts into knowledge structures (Khoo & Na, 2006), relations are needed not only for cementing concepts into a coherent structure but they are also crucial for reasoning and inferencing. Semantic relations in the UMLS are meaningful associations between concepts in the Meta and between concepts in the Meta and the semantic types, and, of course, between types in the SN.

The UMLS has 54 semantic relations (Appendix B presents the semantic relations) which are organized under two root nodes namely *isa* and *associated_with* (U.S. National Library of Medicine, 2006). The *isa* relation type organizes semantic types into a hyponymy hierarchy where types in the hierarchy can inherit properties from higher-level nodes. The *associated_with* relations, on the other hand, are non-hierarchical relation types that are organized under five categories which themselves

are relationships, i.e., "physically related to"; "spatially related to"; "temporally related to"; "functionally related to"; and "conceptually related to" (McCray & Nelson, 1995). The semantic relations are binary in nature because they establish a link between two semantic types in the SN. One more characteristic of these relations is that the arguments of these binary relations are ordered, and most of the relations are asymmetric, meaning the relation never holds in the opposite direction. Without considering the *isa* relation type, it is estimated that there are about 7000 semantic (non-hierarchical) relationship instances in the remaining 53 associative relation types (Zhang, 2004).

In the literature of linguistics and psychology, there is as well a great deal of treatment about semantic relations. Semantic relations can refer to relations between concepts in the mind (called conceptual relations), or relations between words (lexical relations) or text segments (Khoo & Na, 2006; Lyons, 1977). It can be argued, however, concepts and relations are so inextricably bound with language and text, and it is difficult to analyze the meaning of concepts and relations apart from the language that expresses them. Wittgenstein has been quoted saying, "When I think in language, there are not "meanings" going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought." As a result, the distinction between conceptual relations (psychological) and lexical relations (linguistic) are irrelevant and researchers use the term lexical-semantic relations to refer to relations between lexical concepts denoted by words (Khoo & Na, 2006).

There is strong evidence that semantic relations play a critical role in how we represent knowledge psychologically, linguistically, and computationally, and the

first thing for any knowledge representation system to do is to specify the distinction between entities and relations (Green, 2001, p.6). Relations are also considered as constraining factors in the knowledge structure. Relations are normally involved as we combine simple entities to form more complex entities, as we compare entities, as we group entities, as one entity performs a process on another entity, and so forth (Green, 2001, p. 3). Semantic relations are also characterized by three major logical properties, namely reflexivity, anti-symmetry, and transitivity (Burgun & Bodenreider, 2001; Cruse, 2004; Sowa, 1984, p. 381).

Research in experimental psychology has a great deal of evidence that accounts how humans recognize and perceive semantic relations. Chaffin and Herrmann (1987; 1988) carried out a series of studies to demonstrate that people can distinguish between different types of relations, identify instances of similar relations, express relations in words, recognize instances of relation ambiguity, and create new relations. In an aim to answer if semantic relations can be treated as concepts, Chaffin and Herrmann (1988) conducted a study and concluded that relations have the main characteristics of concepts and, according to their research, relations are abstract concepts. They identified four characteristics that relational concepts share with concrete concepts: (a) relations can be analyzed into more basic elements or features; (b) a new relation may be an elaboration or combination of other relations; (c) relations have graded structure (i.e., some instances of relations, represented by word pairs, are more typical of a particular relation than others); and (d) relations vary in the ease with which they can be expressed.

Types of Semantic Relations

There are different types of semantic relation categories and there is no complete taxonomy of relation types in the extant literature. The most widely recognized semantic relation types in controlled vocabulary or thesaurus relationships are three: equivalence (synonymy), hierarchical relations (narrower than, broader than, part-whole), and associative relations (related along some dimension) (Bean & Green, 2001; McCray & Nelson, 1995; NISO, 2005). Another classification of relation types includes paradigmatic and syntagmatic relations (Cruse, 2004; Khoo & Na, 2006). According to Ferdinand de Saussure, paradigmatic relations (the equivalent of associative types) are relations between pairs of words or phrases that can occur in the same position in the same sentence (de Saussure, 1959, as cited in Asher, 1994, v.10, p.5153). Syntagmatic relations refer to relations between words that co-occur (often in close syntactic positions) in the same sentence or text (de Saussure, 1959, as cited in Asher, 1994, v.10, p.5178).

Another category of relation types constitutes five well-known paradigmatic relations: hyponym-hyperonym relation, troponymy relation, meronym-holonym relation, synonymy, antonymy; and one syntagmatic relation type that is known as cause-effect relation (Cruse, 2004; Khoo & Na, 2006; McCray & Nelson, 1995).

Data-Analytic Methods

In response to the cognitive views and theories of concept representation and categorization tasks, several data-analytic methods have been developed and empirically validated. These data-analytic methods are designed to provide the necessary tools to explicate subjective as well as objective data gathered from

cognitive performance tasks. Under the general term “representation of proximity data,” the data analysis methods for concept representation and categorization generally deal with issues of aggregation and characterization (Feger & De Boeck, 1993). While the aggregation issue addresses the question of which objects and entities form a comprehensible class the characterization issue deals with the intension of the category (theories, rules, attributes, or explanations one uses to characterize the category). In view of human cognition, the issue of whether a category is discrete or continuous is a source of contention. However, for the purpose of the aggregation issue to be explored by the data analysis methods, categories are required to be treated as discrete phenomena.

Much of the existing data analysis methods that provide operationalization of the representation of proximity relations focus largely with the development of theories that address similarity relations and the construction of scaling procedures for describing and displaying such similarities between entities. These representations of proximity data are generally divided into two major categories, i.e. spatial and network models (Sattath & Tversky, 1977). The classes of spatial models (also known as multidimensional scaling or MDS) represent entities as points in a low-dimensional coordinate space so that the metric distances between the points reflect the observed/experimental proximities between the entities. On the other hand, the network models represent entities as a node in a connected graph such as a tree so that the relations between the nodes in the graph reflect the observed proximity relations among the entities.

A further review of the literature about data analytic methods for proximity relations distinguishes four types of models, i.e., spatial/dimensional, cluster, set-theoretic, and graph-theoretic models (Corter, 1996). Because there exists a series of models for cognitive theories in the study of concepts and categories, selecting a model that relates to the underlying cognitive view about concept and category relationship is very important. The following section discusses the four classes of models used in the study of proximity relations together with their main features and weaknesses.

Spatial/Dimensional Models

These models represent proximities among objects by locating the objects as points in a low-dimensional geometric space (Corter, 1996; Davison, 1983). The methods that fall under the general name “spatial/dimensional models” include multidimensional scaling (MDS), principal component analysis (PCA), and correspondence analysis which define proximities using variance and covariance (Murtagh, 1993). The underlying assumption in MDS scaling is that dissimilarities and distance are monotonically related, that is, the larger the distance in the configuration of points, the larger the dissimilarities of the experimental data, and vice versa (Kruskal, 1964a, 1964b). The degree to which concepts are related to one another is expressed using words like semantic similarity, semantic relatedness, and semantic distance. The distance metaphor comes from an analogy to a multidimensional space where concepts are located according to their values on various dimensions of meaning (Schvaneveldt, Durso, & Mukherji, 1982). Concepts

near one another in multidimensional space are treated to be more closely related to one another than are concepts that are farther apart in the space.

In his seminal work on MDS, Kruskal (1964a, 1964b) gives a quantitative measure for nonmonotonicity by performing a monotone regression of distance upon dissimilarity, and uses a normalized residual variance. Kruskal calls this normalized residual variance (or the residual sum of squares) stress or objective function.

Kruskal's Stress is given by the formula:

$$\text{Stress} = \left[\sqrt{\frac{\sum_i \sum_j [f(\delta_{ij}) - d_{ij}]^2}{\sum_i \sum_j d_{ij}^2}} \right]$$

where inter-point distance (d_{ij}) is expressed as a function of the ranked dissimilarities $f(\delta_{ij})$ (Kruskal, 1964a, 1964b). Based on Kruskal's procedure, n points are positioned in an m -dimensional space where an attempt was made to find the best possible approximation to a monotonic relationship between $\binom{n}{2}$ inter-point distances and an experimentally obtained ranking of the dissimilarities among all pairs of the N objects. A lower Stress is regarded as a better fit. Kruskal (1964a) provides the following verbal evaluation of Stress, which is normally a "residual sum of squares" and is positive.

A stress is always a positive number and dimensionless. According to Kruskal (1964a), Stress can be expressed as a percentage and he presents the following verbal evaluation: 20 % stress = poor goodness of fit; 10% stress = fair goodness of fit; 5% stress = good goodness of fit; 2.5% stress = excellent goodness of fit; and 0% stress = perfect goodness of fit.

One widely used method to obtain proximity data, especially for large stimulus sets (between 50 -100 objects) is to ask research subjects to sort the stimuli according to perceived similarity (Burton, 1975; Kruskal & Wish, 1978; Rosenberg & Kim, 1975; van der Kloot & van Herk, 1991). The method of sorting is relatively direct and has the advantage of convenience of administration and minimal effort for the research subjects (Miller & Johnson-Laird, 1976, p.254). Traditionally, a list of words are selected and each word is printed on a separate card. The pack of cards is then given to research subjects who are asked to sort them into piles on the basis of similarity of meaning. If n words are used, a matrix containing $n(n-1)/2$ entries (one for each of the pairs of words) is constructed, and the number of times each pair is put into the same pile is counted. after many judges have performed the sorting, the entries in the matrix give the number of judges who thought each pair similar enough to put them in the same pile; those entries can be regarded as measures of semantic proximity between all pairs of words on the list and can be analyzed by any of several alternative procedures to discover the underlying structure (Miller & Johnson-Laird, 1976, p.254).

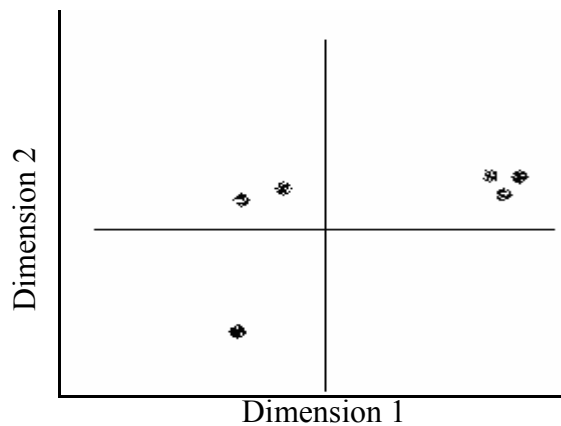


Figure 2. Spatial/dimensional model.

Clustering

There are at least three major types of representational models that fall under cluster models. These are (1) partitioning methods, (2) hierarchical clustering, and (3) overlapping nonhierarchical clustering (Corter, 1996). The overriding factor in all of these models is to group data into a set of clusters in which each set contains relatively similar data or objects. Partitioning methods are designed to find a set of clusters that correspond to mutually exclusive and exhaustive subsets of the set of objects being analyzed. In a partition, each object is a member of exactly one cluster (Corter, 1996).

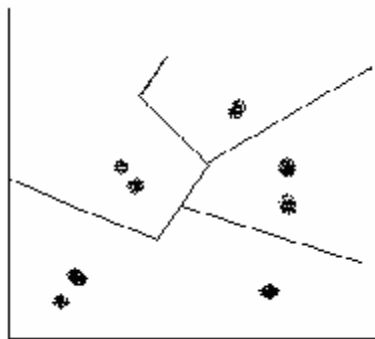


Figure 3. Partition model.

Hierarchical clustering methods are techniques to find sets of clusters that are restricted to be nested, that is either each pair of clusters must be disjoint (i.e., have no objects in common) or one cluster must be included in the other (i.e., the objects composing one cluster are a subset of the items in the other cluster) (Carroll & Corter, 1995). The nested set of clusters resulting from a hierarchical clustering method can be represented by a tree graph, or “dendrogram.” (Corter, 1996).

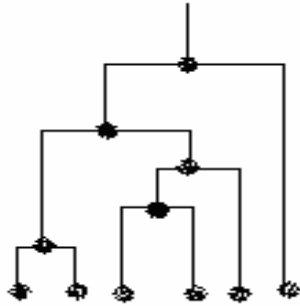


Figure 4. Hierarchical clustering model.

Overlapping non-hierarchical clustering methods fit sets of clusters that are not necessarily restricted to this hierarchical or nested relationship; instead, the clusters may overlap in arbitrary patterns. A partition is a special case of a hierarchical clustering solution, which is a special case of an overlapping nonhierarchical clustering. (Corter, 1996, p.6-7).

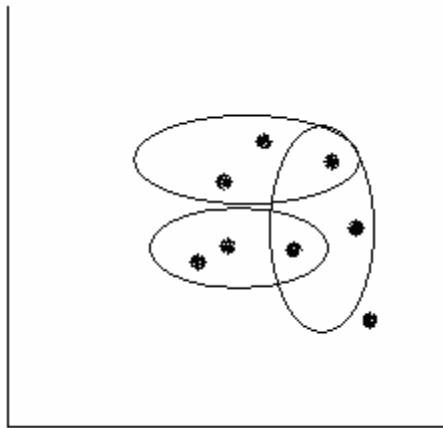


Figure 5. Overlapping nonhierarchical model.

Set-Theoretic Models

Tversky (1977) presented a mathematical model of similarity relations, termed the contrast or “feature-matching” model. The contrast model analyzes the similarity between two objects as a function of the number and salience of the discrete features shared by the objects (their “common features”) and the number and salience of

features that each object has that the other does not. The contrast model expresses the dissimilarity between two objects x and y as

$$d(x, y) = \Theta g(X \cap Y) + \alpha f(X - Y) + \beta f(Y - X)$$

where X and Y represent the feature sets associated with objects x and y respectively, and α and β are nonnegative weights. The set functions g and f define the saliences or weights of individual features and how they are combined to yield the overall contributions of the three relevant feature sets. These sets are $X \cap Y$, which denotes the common features of X and Y ; $X - Y$, which denotes the distinctive features of x (with respect to y); and $Y - X$, which denotes the distinctive features of y (Corter, 1996, p.8).

Graph-Theoretic Models

Semantic network representations are usually modeled using graph-theoretic approaches. A graph is a type of representation composed of nodes and directed arcs or lines connecting the nodes (Arocha, Wang, & Patel, 2005). In a semantic network, knowledge is represented as a net-like graph, where nodes represent conceptual units and the directed links or arrows between the nodes represent relations between the units (Lehmann, 1992). The proximity between two objects is usually modeled by the distance, defined as the length of the minimum-length path between the two corresponding nodes. In a typical application that seeks to model a set of proximities by a graph, both the set of arcs that are defined between pairs of nodes and the weight of each arc are parameters to be estimated.

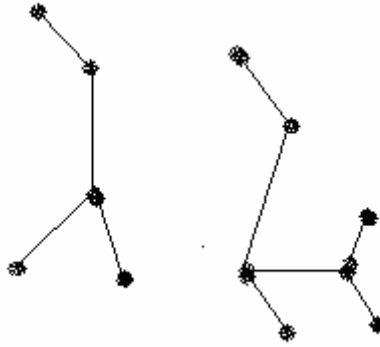


Figure 6. Graph-theoretic models.

Summary

This Chapter has presented a synthesis of extant literature pertaining to the statement of the problem of this research. In general, an attempt has been made to provide the context and theoretical framework within which this research falls. By expounding the statement of the problem in Chapter 1, the literature review helped to shape and delineate the nature of this study from related studies. The emphasis has been to critically evaluate the literature in the area of cognitive theories that provide accounts on the representation of concepts and categories. The theory-based approaches to concept representation offers a more comprehensive theoretical view of conceptual coherence.

The notion that a human conceptual system is profoundly flexible, dynamic, contextual, non-modular, and modal (Barsalou, 2001) is central to this research. Through the review of the literature it has been made clear that existing cognitive theories such as the classical, prototype, and even similarity approaches do not provide sufficient account of the type of conceptual system that is needed to offer required “common-sense” to systems. Discussions were also made regarding issues surrounding knowledge representation and knowledge structure in connection with

the UMLS knowledge sources. Finally, range of data-analytic methods including the spatial/dimensional model, clustering model, set-theoretic model, and graph-theoretic model were discussed. The lesson from the review of literature is one that concepts and categories are fundamental building blocks to both human and systems knowledge bases and only a rich and dynamic theoretical view can provide a better explanation of the profoundly flexible knowledge structure.

CHAPTER 3

MATERIALS AND METHODS

Introduction

This Chapter presents the steps and procedures that were followed to address the research questions discussed in Chapter 1. From the cognitive theories reviewed in Chapter 2, the “theory-based approaches” to concept representation (Medin, 1989; Murphy & Medin, 1985) provided the general framework for the design of this research. The data analysis methods that were considered to offer a better account to the formalization of the theory-based approach and thereby address the issue of aggregation and characterization (Feger & De Boeck, 1993) were Kruskal’s nonmetric multidimensional scaling (1964a, 1964b), Ward’s minimum variance hierarchical clustering method (Hartigan, 1967; Romesburg, 1984, p.30), and content analysis (Weber, 1990).

Research participants generally performed two sets of cognitive tasks online, for which software is being licensed (websort.net). The first set of task required research participants to sort 60 food names into piles/categories followed by description of each piles/categories. In order to understand the rules or criteria subjects use to base their categorization task, the text descriptions (corpus) was analyzed using the method of content analysis (Weber, 1990). The second task involved sorting/classifying relation types in the associative family into five categories identified a priori and was guided by the psychological theory of semantic relations (Chaffin & Herrmann, 1987; Chaffin & Herrmann, 1988).

Stimulus Materials

The UMLS knowledge sources are complex structures. This study based its exploratory investigation on two of the UMLS knowledge sources components, i.e., the Metathesaurus (Meta) and the semantic network (SN). The 2006AC version of Meta contains more than 1.3 million unique concepts from more than 100 separate vocabulary sources (U.S. National Library of Medicine, 2006). The semantic network, on the other hand contains 135 semantic types and 54 semantic relations (U.S. National Library of Medicine, 2006). This is a complex knowledge structure and even the smallest amount of sample size for investigation by humans is overwhelming. As a result, the selection of the 60 food names was based on the theory of indexing (Salton, 1997) where atomic elements are favored to characterize document objects and based on the atomistic theory of concepts (Fodor, 1998).

The stimulus materials for this study were composed of 60 food names taken from semantic type known as “food,” and all of the 47 semantic relations from the “associated_with” relation categories (U.S. National Library of Medicine, 2006). Regarding the selection of the food names, the question of how many and which of the food names to include is addressed by selecting “most common” food names and appropriate number of food names that would not impose cognitive strain on the part of the participants (Coxon, 1999, p.12). Because “most common” (culturally shared) as opposed to idiosyncratic ones is ambiguous on its own, the initial 60 food names selected were given to three graduate students to review them if there are food names they are unsure of. As a result of this pre-test exercise, three food names were replaced.

On the question of how many food names, because the task involves description of derived categories, and because 50-100 objects are considered as large stimulus sets for sorting (Kruskal & Wish, 1978), 60 food names were selected to be an appropriate size. The notion of domain specification is also given due consideration because the set of stimulus items selected needs to refer to a conceptual coherent boundary, or that they jointly refer to a single conceptual sphere (Coxon, 1999, p.9). This assertion is supported by the fact that all of the 60 food names share and refer to a specified domain, or single conceptual sphere, i.e., food.

Food is defined in the semantic network as ‘any substance generally containing nutrients, such as carbohydrates, proteins, and fats, that can be ingested by a living organism and metabolized into energy and body tissues. Some foods are naturally occurring, others are either partially or entirely made by humans’ (U.S. National Library of Medicine, 2005). Appendix C presents the 60 food names in alphabetical order. The tree hierarchy of the semantic type “food” in the semantic network of the UMLS is presented below:

- A Entity
- A1 Physical object
- A1.4 Substance
- A1.4.3 FOOD

The second set of stimulus material comprised semantic relations from the associative type relations in the SN of the UMLS. The 2006AC version of the semantic network (SN) has 54 semantic relations organized under two root nodes, i.e., “*isa*” (hierarchical) and “*associated_with*” (nonhierarchical), which themselves are

relations. The associated_with relation type is a family of relations (a total of 53) grouped under 5 categories, i.e., physical relations (e.g., *connected_to* as in body space or junction *connected_to* tissue), spatial relations (e.g. *location_of* as in anatomical structure *location_of* virus), functional relations (e.g., *causes* as in fungus *causes* pathologic function), temporal relation (e.g., diagnostic procedure *precedes* therapeutic or preventive procedure), and conceptual relation (e.g., *property_of* as in amino acid sequence *property_of* gene or genome).

The following structure (see Figure 7) depicts the five categories of associative relation types together with their definition (U.S. National Library of Medicine, 2005). Using the 53 associative type semantic relations, it is estimated that there are about 7000 categorical links between the 135 semantic types (Zhang, 2004).

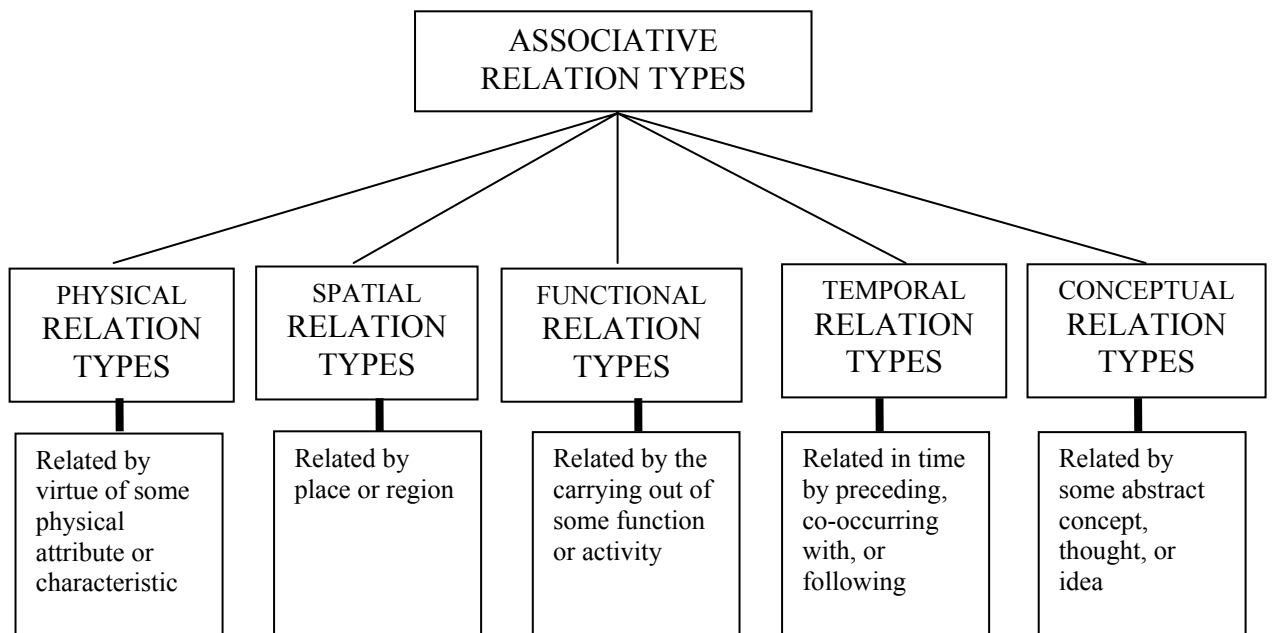


Figure 7. Categories of the associative relation types.

Research Participants

In cognitive and experimental psychology, the participation of human subjects to perform cognitive tasks is a common practice. The purpose of this study was to understand how humans use, organize, and represent concepts and the participation of human subjects was central to the conduct of the research. Examples of related cognitive tasks in semantic organization include sorting 60 personality trait adjectives (van der Kloot & van Herk, 1991), sorting names of behaviors and names of occupations (Burton, 1975), sorting 15 mutually exclusive kinship terms (Rosenberg & Kim, 1975).

A total of 89 graduate students and faculty from the School of Library and Information Sciences, the school of Hospitality Management, and Biology department at the University of North Texas participated in this study. Divided into two groups, participants performed two tasks. The first group, comprised of 49 participants, sorted 60 food names into categories followed by the description of derived categories. The second group, having 40 participants, performed sorting/classification of 47 semantic relations into 5 categories known a priori.

For the type of cognitive performance employed, the number of participants who took part in this study is above the recommended number to obtain a reasonable structure (Tullis & Wood, 2007). Despite continued effort to create a balance between male and female participants, the majority of the participants remained to be female (73% and 70% respectively for the two tasks). Participants were all native U.S. English speakers. Recruiting participants from biology and hospitality

management with food science background helped to enrich the data collection and analysis experience.

Procedure

Research participants were formally invited by letters (Appendix D & E). The researcher used the opportunity of the classes he assists as a teaching assistant to explain about the nature of the task and to distribute the invitation letter. Some distance students were emailed the invitation letter with a how to perform guideline. Addresses of web sites where participants visit to perform the tasks were given together with the invitation letter. Contacting participants resumed immediately after the IRB approval letter is obtained (Appendix K). Three weeks after the initial distribution of the invitation letters, the researcher took another opportunity of the class he taught to remind students to complete the task

Without going into further details not to bias participants on how they perform the task, the researcher used the opportunity of the face to face class to demonstrate the mechanics of sorting and providing descriptions of categories. Only relevant mechanics of the overall task performance were shown. The first task was based on free sorting technique where participants were free to sort, re-shuffle, and re-organize all over again. More over, they were also required to provide a name for the categories they are creating followed by a description of their rationale about their judgment of categorization. The second sorting task was based on closed sorting where participants were only required to sort the 47 semantic relations into 5 classes defined for them. The definitions of the five classes of relations were intended to facilitate

easier sorting of the relation names which in experimental psychology research is known as the “effect of priming” (Rosch, 1975).

In addition to the face-to-face demonstration of the online sorting procedure, detailed instructions explaining the purpose of the study and how to perform the cognitive task was attached to the invitation letter. The software used for the sorting task also allows participants to view an animated demo before starting the task. The data collection was fully launched after a pre-test was done involving 7 students. The pretest helped to better reword the invitation letter and the instruction guideline, which also required obtaining a second approval from IRB. The importance of pretest in the use of sorting as a data collection method is highly emphasized because it is believed that it will offer important clues on the reliability (stability) of the sorting and the ability of subjects to produce similar sortings on the same criterion (Coxon, 1999, p.15).

Data Collection

The data collection was premised on two psychological and cognitive theories that guided this study. The first theory, the theory-based approach to categorization (Medin, 1989; Murphy & Medin, 1985), states that human’s theories of the world, (their mental explanations) embody conceptual knowledge such that concepts are organized by those theories. According to Murphy and Medin (1985), the notion of “theory” manifests five general properties, that are, (a) “Explanations” of a sort, specified over some domain of observation, (b) simplify reality, (c) have an external structure – fit in with (or do not contradict) what is already known, (d) have an

internal structure – defined in part by relations connecting properties, and (e) interact with data and observations in some way.

The second theory that guided this study and the data collection method is known as “psychological theory of semantic relations” (Chaffin & Herrmann, 1988, p.289), which prescribes certain phenomena that such a theory should account for. The phenomena that are of interest to this study were the ability of humans (a) to judge some relations as more similar than others, (b) to distinguish one relation from another, and (c) to identify instances of common relations (Chaffin & Herrmann, 1988, p.289). Together, these two theories helped to frame the selection of appropriate data collection and analysis methods, which are essential components of an exploratory study, i.e., results and interpretations are guided by the methods and analytic tools. These methods are sorting, category description, and sorting or classification of semantic relations, which are described below in the order.

Unconstrained Sorting

Sorting has been widely used as a data collection method for various types of cognitive tasks. Sorting is based on the underlying assumption of discrimination and classification of stimuli and is in general close to the very basic operations in cognition and language (Coxon, 1999, p.1). The use of sorting as a data collection method, particularly when dealing with large number of objects is justified due to its economy and simplicity (Kruskal & Wish, 1978, p.10; Rosenberg & Kim, 1975). A pair wise comparison of 60 food names would have required participants to perform 1770 comparisons ($60 \times 59 / 2$), compared to the unconstrained sorting used in this study. Unconstrained or free sorting is one of several cognitive methods which can be

used to obtain judgmental data about semantic organization in a domain of knowledge or for collecting data about cognitive structures and concepts (Burton, 1975; Coxon, 1999; Harloff, 2005). Examples of important semantic organization studies that used unconstrained sorting method include the work of Rosenberg and his colleagues on implicit personality theory (Rosenberg & Kim, 1975), Burton's study of occupation names (1975) and Miller's work on English nouns (1969).

Traditionally, participants were asked to sort a deck of cards on which the stimuli were written so that stimuli which appear to the participant to be similar in meaning are placed in the same pile. There is no restriction on the number of piles of cards or on the number of cards per pile. In the language of set theory, each participant, i , produces a partition, P_i , in the set, S , of stimulus elements (Burton, 1975). This study employed a web based sorting method and the software allowed participants to drag and drop stimuli items to categories, re-shuffle categories anew, delete categories, remove items from categories, verbalize the criteria or rules they used for category judgment.

In this study, research participants were given 60 food names from the unified medical language system Meta (U.S National Library of Medicine, 2006). Online sorting software was licensed for this task (www.websort.net/go/foodsorting) (Appendix F). Participants were automatically redirected to the online sorting site from a web survey site created using the popular survey monkey tool for collecting biographical information.

(<http://www.surveymonkey.com/s.aspx?sm=0BffJcH%2fgJX%2fEs2ezTIN3%2bDUGfrxtLBF%2bRzIVU3Zncc%3d>). Food names were presented randomly for each participant and no two subjects had seen items in the same order.

Description of Categories

Following the sorting task, research participants were required to verbalize or describe the criteria, rule or reason they used to sort food names into a pile. The description statement is intended to elicit humans' explanatory principles or theory upon which they based their categorization judgment. The online software licensed (websort.net) provide a separate window where participants entered their description/verbalization. Participants were encouraged to verbalize in a form of a statement where they can freely express their "theory" why certain food names should belong to a particular pile.

Classification of Semantic Relations

The classification and categorization of semantic relations into five predefined categories was the second task the second group of participants performed. Unlike the food sorting task, the sorting of semantic relations was a closed or constrained type of sorting. This study focused on the associative type of relations (Appendix B), a total of 47 relation types, which are classed under five faceted dimensions known as "Physical, Spatial, Functional, Temporal, and Conceptual," which themselves are relations (U.S National Library of Medicine, 2006).

The group that were assigned the classification and sorting of semantic relations were similarly invited by letters on which was noted a similar but different web address (www.websort.net/go/relationssort) (Appendix G). Participants were

again redirected from web survey site created using survey monkey software to collect biographical data (<http://www.surveymonkey.com/s.aspx?sm=DVKjillYvGA5ACkp4sFDfznjRAmLHO7DGH0aagHUeHo%3d>). The same software was licensed for the sorting task (websort.net) and the second group had the same interface as the first group. The only difference in this task is that participants were required to drag and drop relation types into the five categories which are defined a priori. In order to facilitate easier sorting of the semantic types, the technique of priming (Rosch, 1975) was used to provide advance information. As a result, definitions of the five families of relations were given as they appear in the UMLS knowledge sources documentation (U.S. National Library of Medicine, 2006).

Data Analysis

Corresponding to the data collection methods employed above, appropriate data analysis methods were selected. These data analysis tools were: (1) nonmetric multidimensional scaling (Kruskal, 1964a, 1964b), (2) content analysis (Weber, 1990), and (3) hierarchical clustering (Ward, 1963). The data collection and analysis methods together provided a solid foundation to the exploratory nature of this study.

Text (entire corpus) obtained as a result of the verbalization or description of the sorting task was analyzed using the method of text analysis. The sorting and classification of semantic relations data was submitted to a widely used clustering solution known as the Ward's minimum variance hierarchical clustering method. Preliminary data standardization task was first done as explained below.

Sorting Data

Sorting as a data collection method is a widely used cognitive activity that provides insight into human conceptual organization. Co-occurrence matrix of sorting data of the food names was used as input to the nonmetric MDS solution using SPSS 14.0 for windows (SPSS Inc., 2004) and to Polyanalyst™ 5.0 (Megaputer Intelligence, Inc., 2002). Although an attempt was made to investigate the role of participants' background, gender, and education, finding a collective representation of the 60 food names was the overall goal in the analysis of the sorting data (Coxon, 1999, p.55).

According to Coxon (1999, p.56), the issue of representing sorting data arises from two basic assumptions, that is, to maximize homogeneity between objects in a category and heterogeneity between categories, and that (a) all objects in the same category are considered to have a higher similarity to each other than they do to the other objects, and (b) the categories themselves are considered to be maximally distinct and separated. Although several psychological theories of concepts and categories identify border-line cases and graded structure in categories (Barsalou, 1992, p.177), sortings are treated as discrete classes, an exclusive and exhaustive set of categories, which dictates that each object must be sorted into one, and only one category (Coxon, 1999, p.55).

Standardizing Sorting Data

The notion of finding adequate representation of sorting data arises when we have disagreement on some of the partitions. Several attempts to define consensus in sorting have been investigated to explain rules that a consensus structure should have.

In order to account for individual variability in the sorting task, partitions were characterized in terms of what is known as the “height measure.” The height measure is a result of the two extreme types of sortings, the *splitter* (at one extreme, where each object/item forms in its own group, so there are as many groups as there are objects), and the *lumper* (at the other extreme, where all objects are in the same, single group), a distinction often known as fine discrimination versus gross discrimination (Burton, 1975; Coxon, 1999, p.29). These two extremes together define a continuum known as the “lumper-splitter axis” (Coxon, 1999, p.29).

According to Boorman and Arabie (1972), the height measure is defined on a scale from 0 to 1, where a partition with a height of zero has one cell for each stimulus element, and a partition with a height of one has all stimulus elements in the same cell. However, critics argue that there is more than a height measure in individual variability in the kinds of partitions they make because height only addresses structure issues (the number of categories one forms and the relative size of the groups) rather than content (the actual composition of the categories).

The results of the sorting data in this study were analyzed in terms of their height and normalized height measures. For each of the 49 research participants, the height measures were calculated using the formula $h(p) = \sum_i c_i(c_i - 1)/2$, which is the sum overall groups of the number of pairs in each of the groups (Coxon, 1999, p.30). Similarly, the normalized height measures were calculated for all of the sortings by the 49 participants using the formula:

$$\text{Normalized height} = \frac{\sum_i (c_i)(c_i - 1)/2}{p(p - 1)/2}$$

The result of the height and normalized height measures provided an overall insight into the nature of the sorting task. They were not, however, true measures to base the investigation. A powerful and advanced machine learning software known as Polyanalyst 5.0 (Megaputer Intelligence, Inc., 2002) was used to better standardize the sortings data.

Analysis of Sorting Data

After a preliminary inspection and standardization step taken above, the sorting data matrix was submitted to a spatial model known as multidimensional scaling (MDS). The Alscol procedure in SPSS (SPSS Inc., 2004) was used to obtain the MDS configuration. Although the sorting data were treated as nonmetric input, because of the ranked order of sortings in the matrix, MDS solutions provide a metric output (Hair, Black, Babin, Anderson, Tatham, 2006). The nonmetric MDS distance model is, however, selected because of its wide usage for representing and analyzing sorting data (Coxon, 1999; Kruskal & Wish, 1978). The values in the cells of the co-occurrence matrix for the 60 food names were frequency numbers indicating the number of individuals who sorted two food names together. The goodness of fit for the MDS solution is provided by a measure called stress. Kruskal's stress or objective function is given by the formula:

$$\text{Stress} = \left[\sqrt{\frac{\sum_i \sum_j [f(\delta_{ij}) - d_{ij}]^2}{\sum_i \sum_j d_{ij}^2}} \right]$$

Where d_{ij} (inter-point distance) is expressed as a function of the ranked dissimilarities $f(\delta_{ij})$ (Kruskal, 1964a, 1978). Based on Kruskal's procedure, the 60 food names (N points) are positioned in an 2-dimensional space where they were

configured to show a monotonic relationship between all pairs (d_{ij}) inter-point distances in the MDS space and ranking of the dissimilarities among all pairs of the 60 food names (the experimentally obtained). Kruskal (1964a) provides the following verbal evaluation of stress (see table 1), which is normally a “residual sum of squares” and is positive.

Table 1

Kruskal’s Goodness of Fit Measures

| Stress | Goodness of Fit |
|--------|-----------------|
| 20 % | Poor |
| 10 % | Fair |
| 5 % | Good |
| 2.5 % | Excellent |
| 0 % | Perfect |

A lower stress means a better fit, which offers a better monotonicity between the inter-point distances and the experimentally obtained rank order of pairwise dissimilarities. A “perfect” fit means there is a perfect monotone relationship between dissimilarities and the distances. The experimentally obtained dissimilarity between objects i and j is denoted by δ_{ij} . It is also assumed that the experimental procedure is inherently symmetrical, so that $\delta_{ij} = \delta_{ji}$. The self-dissimilarities are ignored, i.e., δ_{ii} . thus with n objects, there are only $n(n-1)/2$ numbers, namely δ_{ij} for $i < j$; $i = 1, \dots, n - 1$; $j = 2, \dots, n$. Another assumption is that the possibility of ties is ignored, i.e., it is assumed that no two of these $n(n-1)/2$ numbers are equal. We also left variables to remain tied when we did the analysis in SPSS.

Analysis of Category Descriptions

One important research method widely used to analyze and make valid inferences from text is content analysis (Weber, 1990, p.9). Responses from research

participants regarding the verbalization or description of categories were analyzed using the method of content analysis. During the first stage of the content analysis, the entire responses from the research participants (the Corpus) was saved as a text file organized in a form of table by each research participant. The text was then checked for spelling errors and certain transformations were made, like “carb” into “carbohydrate.”

The text file further underwent a process known as lemmatization to reduce the many words of texts and inflexions into a smaller lexicon or fewer numbers of content categories. The coding rule was selected to be word senses, phrases, and synsets. A dictionary was built based on the text analysis result of the Polyanalyst exploration engine. Important issues in content analysis such as validity (of the classification scheme) and reliability (stability, reproducibility, and accuracy of the content classification) were taken into account (Weber, 1990, p.17-18).

Reducing the initial lexicon into fewer content categories (content classification) required to establish a coding scheme where each written word from is replaced by its base form or root; singulars and plurals are grouped together, the different inflexions of the verb “to be,” for instance, are replaced by the infinitive, etc. In creating the coding scheme, the initial text classification was verified by additional English major graduate student and further changes were made (Weber, 1990, p.21-24). Two text analysis procedures were done, one for the category names and second for the actual descriptions of the categories themselves. Three exploration engines from Polyanalyst 5.0, advanced machine learning software, were used to conduct the

content analysis (Megaputer Intelligence, Inc., 2002). These exploration engines were text analysis, decision tree, and link analysis.

Analysis of Semantic Relations

Hierarchical clustering method or tree structure was used to analyze data obtained from the sorting and classification of semantic relations (Hartigan, 1967). Hierarchical clustering or tree structure is one of the most frequently used multivariate methods of classification or grouping (Hartigan, 1967; Romesburg, 1984, p.30). In this method, objects are arranged according to a sequence of successively larger superordinate categories that form a tree. The higher-level categories of relation types were predetermined for participants and they serve as the top-level superordinate categories in this study. The general assumption in hierarchical clustering is that it aims to maximize homogeneity within cluster members and heterogeneity between clusters. For this assumption to hold it is important that at each level the set of groupings is disjoint and hence non-overlapping (a partition) and objects within a given grouping are related at that level of similarity.

In the hierarchical clustering method, there are several algorithms with differing performance results. Based on the review of the extant literature (Edelbrock & McLaughlin, 1980), and based on a pilot test conducted we chose Ward's minimum variance technique (Ward, 1963) for its better performance. The pilot test involved 15 participants (SLIS graduate students, 10 women and 5 men) who were given instruction to sort the 47 semantic relations into the five categories established a priori. Data matrix obtained as a result of the sorting was analyzed using ward's minimum variance clustering solution and the result comparing the semantic relations

model in the SN of the UMLS and the human relation model (a result of the humans sorting or classification of semantic relations) were documented as a dendrogram tree (Appendix H & I).

Summary

This Chapter outlined the design specification of the study. Research participants, procedures for conducting the cognitive test, data collection methods, and data analysis methods are discussed. This study contains aspects of quantitative and qualitative methods and issues regarding both approaches were described. Three data collection methods (widely used in cognitive studies) namely, unconstrained sorting, category descriptions, and sorting/classification of semantic relations were discussed. Data obtained from the data collection methods were used as input to three data analytic techniques (appropriate in cognitive research). These data analytic techniques are nonmetric MDS, content analysis, and hierarchical clustering or tree structure (particularly the Ward's minimum variance clustering solution). The selection of data collection and analysis methods are all supported by empirical evidence and most of all are known to address human's conceptual organization and have a strong appeal to human cognition and similarity judgments.

CHAPTER 4

ANALYSIS OF DATA, RESEARCH FINDINGS, AND DISCUSSION

Introduction

The idea that humans grasp knowledge through concepts, categories, and relations is widely prevalent in the extant literature. However, how exactly humans use, represent, and organize concepts is not clear and is a source of numerous studies. In view of this thesis, the purpose of this study was to investigate the nexus between human concept cognition and concept representation in the unified medical language system (UMLS) knowledge structure. The present study is an exploratory investigation to understand human concept representation in view of an established knowledge structure known as the UMLS. Guided by a set of cognitive theories and data analytic tools, it is to determine if humans' theory/explanatory principle provides the necessary and sufficient ingredient for how humans use, represent, and organize concepts.

The tasks designed to collect data from research participants significantly involve cognitive performance and support the overall goal of this study, which is, how humans use, represent, and organize concepts (Harnad, 1987, p.1). These cognitive tasks were sorting, classifying, and verbalizing the rationale for the derived categories and are all based on the principles of discrimination and classification (Coxon, 1999). A total of 89 subjects participated in two sets of cognitive tasks. The research participants were divided into two groups. The first group was assigned the first task, which required sorting 60 food names followed by a description of each

category. Participants were instructed to sort food names into as many categories as they found it appropriate based on how they thought food names related to one another. Important tasks in the sorting also include providing a category/class name and describing or verbalizing in free statement why they think certain food names belong to a category/class. The second group was assigned the second task which required participants to sort 47 associative relation types into five categories known a priori.

Quantitative and qualitative data analysis tools were used to analyze data collected as a result of the two tasks. Non-metric MDS (Kruskal, 1964a, 194b) was used to analyze sortings data; the free text obtained from verbalization or description of categories was analyzed using the method of content analysis (Weber, 1990); and classification of the associative relation types were analyzed using the method of hierarchical clustering (Ward, 1963). SPSS 14.0 for Windows (SPSS Inc., 2004), a statistical software, and Polyanalyst™ 5.0 (Megaputer Intelligence, Inc., 2002), advanced machine-learning knowledge discovery system, were used to analyze sortings data, unstructured textual data, and classification of relation types.

This Chapter presents the result of the analysis of data from the two cognitive tasks. In view of the four research questions outlined in Chapter 1, the researcher attempts to discuss the results and findings of the study. Formalization of the research questions and profile of the research participants are presented first.

Operational Definitions

The underlying premise in this study is that the tasks research participants are asked to perform involve significant cognitive processing, for through data analytic

and visual presentation methods, it is generally possible to reveal the nature of how humans understand, use, and organize concepts. The following specific assumptions and operational definitions help to further qualify the analysis task.

- Given 60 food names (n), and 49 participants (s) sorting the n food names into as many categories as they find it appropriate, there may be any number of partitions ranging from one to n .
- Each partition/category may contain any number of food names, and with 49 (s) subjects performing the sorting, we can have z divisions of the set n into non-overlapping partitions.
- In the nonmetric MDS approach we are pursuing, relatedness between any two food names, i and j , is defined as a function of the frequency of these two food names entering the same class, according to the decisions made by all subjects who participated.
- Through the description statements of the derived categories, it is possible to reveal hidden relationships between categories and their members thereby predicting common criteria for conceptual coherence.
- Humans' theory or explanatory principle remains stable across members of a category.
- Humans' theory or explanatory principle (intension) of a category captures the commonalities that exist among members of a category (extension).
- By way of a psychological distance measure, it is possible to re-create the human conceptual representation in a multi-dimensional space.

- The inter-point distance between food names in the low-dimensional space can be accounted for by the experimental data obtained as a result of the cognitive performance.
- It is assumed that human relation model will strongly support the current classification of relation hierarchy in the SN of the UMLS.

Description of Participants

A total of 89 research participants took part in the two cognitive tasks. The criteria for participant selection have been to recruit primarily graduate and to a lesser extent undergraduate students enrolled at the University of North Texas who are also native U.S. English speakers. Attempt has been made to recruit participants from areas of study that potentially can have a moderating effect on the task at hand. The participants are equally divided into two groups to perform two different but related tasks.

Forty-nine participants sorted food names followed by the description of categories, and 40 participants performed the classification of semantic relations, a significant and acceptable subject size in both groups (Tullis & Wood, 2007). Table 2 shows the summary of the of research participants. Both groups answered similar demographic questions with the first group performing the sorting of food names having one different question that relates to any education or training in food science or culinary arts that they may have. The second group also has one different question that the first group does not, which relates to the frequency of use of medical databases.

Table 2

Summary of Research Participants Profile

| Participants profile | Food Sorting & Category Description Task | | Sorting Semantic Relations Task | |
|---|--|------|------------------------------------|------|
| | Count | %age | Count | %age |
| Gender: | | | | |
| M | 13 | 26.6 | 12 | 30 |
| F | 36 | 73.4 | 28 | 70 |
| Age Group : | | | | |
| < 21 | 0 | 0 | 2 | 5.0 |
| 21-25 | 7 | 14.3 | 11 | 27.5 |
| 26-30 | 12 | 24.5 | 5 | 12.5 |
| 31-35 | 7 | 14.3 | 7 | 17.5 |
| 36-40 | 11 | 22.4 | 10 | 25 |
| 41-45 | 6 | 12.2 | 2 | 5.0 |
| 46-50 | 2 | 4.1 | 2 | 5.0 |
| 51-55 | 1 | 2.0 | 0 | 0 |
| | 3 | 6.1 | 1 | 2.5 |
| Major: | | | | |
| Lib./Info. | | | | |
| Science | 37 | 75.5 | 30 | 75.0 |
| Hospitality Management/Food Science | 3 | 6.1 | 0 | 0 |
| Biology | 3 | 6.1 | 5 | 12.5 |
| Other | 6 | 12.2 | 5 | 12.5 |
| Highest Degree Completed: | | | | |
| Bachelors | 26 | 53.1 | 25 | 62.5 |
| Masters | 14 | 28.6 | 10 | 25.0 |
| Ph.D. | 3 | 6.1 | 0 | 0 |
| Other | 6 | 12.2 | 5 | 12.5 |
| Current Program of Study: | | | | |
| Bachelors | 0 | 0 | 0 | 0 |
| Masters | 32 | 65.3 | 33 | 82.5 |
| Ph.D. | 6 | 12.2 | 0 | 0 |
| Other | 11 | 22.4 | 7 | 17.5 |
| Education/Training in Food Science or Culinary Art | | | | |
| Yes | 10 | 20.4 | | |
| No | 39 | 79.6 | | |
| How often do you use/search medical databases | | | | |
| Daily | | | 2 | 5.0 |
| 3 times a week | | | 1 | 2.5 |
| Twice a week | | | 1 | 2.5 |
| Once a week | | | 2 | 5.0 |
| Once every two weeks | | | 10 | 25.0 |
| Once a month | | | 12 | 30.0 |
| Never | | | 12 | 30.0 |

As shown in Table 2, the majority of the research participants are female students (73% for food sorting and 70% for relation classification), which is representative of the population from which the sample is drawn. Three quarters of the participants in both groups (75.5 and 75%) major in Library and Information Sciences. Only about 20% of the participants had a food science or culinary arts background. The majority of the participants (87.7% for the first group and 87.5% for the second group) are between the ages of 21 to 40.

Analysis of Sorting Data

Standardizing Categories for Analysis

This first cognitive task is based on unconstrained sorting where research participants provide a name/label for each category they formed. Before proceeding to the actual analysis of sorting data to determine the number of partitions, the size of elements in each partition, and the composition of the elements in each category, it is important to arrive at standard category names from the different names provided by research participants.

This required a quick pre-processing or editing of the category names (labels) to check spelling, transform short forms to their standard word forms, such as “carb” to “carbohydrate” and “misc” to “miscellaneous”. After the preprocessing and text cleanup operation, text analysis is run on the category names to extract most frequent category labels. The initial total count of 546 category names were reduced to 76 text rules (see Figure 8) as a result of text analysis performed using Polyanalyst™ text analysis exploration engine (Megaputer Intelligence, Inc, 2002).

| Rule name | Rec Count | % | Description |
|--------------------|-----------|-------|---|
| vegetable | 47 | 8.608 | There are 2 possible descriptions for this node |
| fruit | 47 | 8.608 | * the ripened reproductive body of a seed plant |
| meat | 38 | 6.96 | * the flesh of animals (including fishes and birds and snails) used as food |
| dairy | 38 | 6.96 | * a farm where dairy products are produced |
| beverage | 38 | 6.96 | * any liquid suitable for drinking; "may I take your beverage order?" |
| grain | 32 | 5.861 | There are 2 possible descriptions for this node |
| condiment | 27 | 4.945 | * a preparation (a sauce or relish or spice) to enhance flavor or enjoyment; "mustard and ketchup are condiments" |
| candy | 23 | 4.212 | * a rich sweet made of flavored sugar and often combined with fruit or nuts |
| nut | 22 | 4.029 | There are 2 possible descriptions for this node |
| food | 21 | 3.846 | * any substance that can be metabolized by an organism to give energy and build tissue |
| snack | 18 | 3.297 | * a light informal meal |
| fat | 18 | 3.297 | There are 2 possible descriptions for this node |
| product | 16 | 2.93 | There are 2 possible descriptions for this node |
| sweet | 13 | 2.381 | ** unknown word ** |
| legume | 13 | 2.381 | There are 3 possible descriptions for this node |
| drink | 12 | 2.198 | There are 4 possible descriptions for this node |
| seafood | 12 | 2.198 | * edible fish (broadly including freshwater fish) or shellfish or roe etc |
| fish | 11 | 2.015 | * catch fish or shellfish |
| alcoholic beverage | 10 | 1.832 | |
| bread | 10 | 1.832 | * food made from dough of flour or meal and usually raised with yeast or baking powder and then baked |
| protein | 9 | 1.648 | * any of a large group of nitrogenous organic compounds that are essential constituents of living cells; consist of long strings of amino acids |
| pasta | 8 | 1.465 | * shaped and dried dough made from flour and water and sometimes egg |

Figure 8. Text analysis report of category names.

The result of the text analysis, the 76 text rules, were further sorted into alphabetical order and re-grouped based on meaning and synonyms of words and word senses which later resulted in 21 categories (see Table 3). The grouping of category labels is further determined by checking against category members. For example two participants used “beverage” as a category name for both alcoholic drinks and nonalcoholic drinks. Therefore, verification of members of the two categories was made before deciding to consider the 2 “beverage” counts under one label. This is important because some participants have used, say, the category name “beverage” and have {Tea, Milk} as members and other have used the same category name “beverage” to categorize {Martini, Stout, Lager, Wine} together.

After a through manual inspection, sorting data is organized by the 21 categories. As shown in Table 3, the category names in bold are selected as representative of each category. For example, beans, legumes, and vegetable protein are classed under the category label “Legume.”

Table 3

Summary of the 21 Category Names Identified

| Category/Class Name | Category/Class Name | Category/Class Name |
|---|--|--|
| 1. Alcoholic Beverage | 2. Beverage | 3. Breakfast Cereal |
| <ul style="list-style-type: none"> ▪ Alcohol ▪ Alcoholic Drinks Beverage ▪ Drinks; Types of beer | <ul style="list-style-type: none"> ▪ Drinks ▪ Non-Alcoholic beverages ▪ Non-beer drinks | <ul style="list-style-type: none"> ▪ Breakfast ▪ Breakfast Food ▪ Breakfast Bread ▪ Cereals |
| 4. Candy | 5. Condiments | 6. Dairy |
| <ul style="list-style-type: none"> ▪ Candies; Sweets ▪ Candy Bars ▪ Candy and Sweet ▪ Candy or Sugary Substance | <ul style="list-style-type: none"> ▪ Flavorings ▪ Seasoning ▪ Condiment ▪ Sweeteners | <ul style="list-style-type: none"> ▪ Dairy ▪ Dairy products ▪ Milk Base ▪ Milk Products ▪ Calcium |
| 7. Fats | 8. Fish/Seafood | 9. Fruit |
| <ul style="list-style-type: none"> ▪ Fat, Fat/Oils ▪ Oils and Fat ▪ Cooking Oils | <ul style="list-style-type: none"> ▪ Fish and Seafood ▪ Seafood ▪ Sardine Family | <ul style="list-style-type: none"> ▪ Fruit |
| 10. Fruit and Vegetables | 11. Grain | 12. Grain/Wheat Products |
| <ul style="list-style-type: none"> ▪ Fruit and Vegetables | <ul style="list-style-type: none"> ▪ Grain Family ▪ Grains and Cereals ▪ Grains and Fiber ▪ Grains and Starches ▪ Grains and Nuts | <ul style="list-style-type: none"> ▪ Grain Products ▪ Grains and Foods made from Grains ▪ Wheat Products ▪ Flour Products ▪ Baked Goods, Breads |
| 13. Junk Food | 14. Legume | 15. Meat |
| <ul style="list-style-type: none"> ▪ Junk Food ▪ Sweets/Junk Food | <ul style="list-style-type: none"> ▪ Beans ▪ Beans/Legume ▪ Vegetable Protein | <ul style="list-style-type: none"> ▪ Red Meat ▪ Meats and Proteins ▪ Animal Products; ▪ Pork Products; Poultry |
| 16. Meats and Seafood | 17. Seeds/Nuts | 18. Snack |
| <ul style="list-style-type: none"> ▪ Meats and Seafood | <ul style="list-style-type: none"> ▪ Nuts; Seeds ▪ Seeds and Seed based | <ul style="list-style-type: none"> ▪ Snack Foods ▪ Salty Snacks |
| 19. Starch | 20. Vegetable | 21. Other |
| <ul style="list-style-type: none"> ▪ Carbohydrates ▪ Carb-rich foods ▪ Pasta and Bread ▪ Pasta and Noodles | <ul style="list-style-type: none"> ▪ Roots ▪ Salad | <ul style="list-style-type: none"> ▪ Unknown ▪ I do not know ▪ Have no idea |

It is possible to transform elements so as to consolidate the number of partitions into related classes. However, due to inconsistent responses from

participants and due to borderline confusion, the partitions were treated as provided by subjects. For example, six research participants formed a class called “Fruit and Vegetables” where they combined fruits and vegetables into one category. This category could have been transformed into the categories “Fruit” and “Vegetables.” However, there is no clear-cut consensus on some items where they belong. For example, some classified “Avocado” in “Fruit” and others classified it in “Vegetable.” Likewise, “tomato” is classified by some in “Fruit” and in “Vegetable” by others. Interestingly, however, when text analysis is run for the category description, Polyanalyst text analysis exploration engine combined text entries for “Fruit and Vegetables” category into either “Fruits” or “Vegetables” depending on the synsets of the particular record. As can be seen later, this is consistent with the transformation made to “meat and seafood,” “junk food,” and “fruit and vegetable” categories due to fewer participants forming these categories (a smaller threshold value), which later reduced the number of categories to 17.

Describing Sortings Data

Before producing a co-occurrence matrix that can be serviced to the MDS solution, sorting data has been characterized in terms of the basic structure to account for individual variability in sorting. The structure of the sorting is a synthesis of the number of categories and the size of categories of each individual sorting that is calculated using height and normalized height measures. Although the aim of this study is not to explain individual differences in sorting, the height and normalized height measures of individual sortings (see Table 4) offer important evidence to the overall collective representation of the sorting data.

Table 4

Summary of Individual Sorting Data

| Sub ject | # of categories | H(p) [*] | Normalized Height ^{**} | Sub ject | # of Categories | H(p) [*] | Normalized Height ^{**} |
|-------------|--------------------|-------------------|------------------------------------|-------------|--------------------|-------------------|------------------------------------|
| 1 | 8 | 301 | 0.17 | 26 | 13 | 169 | 0.10 |
| 2 | 15 | 146 | 0.08 | 27 | 13 | 220 | 0.12 |
| 3 | 12 | 152 | 0.09 | 28 | 8 | 249 | 0.14 |
| 4 | 13 | 166 | 0.09 | 29 | 15 | 133 | 0.08 |
| 5 | 10 | 194 | 0.11 | 30 | 9 | 231 | 0.13 |
| 6 | 14 | 146 | 0.08 | 31 | 12 | 174 | 0.10 |
| 7 | 10 | 226 | 0.13 | 32 | 11 | 187 | 0.11 |
| 8 | 10 | 261 | 0.15 | 33 | 11 | 240 | 0.14 |
| 9 | 12 | 154 | 0.09 | 34 | 9 | 248 | 0.14 |
| 10 | 6 | 344 | 0.19 | 35 | 10 | 228 | 0.13 |
| 11 | 12 | 182 | 0.10 | 36 | 12 | 174 | 0.10 |
| 12 | 13 | 144 | 0.08 | 37 | 12 | 175 | 0.10 |
| 13 | 12 | 174 | 0.10 | 38 | 15 | 139 | 0.08 |
| 14 | 9 | 219 | 0.12 | 39 | 13 | 220 | 0.12 |
| 15 | 11 | 230 | 0.13 | 40 | 8 | 233 | 0.13 |
| 16 | 11 | 204 | 0.12 | 41 | 11 | 249 | 0.14 |
| 17 | 11 | 208 | 0.12 | 42 | 14 | 151 | 0.09 |
| 18 | 14 | 140 | 0.08 | 43 | 14 | 147 | 0.08 |
| 19 | 14 | 143 | 0.08 | 44 | 13 | 136 | 0.08 |
| 20 | 15 | 125 | 0.07 | 45 | 10 | 225 | 0.13 |
| 21 | 12 | 186 | 0.11 | 46 | 13 | 197 | 0.11 |
| 22 | 8 | 270 | 0.15 | 47 | 9 | 242 | 0.14 |
| 23 | 11 | 220 | 0.12 | 48 | 17 | 138 | 0.08 |
| 24 | 11 | 173 | 0.10 | 49 | 13 | 134 | 0.08 |
| 25 | 13 | 199 | 0.11 | | | | |

$$*H(p) = \text{Height of Partition} = \sum c_i(c_i - 1)/2$$

$$**\text{Normalized Height} = [\sum c_i(c_i - 1)/2] / [P(P - 1)/2]$$

As shown in Table 4, the height and normalized height measure for the overall sorting by each individual participants exhibit wide variation ranging from six categories with height and normalized height measures of 344 and 0.19 respectively to 17 categories with height and normalized height measures of 138 and 0.08, respectively. The average number of categories for all 49 participants is 11.6.

The variation in the number of partitions and size of partitions indicate the lumpers-splitter continuum between individual sortings, which is usually an indicator of gross discrimination versus fine discrimination. The higher the height measure the more one tends to bring a large number of food names into a category and vice-versa. Conversely, the more the normalized height measure is closer to zero the more the nature of the splitter type of sorting and the more the normalized height is closer to 1 the more the nature of the lumpers. The number of categories also provide a clue to the height and normalized height measure with certain exceptions. A smaller number of categories correlate to the higher height and normalized height measures.

However, the height and normalized height measures are not true indicators of collective representation of the sorting task. Subject 20 (see Table 4) has 15 categories and the height measure is 125 while normalized height is 0.07. Subject 48, on the other hand, has 17 categories with height measure of 138 and normalized height measure of 0.08. In addition, the height measures do not show at all the contents or elements of category members, which is necessary to understand the collective representation of concepts by research participants. A ranked order of sorting co-occurrence data with the number of individuals sorting two food names into a category serving as the weight is needed to overcome the limitations of the height measures. Figure 9 below presents a partial view of the co-occurrence matrix of the 60 food names where the numbers in the cells indicate the number of participants who sorted two food names into one category.

| | V1 | Apricot | Avocado | Bacon | Banana | Beans | Beef | Berries | Bread | Broccoli | Butter | Cabbage | Cashews | Cheese | Chocolate | Corn | Cracker |
|----|-----------|---------|---------|-------|--------|-------|------|---------|-------|----------|--------|---------|---------|--------|-----------|------|---------|
| 1 | Apricot | 0 | 19 | 0 | 49 | 1 | 0 | 49 | 0 | 6 | 0 | 6 | 3 | 0 | 0 | 4 | 0 |
| 2 | Avocado | 19 | 0 | 0 | 19 | 16 | 0 | 19 | 0 | 34 | 1 | 34 | 2 | 0 | 1 | 30 | 0 |
| 3 | Bacon | 0 | 0 | 0 | 0 | 6 | 47 | 0 | 1 | 0 | 1 | 0 | 6 | 2 | 0 | 0 | 0 |
| 4 | Banana | 49 | 19 | 0 | 0 | 1 | 0 | 49 | 0 | 6 | 0 | 6 | 3 | 0 | 0 | 4 | 0 |
| 5 | Beans | 1 | 16 | 6 | 1 | 0 | 6 | 1 | 2 | 23 | 0 | 23 | 8 | 1 | 0 | 23 | 0 |
| 6 | Beef | 0 | 0 | 47 | 0 | 6 | 0 | 0 | 1 | 0 | 1 | 0 | 6 | 2 | 0 | 0 | 0 |
| 7 | Berries | 49 | 19 | 0 | 49 | 1 | 0 | 0 | 0 | 6 | 0 | 6 | 3 | 0 | 0 | 4 | 0 |
| 8 | Bread | 0 | 0 | 1 | 0 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 5 | 38 |
| 9 | Broccoli | 6 | 34 | 0 | 6 | 23 | 0 | 6 | 0 | 0 | 0 | 49 | 0 | 0 | 0 | 39 | 0 |
| 10 | Butter | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 4 | 30 | 4 | 0 | 2 |
| 11 | Cabbage | 6 | 34 | 0 | 6 | 23 | 0 | 6 | 0 | 49 | 0 | 0 | 0 | 0 | 0 | 39 | 0 |
| 12 | Cashews | 3 | 2 | 6 | 3 | 8 | 6 | 3 | 0 | 0 | 4 | 0 | 0 | 1 | 8 | 0 | 7 |
| 13 | Cheese | 0 | 0 | 2 | 0 | 1 | 2 | 0 | 0 | 0 | 30 | 0 | 1 | 0 | 2 | 0 | 0 |
| 14 | Chocolate | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 8 | 2 | 0 | 0 | 4 |
| 15 | Corn | 4 | 30 | 0 | 4 | 23 | 0 | 4 | 5 | 39 | 0 | 39 | 0 | 0 | 0 | 0 | 4 |
| 16 | Cracker | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 38 | 0 | 2 | 0 | 7 | 0 | 4 | 4 | 0 |
| 17 | Ghee | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 3 | 0 | 24 | 0 | 2 | 17 | 2 | 0 | 3 |
| 18 | Grapes | 49 | 19 | 0 | 49 | 1 | 0 | 49 | 0 | 6 | 0 | 6 | 3 | 0 | 0 | 4 | 0 |
| 19 | Honey | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 10 | 0 | 2 | 3 | 13 | 1 | 3 |
| 20 | Kit kat | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 7 | 0 | 39 | 0 | 4 |
| 21 | Lager | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 0 | 4 | 0 | 2 |
| 22 | Lard | 0 | 1 | 8 | 0 | 0 | 7 | 0 | 1 | 0 | 25 | 0 | 4 | 9 | 3 | 0 | 1 |
| 23 | Lentils | 0 | 13 | 3 | 0 | 36 | 3 | 0 | 5 | 18 | 0 | 18 | 8 | 0 | 0 | 20 | 4 |
| 24 | Lettuce | 7 | 33 | 0 | 7 | 22 | 0 | 7 | 0 | 48 | 0 | 48 | 0 | 0 | 0 | 38 | 0 |
| 25 | Margarine | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 43 | 1 | 4 | 26 | 4 | 1 | 1 |
| 26 | Martini | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 4 | 0 | 1 |
| 27 | Melons | 48 | 20 | 0 | 48 | 1 | 0 | 48 | 0 | 7 | 0 | 7 | 3 | 0 | 0 | 5 | 0 |
| 28 | Milk | 0 | 0 | 4 | 0 | 6 | 4 | 0 | 15 | 1 | 0 | 1 | 2 | 0 | 1 | 8 | 11 |
| 29 | Milky way | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 23 | 0 | 1 | 41 | 3 | 0 | 0 |
| 30 | Millet | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 7 | 0 | 39 | 0 | 4 |
| 31 | Muffin | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 38 | 0 | 2 | 0 | 2 | 0 | 3 | 3 | 34 |
| 32 | Mustard | 0 | 4 | 0 | 0 | 2 | 0 | 0 | 1 | 4 | 10 | 4 | 1 | 4 | 2 | 3 | 1 |
| 33 | Noodles | 0 | 0 | 1 | 0 | 3 | 1 | 0 | 38 | 0 | 1 | 0 | 1 | 0 | 0 | 5 | 29 |
| 34 | Nougat | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 9 | 0 | 38 | 0 | 5 |

Figure 9. Co-occurrence matrix of sorting data with frequency.

To better address the content issue of a sorting task and at the same time obtain a nonmetric input for the MDS solution, the height and normalized height measures were further analyzed to produce a rank-order of co-occurrence sorting data for the 21 categories. Polyanalyst’s powerful algorithm, the decision tree exploration engine, was used to determine frequent members of a category. The decision tree algorithm of the Polyanalyst™ 5.0 (Megaputer Intelligence, Inc., 2002) is based on the principle of information gain criteria and is best suited to the problem in this study. Figure 10 Shows the result of decision tree exploration engine for the “Fruit” category where {Banana, Melons, Pineapple, Berries, Apricot, Grapes} are discovered as the most frequent food names categorized by research participants (43 in this case) to belong in the “Fruit” category. Furthermore, the decision tree

exploration engine correctly predicted category members for 12 of the 17 categories (71% consistency rate).

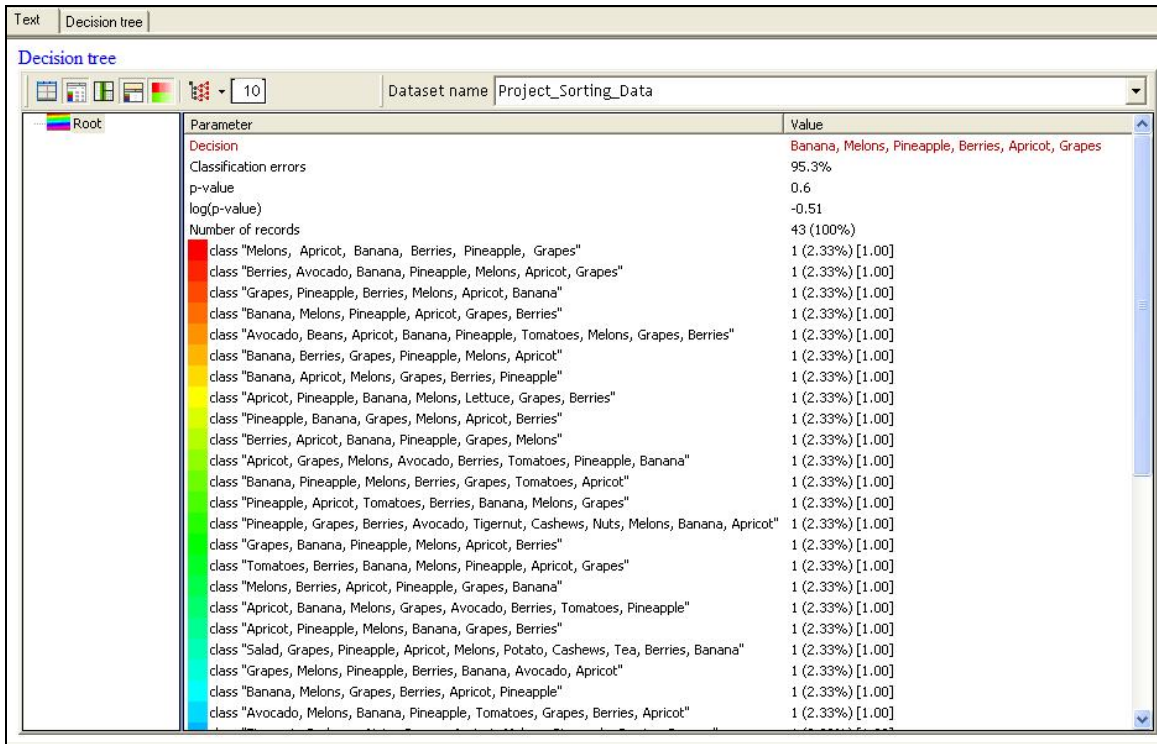


Figure 10. Decision tree diagram for “Fruit” category.

Multidimensional Representation of Sorting Data

After obtaining a nonmetric input data which consists of an ordinal rank of sorting data, SPSS 14.0 for Windows statistical software (SPSS Inc., 2004) was used to plot the 60 food names in two dimensional space. In the MDS solution, data are treated as ordinal type for the measurement purpose. The alternating least-square algorithm (ALSCAL) was used because it is suited to data elements that are dissimilarities at the ordinal level of measurement. Figure 11 presents a two dimensional configuration of 60 food names in MDS space. No missing values were reported, and the chi-square between sets of frequencies is used. The whole parameters involved in the ALSCAL procedure is attached in Appendix J.

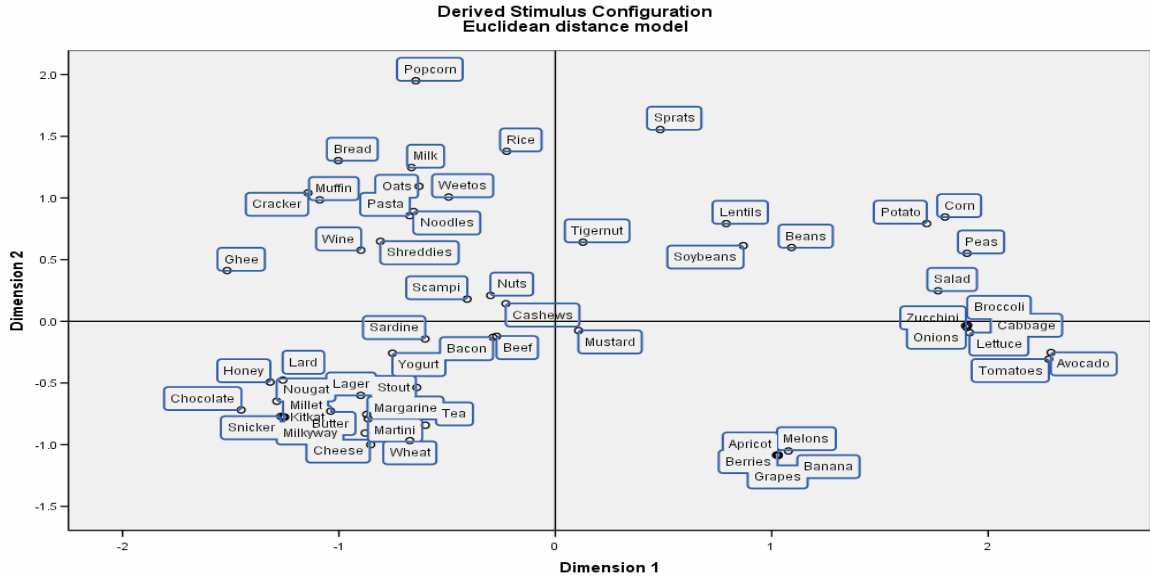


Figure 11. Two dimensional representations of 60 food names.

In the above MDS representation, Kruskal's stress statistic is used and stress and squared correlation (RSQ) of 0.12791 and 0.93092 were obtained, respectively. The stress and RSQ coefficients significantly support the variance in the MDS space because they account for the input frequency data obtained as a result of human sorting task. The scatter plot (see Figure 12) presents a better liner fit between dissimilarities and distance.

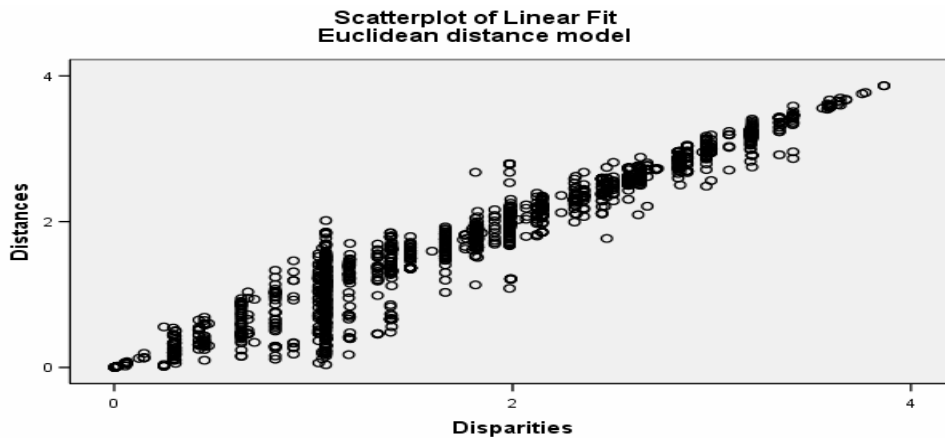


Figure 12. Scatter plot of dissimilarities (sorting data) against distance.

The scatter plot of a liner fit forms a 45 degree line, an indication of a perfect model, which in turn is indicative of a monotonic relationship between an ordinal input sorting data (dissimilarities) and distance in the MDS space. Although, the nonmetric MDS requires nonmetric input, the output is metric because the nonmetric (rank-order) output limits the interpretability of the MDS configuration (Hair, Black, Babin, Anderson, & Tatham, 2006, p. 646). The transformed scatter plot (see Figure 13) provides a better approximation of the experimental data.

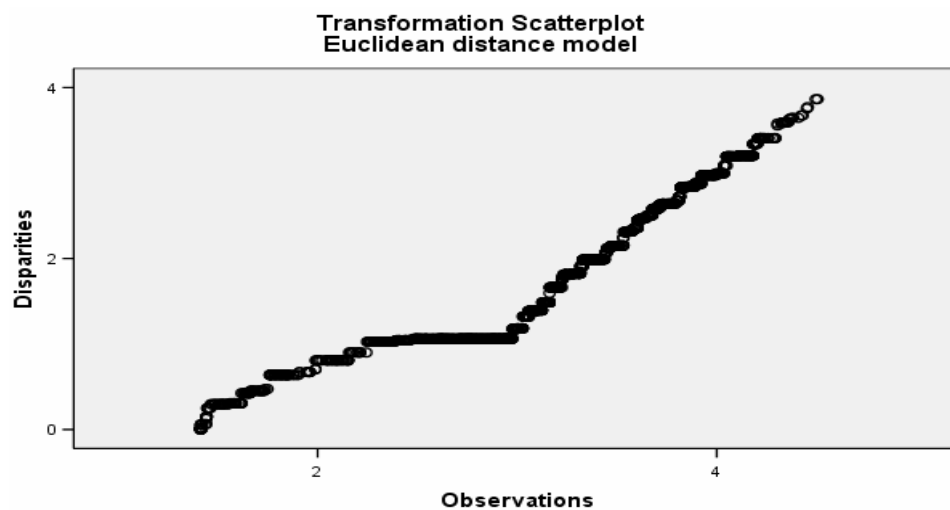


Figure 13. Transformed scatter plot of 60 food names.

The transformed scatter plot (see Figure 13) provides a better approximation of the experimental data.

In addition to a collective representation of the 60 food names above, analysis is made for individual categories to see if the MDS configuration shows some variation. The result for individual MDS representation indeed reveals a marked difference from the collective representation above. The scatter plot for the “Fruit” category is presented below for illustration purpose. From the fruit members identified by the decision tree engine (see Figure 10), “Apricot” is randomly selected

(as a representative for the “Fruit” category) to regress on the remaining 59 food names and a perfect prediction rule is discovered by Polyanalyst linear regression algorithm. Figure 14 displays a strong linear fit between food names in the fruit category.

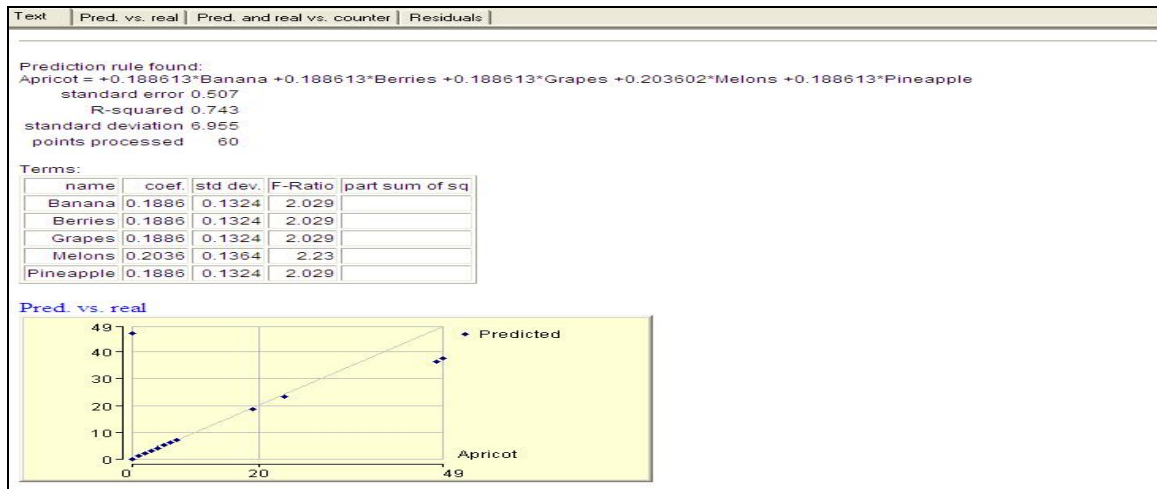


Figure 14. Predicted vs. real graph for “Fruit” category based on linear regression.

Analysis of Category Description Data

After sorting food names into categories, participants were asked to provide description of each category they make explaining the rationale for their category judgment. The task of verbalizing or describing category members resulted in free and unstructured text. In order to make valid and reliable text summarization and analysis that will furnish explication to the categories and their members, descriptive statements were first organized by the 21 categories and by each research participants. After a pre-processing of text content including spelling check, regrouping the different inflexions of a word, and coding and recoding, Polyanalyst’s text analysis exploration engine is used to perform morphological and semantic analysis.

A table of comma-delimited text file that has the 21 categories and the participants profile as columns was created. The values in the cell across the 21 categories contain edited and pre-processed text. The purpose of the text analysis is to extract normalized forms, synsets, and phrases corresponding to each of the 21 categories so as to determine if these concepts capture the commonalities across members of the categories or remain valid across members of the categories. In other words, the purpose is to investigate if humans' explanatory principle (intension) can be used as a decision rule to discriminate category members from non-members. The coding scheme was established within the framework of the theory-based approach such that it will support an explanation of a sort. The unit of recording/coding is, therefore, decided to be word senses, synsets (sets of synonyms representing a meaning of the term rather than a word form, for example "noodle and pasta," "bread and pasta" are treated in the "starch" synset), and phrases that constitute a semantic unit, such as "high carbohydrate food," "fruit-bearing plants," "ovary of a flowering plant," or "beverage" (Weber, 1990, p.21-22).

Using Polyanalyst text analysis (TA) exploration engine, a custom dictionary was first built by importing entries (word senses, synsets, and phrases) as text files. To better organize the creation of the dictionary, content items were first organized by the 21 categories. The content for each category from all participants is compiled as one unstructured text file. After the pre-processing and clean-up operation, word senses, synsets, and phrases were coded and entered into a spreadsheet as comma-delimited text file. This file was later imported as a dictionary into Polyanalyst (see Figure 15).

The text analysis exploration engine from Polyanalyst (Megaputer Intelligence, Inc., 2002) is used to conduct the content analysis. Polyanalyst offers several options to select before running the text analysis, including the selection of “find phrases” which automatically match phrases from the dataset to the predefined dictionary entries.

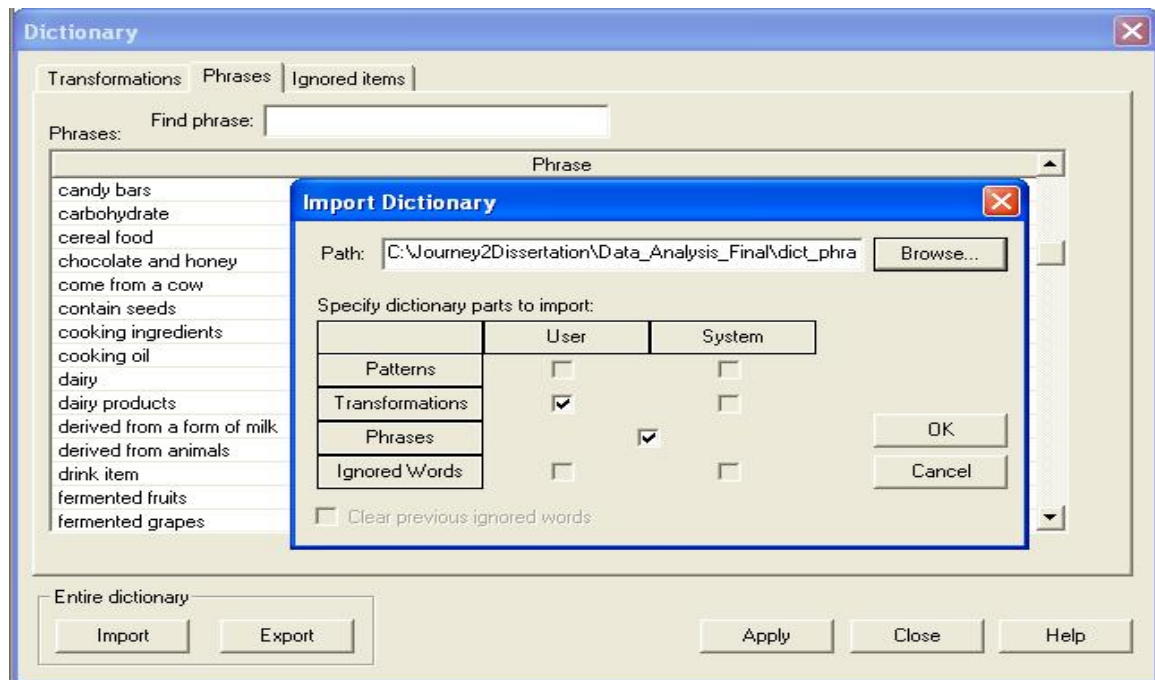


Figure 15. Dictionary building process.

The result of the text analysis report presents all normal forms for all words and predefined phrases from the corpus in each cell for the 21 categories. For every extracted normal form and synsets, Polyanalyst creates a special term-rule and sense count rules, which are used for tagging descriptions/verbalizations across the 21 categories by the corresponding terms present in each cell. These rules are basically Boolean yes/no types indicating whether a given textual attribute contains a specific term. The report of the semantic analysis for all 21 categories is too large to present

here. Figure 16 presents one example of the text analysis report for the “Alcoholic beverage” category.

| Term | Record Count | % of Records | Frequency |
|--------------------------------------|--------------|--------------|-----------|
| alcohol | 5 | 10.000000 | 5 |
| alcoholic | 1 | 2.000000 | 1 |
| alcoholic beverage | 10 | 20.000000 | 10 |
| alcoholic drink | 4 | 8.000000 | 4 |
| beverage | 13 | 26.000000 | 13 |
| category | 1 | 2.000000 | 1 |
| drink | 8 | 16.000000 | 10 |
| drink item | 1 | 2.000000 | 1 |
| fermented fruits | 1 | 2.000000 | 1 |
| good for your health | 1 | 2.000000 | 1 |
| grain | 2 | 4.000000 | 2 |
| liquids affect | 1 | 2.000000 | 1 |
| moderate amount | 1 | 2.000000 | 1 |
| of enjoyment | 1 | 2.000000 | 1 |
| red wine | 1 | 2.000000 | 1 |
| satisfy your urges | 1 | 2.000000 | 1 |
| term | 1 | 2.000000 | 1 |
| types of beer | 2 | 4.000000 | 2 |

| Rec # | Term | Top |
|-------|---|---------------------|
| 1 | Types of Beer. Alcoholic drinks. Alcohol | Top |
| 31 | Alcohol . All alcoholic drinks. | |
| 35 | Alcohol . Moderate amounts of red wine can be good for your health. | |
| 34 | Alcohol . Drinks that satisfy your urges. | |
| 46 | Alcohol . | |
| Rec # | alcoholic | Top |

Figure 16. Text analysis report for the “Alcoholic beverage” category.

The text analysis exploration engine discovered 488 term-rules for all of the 21 categories, the statistical summary of which is shown in Table 5. These text rules are later used as input for creating a visual cyclical graph with directed links where the nodes of the graph are represented by these text rules. The text rules for the 21 categories are concepts and word forms that have frequent counts from respective records for individual categories. The rule names are essentially the actual terms the text analysis engine determined. The record count is the number of records where the term appears.

In the following table (see Table 5), the summary of these text rules and the highest term/phrase/synset counts for each of the 21 categories is presented. The number of participants (# of RP) column shows how many participants created the corresponding category. Before running the text analysis, several decisions were

made that support the classification of text needed for this study. For example, the option “remove identical leaves” is unchecked to include different synsets that can be expressed by identical words.

Table 5

Statistical Summary of Text Analysis Report for the 21 Categories

| Category | # of RP* | Text rules | Highest term/phrase/synset count |
|---------------------|----------|------------|---|
| Alcoholic beverage | 46 | 18 | Beverage; Alcoholic beverage; Drink; Alcohol; Alcoholic drink; Types of beer; Grain; Fermented fruit; Fermented grain |
| Beverages | 42 | 18 | Beverage; Drink; Nonalcoholic beverage; liquid; Non-beer drinks; Tea |
| Breakfast cereal | 11 | 15 | Cereal; Breakfast food; Breakfast cereal |
| Candy | 37 | 50 | Candy; Sweet; Candy bar; Dessert; Sugar; Chocolate |
| Condiments | 40 | 38 | Condiment; Sweetener; Spice; Seasoning; Flavoring; Mustard; Honey; Natural sugar |
| Dairy | 43 | 36 | Dairy; Dairy products; Milk; Made from milk; milk base |
| Fats | 27 | 18 | Fat; Oil; Cooking ingredients; Cooking oil |
| Fish/Seafood | 21 | 23 | Fish; Seafood; Sardine family |
| Fruit | 43 | 25 | Plant that is sweet; Contain seeds; grow on trees; |
| Fruits & Vegetables | 6 | 12 | Fruit; Vegetable; Salad |
| Grains | 37 | 32 | Grain; Grain and cereal; Bread; |
| Grain products | 16 | 14 | Grain product; Baked goods; Wheat products |
| Junk food | 6 | 11 | Junk food; Not so good for you; High sugar |
| Legume | 20 | 27 | Legume; Nuts and beans; bean; beanlike things; Vegetable protein |
| Meat | 39 | 21 | Meat; Protein; Red meat; Animal; Poultry |
| Meats & seafood | 10 | 14 | Meat; Meat and fish; Meat and seafood |
| Seeds/Nuts | 23 | 10 | Nut; Seed; Seed and seed based; |
| Snack | 18 | 22 | Snack food; Salty snack; Sweet snack; in between meals |
| Starch | 34 | 27 | Pasta; Starch; Carbohydrate; Bread & pasta; Noodle; Prepared food |
| Vegetable | 43 | 40 | Vegetable; Root; All tubers; Plant |
| Other | 22 | 17 | I do not know; Miscellaneous; Have no idea |
| Total text rules | | 488 | |

* RP = Research participants who created the corresponding category

Polyanalyst provides powerful visualization tools to visually inspect the nature of relationships between concepts and normal forms identified. Using these visual tools, the approach in this study is to investigate if the theory-based approach to categorization offers coherent explanations for each categories (determined by the existence of meaningful interrelationships among the nodes in the graph). We used “Link Analysis” and “Link Chart” visualization engines in Polyanalyst to present data in a graph. Only selected graphs are presented here for illustration purposes. By way of the link analysis graph, the idea is to visually present and reveal complex patterns of correlation between individual tokens in the corpus of individual categories.

The graphs (see Figures 17, 18) display found association between the extracted normal forms from the text analysis. The web of relationships between the concepts and the type and intensity of the linkage also reveals the hidden structure, and together they explain the individual category in a coherent fashion. Participants’ biographic data are also represented in the graph to show their moderating effect.

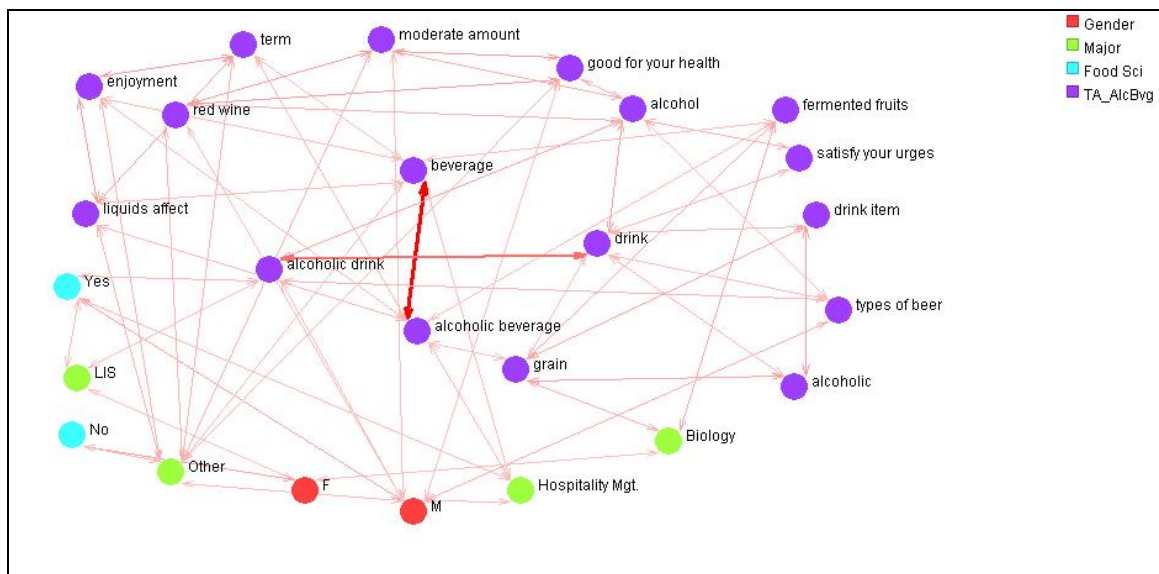


Figure 17. Link analysis graph for “Alcoholic beverage” category.

In the above graph (see Figure 17), normalized forms extracted from the unstructured text for the “Alcoholic beverage” category are shown as nodes. There are 74 link counts among the nodes. The biographical details for gender, major, and food science background are also represented as nodes to reveal the moderating effect they may have on the description task. A quick glance of the graph reveals a strong correlation (as shown by the thickness of the lines) in the center of the graph where beverage, alcoholic beverage, drink, and alcoholic drink are linked to one another. The graph allocates appropriate correlation weights to the links.

Another example shows a similar link analysis graph for the “Candy” Category (see Figure 18) having 169 link counts.

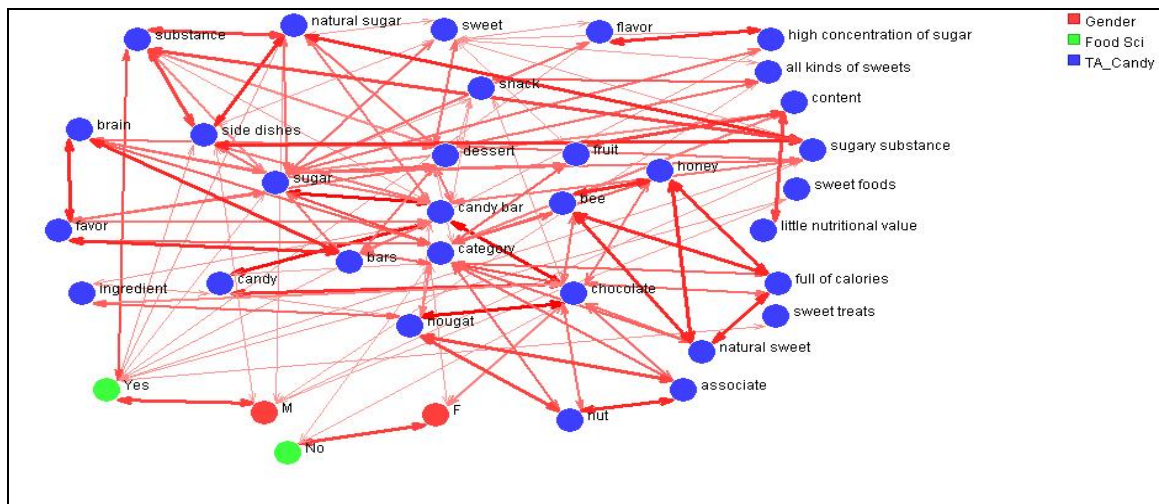


Figure 18. Link analysis graph for “Candy” category.

Link chart, another visual tool, presents the graph in terms of antecedent and consequent relationship (see Figures 19, 20). Red lines indicate positive association between values of attributes and blue line indicate negative association. The thickness and color intensity of each line reveals the strength of positive or negative correlation.

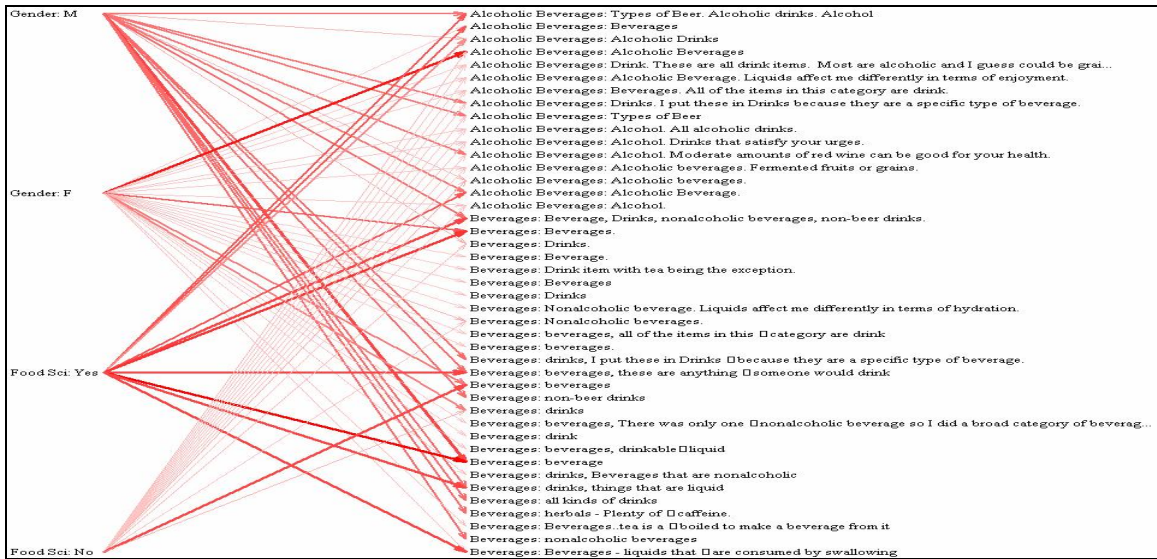


Figure 19. Link chart for “Alcoholic beverage” & “Nonalcoholic beverage” categories.

The link chart above has gender and food science background attributes as its antecedents and the descriptive statements as its consequents. The same intensity and line thickness in the link analysis graph is reproduced by the link chart (see Figure 20) for the “Candy” category combines both positive and negative correlation.

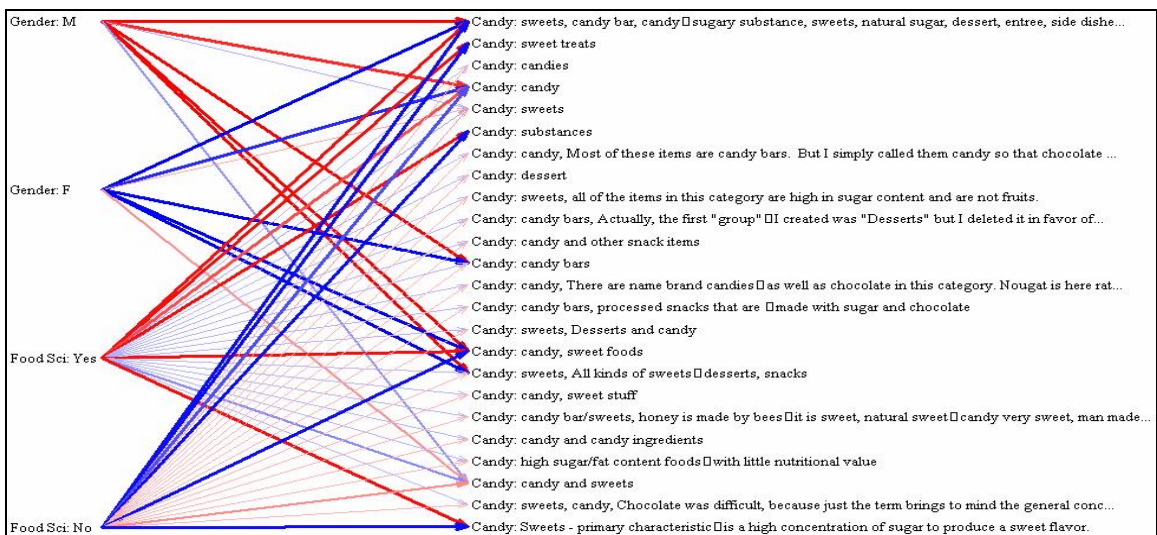


Figure 20. Link chart for Candy category with positive and negative correlation.

For a better comparison, a similar link chart as above (see Figure 20) is presented below (see Figure 21) to show the inverse relationship between the positive and negative links. The positive and negative correlations are better understood in terms of the fact that wherever the red lines (positive correlations) exist the interpretation is that the descriptions or theories are contributed by the antecedent attribute value. Conversely, wherever we have the blue lines (the negative correlations) the antecedent attribute value has no contribution in the consequent theoretical explanation. In a more specific tone, one can easily observe the positive correlation coming out of the “male” and “Food Science background – Yes” participants. This is not an isolated instance for the selected sample categories presented here. The same type of pattern persists across the 21 categories.

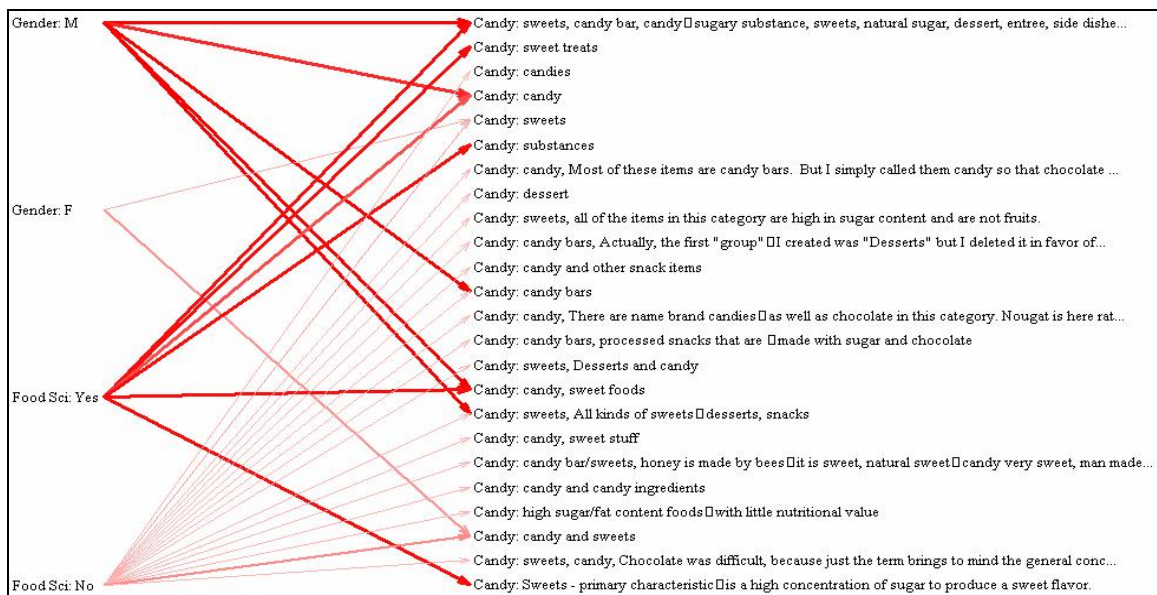


Figure 21. Link chart for the candy category (positive links only).

The color intensity and the weight of the line (observed by the heaviness of the line) from the antecedent to the consequent also represents the strength of the correlation (positive or negative) in the direction from left to right. Thicker and darker

lines show strong positive (if the lines are red) and strong negative (if the lines are blue) correlation.

Analysis of Classification of Semantic Relations

The associative relation types in the SN of the UMLS are used as stimulus materials. There are a total of 47 associative relation types in the SN of the UMLS. Organized under five classes, these semantic relations exist between semantic types (the broad categories), for example, in the physical relation (e.g., *connected_to* as in tissue *connected_to* body space or junction), temporal relation (e.g., *precedes* as in diagnostic procedure *precedes* therapeutic or preventive procedure), functional relation (e.g., *treats* as in Antibiotic *treats* Disease or Syndrome), spatial relation (e.g., *location_of* as in Tissue *location_of* Body Space or Junction), and conceptual relation (e.g. *analyzes* as in Laboratory Procedure *analyzes* Chemical).

Psychological theory of semantic relations provided the framework for the classification of semantic relations in this study. The underlying assumption is that humans have the ability (a) to judge some relations as more similar than others, (b) distinguish one relation from another, and (c) to identify instances of common relations (Chaffin & Herrmann, 1988, p.289). Although we do not have a task for subjects to identify instances of a relation, the classification/sorting of the 47 *associated_with* relation types is supported by the first two views.

The hierarchical clustering method, Ward's minimum variance hierarchical clustering method (Ward, 1963), is used to model the human relation classification data. Hierarchical clustering or tree structure is one of the multivariate methods of classification (or grouping) that is widely used (Hartigan, 1967; Romesburg, 1984,

p.30) and it is selected for its better performance (Edelbrock & McLaughlin, 1980).

The relation classification data is prepared in a matrix of 47 relations by five classes with the cells containing the number of participants who have sorted/classified the relation type in a particular class. There is no missing value and using SPSS 14.0 for Windows (SPSS Inc., 2004), data is used as input for ward's hierarchical clustering solution.

Assuming all 40 research participants would have classified the 47 relation types according to the existing structure in the UMLS SN, the result of the hierarchical clustering diagram is presented in Appendix I. This will help compare the human relation classification model with the UMLS SN relation hierarchy. The result of the clustering algorithm is shown below for each of the five classes separately for ease of presentation.

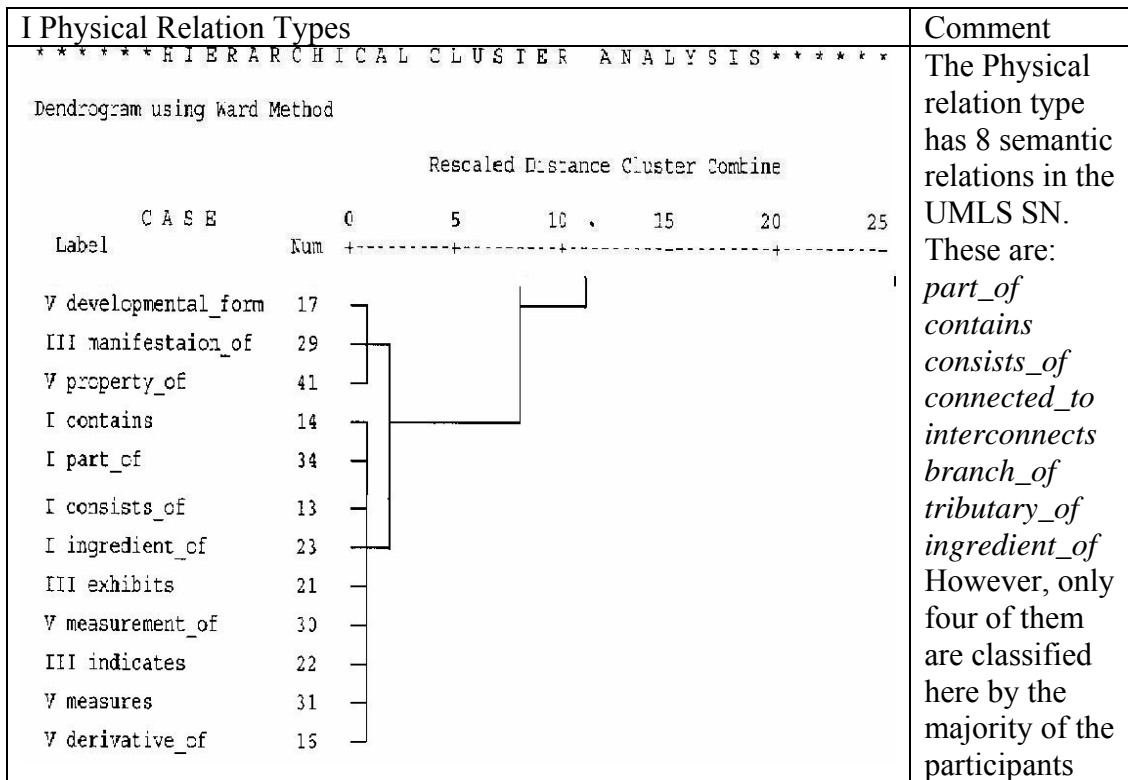


Figure 22. Hierarchical clustering of the physical relation types.

| II Spatial Relation Types | | Comment |
|--|--|---|
| <pre> *****HIERARCHICAL CLUSTER ANALYSIS***** Dendrogram using Ward Method Rescaled Distance Cluster Combine 0 5 10 15 20 25 Label CASE Num II adjacent_to 1 II location_of 27 V degree_of 15 I tributary_of 46 II traverses 44 I connected_to 12 II surrounds 43 I branch_of 5 I interconnects 25 </pre> | | <p>The Spatial relation type has 4 semantic relations in the UMLS SN. These are:</p> <p><i>location_of adjacent_to surrounds traverses</i></p> <p>All 4 are classified here with some more from other classes</p> |

Figure 23. Hierarchical clustering of the spatial relation types.

| IV Temporal Relation Types | | Comment |
|--|--|--|
| <pre> *****HIERARCHICAL CLUSTER ANALYSIS***** Dendrogram using Ward Method Rescaled Distance Cluster Combine 0 5 10 15 20 25 Label CASE Num IV cooccurs_with 9 IV precedes 37 III occurs_in 33 </pre> | | <p>There are 2 semantic relation types in the Temporal class. These are:</p> <p><i>co-occurs_with precedes</i></p> <p>Both are classified here</p> |

Figure 24. Hierarchical clustering of the temporal relation types.

| III Functional Relation Types | | Comment | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|-----|---------|-----|----|----|----|----|----|----|-------|-----|--|--|--|--|--|--|----------------------|---|--|--|--|--|--|--|--------------|----|----------|----|------------|---|-----------------|---|--------------|----|---------------|----|-------------|----|-------------|----|-------------|---|--------------------|----|--------------|----|--------------|----|---------------|----|------------------|---|----------------|----|------------|----|-----------------|----|-----------------|----|-------------|----|--|
| <p>***** HIERARCHICAL CLUSTER ANALYSIS *****</p> <p>Dendrogram using Ward Method</p> <p style="text-align: center;">Rescaled Distance Cluster Combine</p> <table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">C A S E</th> <th style="text-align: left;">Num</th> <th style="text-align: center;">0</th> <th style="text-align: center;">5</th> <th style="text-align: center;">10</th> <th style="text-align: center;">15</th> <th style="text-align: center;">20</th> <th style="text-align: center;">25</th> </tr> </thead> <tbody> <tr> <td>Label</td> <td>Num</td> <td colspan="6" style="border-top: 1px dashed black;"></td> </tr> <tr> <td>V assesses_effect_of</td> <td>4</td> <td colspan="6" rowspan="20" style="vertical-align: middle; text-align: center;"> </td> </tr> <tr> <td>III produces</td> <td>40</td> </tr> <tr> <td>III uses</td> <td>47</td> </tr> <tr> <td>III causes</td> <td>8</td> </tr> <tr> <td>III carries_out</td> <td>7</td> </tr> <tr> <td>III disrupts</td> <td>19</td> </tr> <tr> <td>III result_of</td> <td>42</td> </tr> <tr> <td>V diagnoses</td> <td>18</td> </tr> <tr> <td>III manages</td> <td>28</td> </tr> <tr> <td>III affects</td> <td>2</td> </tr> <tr> <td>III interacts_with</td> <td>24</td> </tr> <tr> <td>III prevents</td> <td>38</td> </tr> <tr> <td>III performs</td> <td>35</td> </tr> <tr> <td>III practices</td> <td>36</td> </tr> <tr> <td>III brings_about</td> <td>6</td> </tr> <tr> <td>III process_of</td> <td>39</td> </tr> <tr> <td>III treats</td> <td>45</td> </tr> <tr> <td>III complicates</td> <td>10</td> </tr> <tr> <td>V evaluation_of</td> <td>20</td> </tr> <tr> <td>V method_of</td> <td>32</td> </tr> </tbody> </table> | | C A S E | Num | 0 | 5 | 10 | 15 | 20 | 25 | Label | Num | | | | | | | V assesses_effect_of | 4 | | | | | | | III produces | 40 | III uses | 47 | III causes | 8 | III carries_out | 7 | III disrupts | 19 | III result_of | 42 | V diagnoses | 18 | III manages | 28 | III affects | 2 | III interacts_with | 24 | III prevents | 38 | III performs | 35 | III practices | 36 | III brings_about | 6 | III process_of | 39 | III treats | 45 | III complicates | 10 | V evaluation_of | 20 | V method_of | 32 | <p>The Functional relation type has 20 semantic relations in the UMLS SN.</p> <p>These are:</p> <p><i>manifestation_of</i></p> <p><i>affects</i></p> <p><i>interacts_with</i></p> <p><i>disrupts</i></p> <p><i>prevents</i></p> <p><i>complicates</i></p> <p><i>manages</i></p> <p><i>treats</i></p> <p><i>occurs_in</i></p> <p><i>process_of</i></p> <p><i>uses</i></p> <p><i>indicates</i></p> <p><i>result_of</i></p> <p><i>brings_about</i></p> <p><i>produces</i></p> <p><i>causes</i></p> <p><i>performs</i></p> <p><i>carries_out</i></p> <p><i>practices</i></p> <p><i>exhibits</i></p> <p>16 of them are classified here by research subjects</p> |
| C A S E | Num | 0 | 5 | 10 | 15 | 20 | 25 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Label | Num | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| V assesses_effect_of | 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III produces | 40 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III uses | 47 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III causes | 8 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III carries_out | 7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III disrupts | 19 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III result_of | 42 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| V diagnoses | 18 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III manages | 28 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III affects | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III interacts_with | 24 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III prevents | 38 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III performs | 35 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III practices | 36 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III brings_about | 6 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III process_of | 39 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III treats | 45 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| III complicates | 10 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| V evaluation_of | 20 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| V method_of | 32 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Figure 25. Hierarchical clustering of the functional relation types.

| V Conceptual Relation Types | | Comment |
|---|--|--|
| <p>***** HIERARCHICAL CLUSTER ANALYSIS *****</p> <p>Dendrogram using Ward Method</p> <p>Rescaled Distance Cluster Combine</p> | | <p>Conceptual relation class has 13 semantic relations in the UMLS SN. These are:</p> <p><i>property_of</i> <i>conceptual_part_of</i> <i>evaluation_of</i> <i>measures</i> <i>diagnoses</i> <i>issue_in</i> <i>derivative_of</i> <i>developmental_form_of</i> <i>degree_of</i> <i>measurement_of</i> <i>method_of</i> <i>analyzes</i> <i>assesses_effect_of</i></p> <p>Only 3 are classified here.</p> |

Figure 26. Hierarchical clustering of the conceptual relation types.

Research Findings and Discussion

This study is entirely based on cognitive tasks. It is an exploratory investigation, whose goal is to reveal human conceptual representation with a view to compare human cognitive map with the UMLS knowledge structure. As outlined in Chapter 3, important consideration has been rendered to select appropriate data analytic tools that will offer a better formalization of the datasets obtained as a result of the cognitive performance. In an exploratory study of this nature, selecting the appropriate data collection and analysis method will help to interpret results and findings. In order to enhance the quality and validity of the input data, datasets that resulted from the cognitive performance were first subjected to prior inspection and data screening before the analysis is started. Based on multivariate analysis, a

summary of descriptive statistics is generated for the 60 food names and 47 relation types.

The explore procedure in SPSS is used to check for missing data, outliers, and extreme values. No missing data for either food name sorting and relation type classification were reported. Because the sorting data for the food names (unconstrained or free-sorting technique) were based on frequency counts (ordinal type), the statistic for skewness and other normality measures did not reflect the actual nature of the distribution. For obvious reason, all fruit names will have the highest frequency with one another and should have a very small or zero value when compared with other food names. For example, SPSS reports the highest five extreme cases to “Apricot” as {Banana, Berries, Grapes, Pineapple, Melons} with 49 frequency value and reports the lowest 5 extreme cases to “Apricot” as {Wine, Weetos, Twix, Stout, Soybeans} with zero frequency value. SPSS treats such high and low frequency values as extreme values, and does not contradict the type of data we have.

The relation sorting is based on a closed sorting technique where participants had only the option of sorting the 47 relations into one of the five classes. As a result, descriptive summary statistics are generated using the explore option in SPSS. As shown in Table 6, the descriptive statistic for the semantic relation classification indicates a normal distribution. The other data set, the free and unstructured text from category description task, is thoroughly inspected for different word forms, and the coding by the principal investigator is verified by a graduate English student for consistency and validity. In addition, definitions of the 21 categories are consulted

from the UMLS knowledge server to arrive at the final text rules or normal forms to be used for the text analysis.

Table 6

Summary of Descriptive Statistics for Semantic Relation Classification

| Category Type | N | | | | |
|--------------------|-----------|-------|--------|----------|----------|
| | Statistic | Mean | SD | Skewness | Kurtosis |
| Conceptual | 47 | 6.83 | 7.185 | 2.134 | 6.155 |
| Functional | 47 | 13.51 | 10.217 | 0.256 | -1.300 |
| Physical | 47 | 8.36 | 6.569 | 0.933 | 0.281 |
| Spatial | 47 | 6.28 | 8.032 | 1.493 | 1.609 |
| Temporal | 47 | 5.02 | 6.509 | 2.814 | 8.776 |
| Valid N (listwise) | 47 | | | | |

The boxplot data for the semantic relation (see Figure 20) visually presents similar descriptive statistics where the extreme values and outliers are shown for all of the relation classes except the “functional” class. The outliers shown for the classes

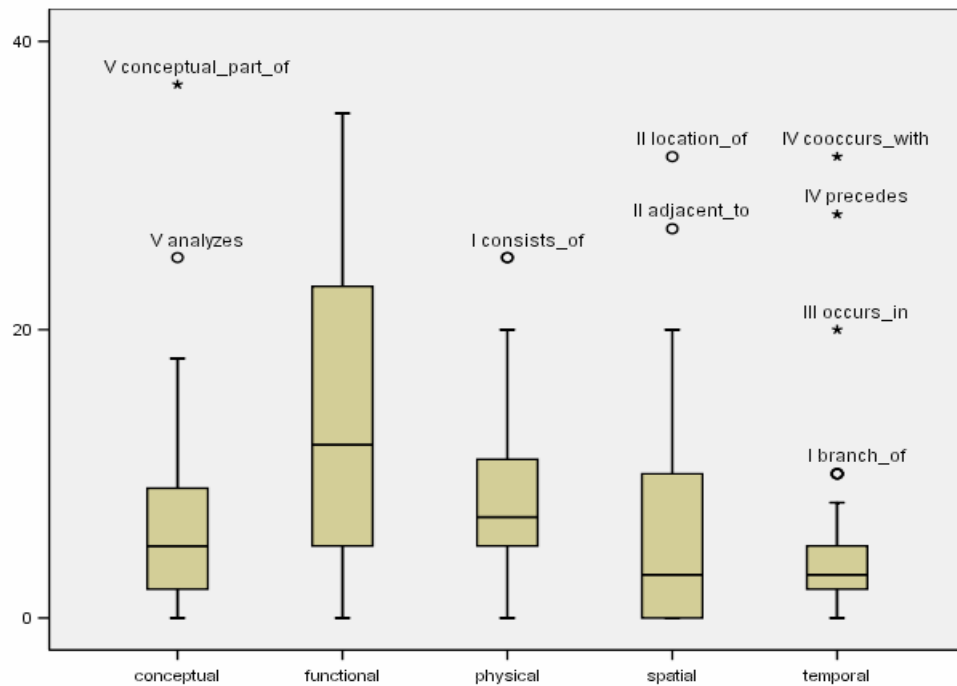


Figure 27. Boxplot data summarizing relation types by the 5 classes.

conceptual, physical, and spatial are treated as outliers because they have higher number of frequencies. Otherwise, the outliers belong in the classes where they are treated as outliers. The same is true for the two IV's (extreme values) shown for the temporal class. The relation type "III occurs_in" is treated as extreme case in the temporal class because subjects consider "occurs_in" to be more of a temporal dimension than a functional one.

Categorization Basis and Coherence Criteria (RQ1)

The first research question (RQ1) seeks to find a plausible explanation about the most frequent criteria people use in the categorization task. RQ1 is thus stated in Chapter 1 as: What are the common coherence criteria (categorization basis) humans' use in categorization? In Chapter 2, the extant literature that offers explanation on what is a comprehensible class is reviewed and discussed, including the classical theory (Medin, 1989; Smith & Medin, 1981), the prototype (Hampton, 1993; Rosch & Mervis, 1975), exemplar (Medin & Schaffer, 1978; Medin & Smith, 1984; Smith & Medin, 1981), the two-tiered approach (Michalski, 1989; 1993), and the theory-based approach (Medin, 1989; Murphy & Medin, 1985). In this study, the theory-based approach to categorization guides the exploratory investigation because it is believed will offer a richer better account to the complex human conceptual representation system.

One significant role of the theory-based approach in cognition and its explanation in conceptual coherence is its property of "explanations" of a sort, specified over some domain of observation (Murphy & Medin, 1985). By allowing participants sort food names freely followed by verbalizing the derived categories,

this study revealed that humans commonly use naïve explanations and simplified rules, largely similarity-based feature representation when they create categories. For example, for the “alcoholic beverage” category, not a single participant mentioned the word “ethanol” which is commonly used by the three vocabulary sources in the UMLS that defined alcoholic beverage: the MeSH (Medical Subject Headings), the CRISP thesaurus (Computer Retrieval of Information on Scientific projects), and the NCI thesaurus (National Cancer Institute), (U.S. National Library of Medicine, 2006).

The “explanations” of a sort or the simplified rules are evident in the subjects’ statements like “food that swims” when they describe the “fish” category; “not good for you” for the “junk food” category; “plant that is sweet” for the “fruit” category; or “vegetable protein” for the “legume” category. Though they may lack the technical terminologies, for instance, “polysaccharides” to describe the “starch” category, or “ethanol” to describe the “alcoholic beverages,” participants’ have their own way of simplified theory within which their knowledge is embodied.

However, despite their simplified and naïve theories, human in general can form a comprehensible class that contains related items in it. As shown in Table 7, participants, to a large extent, form consistent and coherent categories, identify common members of the categories, and describe the derived categories adequately well. The analysis of sorting data and the analysis of text together provide significant evidence to the above assertion. As shown in Table 7, the decision tree exploration engine (an algorithm that helps solve the problem of classifying cases into multiple categories) identified the most frequent members of the category in 12 out of the 21 categories (57% consistency rate).

It is also important to note that the fact there is a category labeled as “other” for the “I do not know” type of food names created by 22 of the research participants (about 45% of the participants) indicate the role of language in the task of categorization. Even when research participants were selected to be native English speakers, this has remained to be a problem.

Table 7

Decision Tree Report for each Category

| Category | Decision | p-value | Log (p-value) |
|----------------------|---|---------|---------------|
| Alcoholic beverage | Lager, Martini, Stout, Wine | 0.575 | -0.553 |
| Beverage | Tea | 0.574 | -0.541 |
| Breakfast cereal | Shreddies, Weetos | 0.616 | -0.485 |
| Candy | NA | 1 | 0 |
| Condiment | Honey, Mustard | 0.556 | -0.586 |
| Dairy | NA | 1 | 0 |
| Fats | Lard | 0.577 | -0.475 |
| Fish | Sardine, Scampi | 0.586 | -0.534 |
| Fruit | Apricot, Banana, Berries, Grapes, Melons, Pineapple | 0.6 | -0.51 |
| Fruit & vegetables | NA | 1 | 0 |
| Grain | NA | 1 | 0 |
| Grain/Wheat products | Muffin, Cracker, Bread | 0.612 | |
| Junk food | NA | 1 | 0 |
| Legume | Beans, lentils, Peas, Soybeans | 0.611 | -0.493 |
| Meat | Bacon, Beef | 0.558 | -0.584 |
| Meat & seafood | NA | 1 | 0 |
| Seeds | Cashews, Nuts, Tigernut | 0.577 | -0.549 |
| Snack | NA | 1 | 0 |
| Starch | Noodle, Pasta | 0.58 | -0.52 |
| Vegetable | NA | 1 | 0 |
| Other | NA | 1 | 0 |

The p-value for all discovered categories of related food names is a probability that the detected joint occurrence of food names in their respective category is a mere statistical fluctuation (accidental). However, the interpretation is

that as the numbers get closer to zero, it is an indication that the correctness of the relatedness of the food names in a category is more credible. NA's are the categories for which the decision tree exploration engine could not discover a decision rule. Due to lower threshold value (fewer participants sorting), members of the "fruits and vegetable" category were transformed into "fruit" and "vegetable" and members of "meat and seafood" into "meat" and "fish/seafood." Moreover, the "junk food" category is further re-classified for the same reason (only 6 participants creating this category) and the "other" category is excluded from further analysis. The transformation process reduced the number of categories to 17, the rate with which humans consistently classify food items increases to 70% (12 categories out of 17). It is a normal procedure in categorization task to introduce a threshold value beyond which to treat items as "noise" (Frumkina & Mikhejev, 1996, p.87).

In order to display the exact value, the base 10 logarithm of the p-value is shown (see Table 7), since the probability of an accidental association of food names is always < 1 , the respective logarithms are negative. The relatedness of food names discovered by the decision tree engine are further verified in the UMLS knowledge structure and that also supports the human cognitive performance of the sorting task.

Concept as a Decision Rule (RQ2)

The second research question is concerned with finding answers for the characterization issue in the study of categories (Feger & De Boeck, p.204). As stated in Chapter 1, the research question is "What is the role of human theory/explanatory principle (intension) in discriminating category members (extension)?" The text analysis of the verbalization/description task reveals a significant association between

humans' explanatory principles (their theories) and the category members. As shown in Table 5, the normalized forms extracted from the text analysis for the category "Alcoholic beverage," that has members {Lager, Martini, Stout, Wine} include beverage, alcoholic beverage, drink, alcohol, alcoholic drink, types of beer, fermented grain, and fermented fruit (the highest synset counts).

The link analysis graph (see Figure 17), also reveals a strong relationship between the synsets beverage, alcoholic beverage, drink, and alcoholic drink for the "alcoholic beverage" category. This is shown by the thicker and heavier lines between these feature representations in the graph. The link analysis graph (see Figure 17) and the link chart (see Figure 19), together reveal another important discovery. The participants variables (gender, major, and food science background) are allowed to be represented in the graph together with the concept representation and it can easily be seen that the source of the thick and heavier lines (an indication for a stronger and positive relationship) is from those who have "Yes" for "Food Science Background" and whose major is "Biology" and "Hospitality Management" as their major field of study.

One more additional link analysis graph and link chart for "Condiment" and "Fish" categories (see Figures 28, 29) reveal clearly the nature/magnitude of the theory/intension individuals provide when describing categories. The Biographic details, which are, "major program of study" and "food science background," are represented in the graphs and that again show the higher proportion of description coming from participants whose major is biology and hospitality management and from those who have said "yes" for food science background. These are interesting

findings given that the numbers of participants from biology, hospitality management are very small, just 12%, compared to participants from the Library and Information Science program, who comprise about 76% of the participants; or compared to those who have indicated as having background in food science number just 20%.

The link chart for “Fish” category (see Figure 29) also reveals the same phenomena in that the bulk of the description/verbalization is from those whose highest degree is Ph.D., Masters, “Other,” and from those who have said “Yes” for food science background. Those who said “No” for food science barely contributed to the description statements for the “Fish” category (as illustrated by the sparse lines). The major finding in this section of the data analysis is that participants indeed use their own explanatory principles/theories to discriminate category members, and these explanatory principles, represented as nodes in the graph, perfectly form a coherent structure.

The link analysis exploration engine of Polyanalyst system visually presents complex patterns of correlation between the nodes in the graph (represented here by the normalized forms of the theories) and it can be in general said these concepts and descriptions/intension are largely shared by the category members of the category “condiment” (for example), i.e., {honey, mustard} (identified by the decision rule, see Table7).

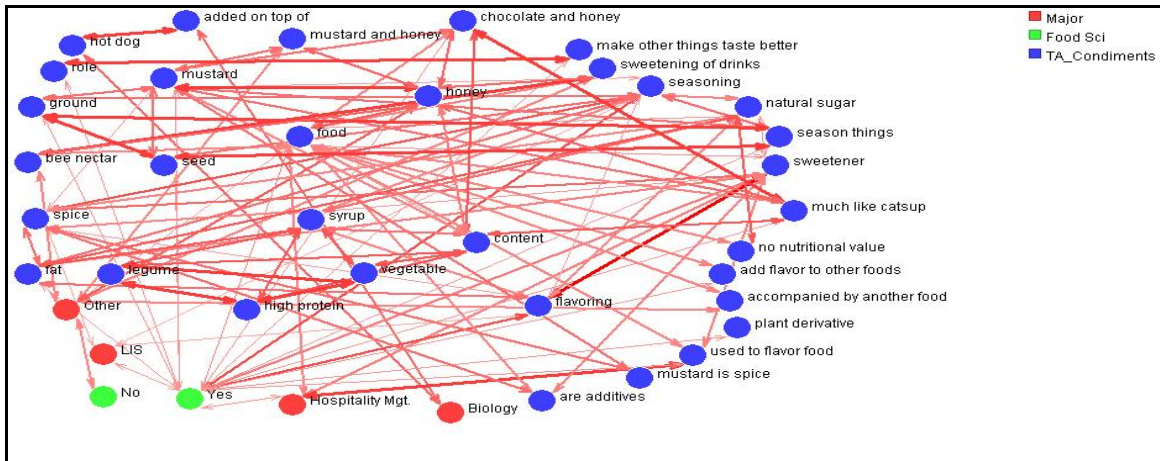


Figure 28. Link analysis graph for “Condiment” category.

All the blue nodes are the normal forms (synsets and conceptual units) resulting from the text analysis. These normal forms are treated as term rules in the convention of Polyanalyst. We generated Boolean (yes/no) and sense-count rules, which we later applied them to the dataset that has the complete text description. The rules are in turn used to tokenize, or tag, the text for each record and each category with patterns of encountered terms, phrases, and synsets as defined in the dictionary. The whole link analysis and link chart graphs are based on the algorithm known as symbolic rule language (SRL).

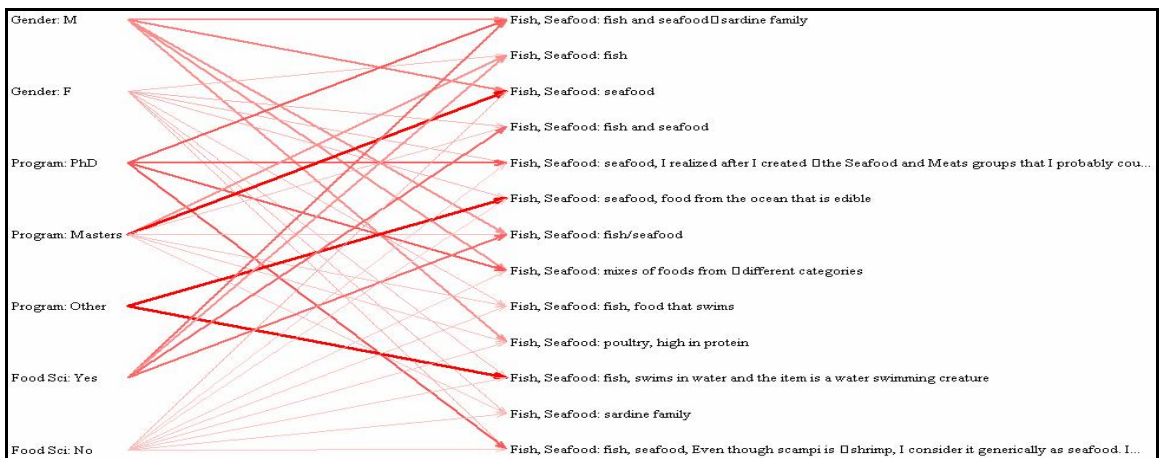


Figure 29. Link chart for the “Fish” category.

The Relationship between Human Cognitive Map and Knowledge Structure (RQ3)

The third question is “To what extent does human concept representation match a sample of concept representation in the unified medical language system (UMLS)?” There is strong evidence that the task of sorting is closely related to the cognitive process and can be used to obtain judgmental data about semantic organization (Burton, 1975; Harloff, 2005; van der Kloot & van Herk, 1991). The underlying rationale behind the task of sorting is that individuals have a map-like representation of stimulus items in a given domain of knowledge and that they use the distances between the stimulus items in this map to generate their own sortings (Kruskal, 1964a; van der Kloot & van Herk, 1991).

Based on this thesis, this study asked research participants to sort 60 food names into categories based on how they think food names relate to one another. The nonmetric MDS solution (Kruskal, 1964a, 1964b) is used to analyze the result of the sortings data. Using SPSS 14.0 for windows (SPSS Inc., 2004), the ALSCAL (Alternating Least Square Approach in Scaling) program is used to represent humans’ sorting data in two-dimensional space. The data are treated as ordinal for the level of measurement because the values in the co-occurrence matrix are frequency values showing how many individuals put two food names together. The complete summary of the ALSCAL procedure is presented in Appendix J.

Kruskal’s stress measure and the squared correlation (RSQ) coefficient of 0.12791 and 0.93092, respectively, are obtained for the MDS configuration of the 60 food names in a low-dimensional space. The stress measure is an indication of how well the configuration in the MDS space matches the experimental data obtained from

the input sorting data. A stress coefficient of 0.12791 or 12% is a little over the “fair” goodness-of-fit level of measurement (which is 10%) (Kruskal, 1964a). The RSQ coefficient of 0.93092 indicates that about 93% of the variance in the MDS space configuration is accounted for by the input frequency of ordinal data obtained as a result of human sortings. The linear fit scatter plot and the transformed scatter plot (see Figures 12 and 13) also show a modest monotonic relationship between dissimilarities in the sortings data and distance in the MDS space.

The PROXSCAL (proximity scaling) procedure is also used in SPSS 14.0 for Windows to verify the result obtained from the ALSCAL procedure. A similar result is obtained with Stress-I of 0.12767 and normalized raw stress of 0.01630 (because PROXSCAL minimizes the normalized raw stress). The common space points plot the 60 food names in MDS space (see Figure 30) and is closer to the result from the Alscal procedure output shown in Figure 4.

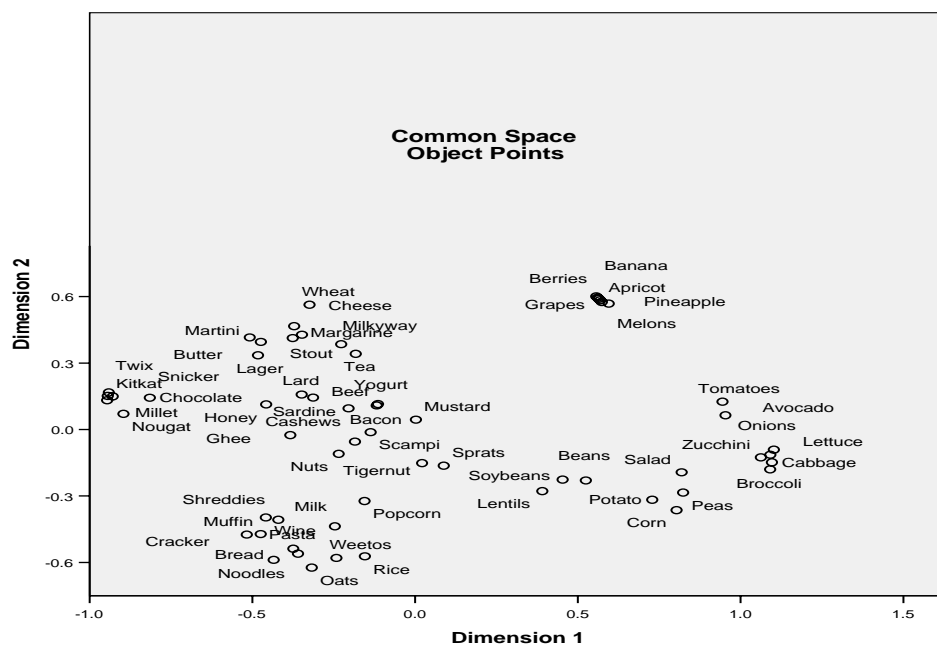


Figure 30. Common space points for the 60 food names.

The Relationship between Human Relation Classification and Relation Structure in
the Semantic Network of the UMLS (RQ4)

While the discussion of concepts, categories and their representation is important to understand how humans use, represent, and organize concept, their discussion without semantic relation is incomplete. The fourth research question aims to gain human cognitive view of the relation classification to find a desirable correspondence between humans' relation classification and the existing relation structure in the UMLS SN. The fourth research question is "What is the relationship between human relation classification and relation structure in the semantic network (SN) of the unified medical language system (UMLS)?"

As shown in Figures 22 through 26, of a total of 47 semantic relations, research participants classified only 29 relations (about 62%) according to UMLS structure. The result shows a remarkable difference from one class to another. In the "Functional" relation category, there are 20 semantic relations and research participants classified 16 of them (80%) consistently and according to the UMLS structure (see Figure 25). The least consistent class where participants poorly performed is in the "Conceptual" relation type category. The conceptual relation class has 13 semantic relations, of which subjects classified only 3 (23%) accurately (see Figure 26). The result for the "Temporal" (time dimension) class also reveals important discovery to consider. While participants classified both "co-occurs_with" and "precedes" correctly in this class (as it exactly exists in the UMLS SN), they also added "occurs_in" in this class (which was in the functional class originally).

The result of the semantic relation classification suggests the nature of spatio-temporal and pragmatic dimensions humans tend to respond or associate with fairly well. At the same time, the result is indicative of the difficulty individuals have with the conceptual type of semantic relations. Although there is no complete inventory of semantic relation types, the most common relation types such as the part-whole, similars, contrasts, queuing, case-relations, and the lexical-semantic relations may provide a better coherence to the knowledge structure (Burgun & Bodenreider, 2001; Chaffin & Herrmann, 1988; Chaffin & Herrmann, 1987). If we are to construct a richer knowledge representation system, the relation types are required to exhaustively respond to the fundamentals of human cognition and conceptual representation (Markowitz, Nutter, & Evens, 1992).

Summary

This Chapter began with an introductory remark and operational definition of the four research questions. The operational definitions were intended to better qualify the research questions so as to guide the analysis of data and presentation of results. Due to the exploratory nature of this study, the operational assumptions provided direction to the overall conduct of the analysis task. We also presented the profile of the research participants by age, gender, current program of study, highest degree earned, and background related to food science.

The necessary procedures to inspect the data collected for its completeness and normality are outlined. Corresponding to the three sets of data - sorting of food names, verbalization text, and sorting of semantic relations - analysis steps, results, and findings were discussed. The findings were discussed for each of the four

research questions. The results were presented using the data analytic tools employed: (a) representation of food names in a low-dimensional space; (b) representation of conceptual units of text using link analysis graph and link chart; and (c) the presentation of semantic relations using the hierarchical tree (dendrogram). Together with the presentation of the results of the analysis, important interpretations were also made that potentially explain clear findings.

CHAPTER 5

SUMMARY AND CONCLUSION

Introduction

This study may have ventured on a complex subject. The connection between how humans represent concepts inside their heads and how this compares to an established knowledge structure has, however, far-reaching implications in how knowledge representation can be improved in information retrieval systems. In a small way, this study aimed to employ sets of related cognitive theories and data analytic tools to shed light on how humans represent concepts and relations. The unified medical language system (UMLS) is a well established knowledge structure containing millions of concepts and relations between the concepts. By taking samples of concepts and relations from this established knowledge structure, this study involved a total of 89 research participants to perform two sets of cognitive tasks.

Four research questions were carefully crafted to address the overall objective of this study. These four questions in general address important issues in human concept representation. In other words, the issues concern the question of aggregation and characterization. Several of the cognitive theories that focus on information processing as a subject (Carpineto & Romano, 2004; Feger & De Boeck, 1993; Priss, 2006; Quillian, 1968) deal with issues of aggregation and characterization by asking subjects how they conceptualize a given domain (categorization, classification) and

by eliciting descriptions of the given objects in terms of perceived attributes or theoretical explanations.

Following a similar approach, using the theory-based approach to concept representation (Keil, 1989; Medin, 1989; Murphy & Medin, 1989) and the psychological theory of semantic relations (Chaffin & Herrmann, 1987; Chaffin & Herrmann, 1988) as a general framework, this study aimed to explore and find answers to four related issues which can be summarized as: (a) coherence criteria in categorization, (b) concept as a decision rule, (c) the relationship between human cognitive map and knowledge structure in the UMLS, and (d) the relationship between human relation classification and relation structure in the SN of the UMLS.

Divided into two groups, a total of 89 participants took part in two cognitive tasks. The first group of 49 participants sorted 60 food names into categories followed by simultaneous verbalization of the derived categories. The second group of 40 participants participated in the classification of the associative type semantic relations into five categories. As a result of the cognitive tasks, three datasets were obtained: food-sorting data, category-description text, and relation-classification/sorting data.

Corresponding to the three datasets, appropriate data-analytic tools were used for analysis after the proper data screening and pre-processing was carried out. The food-sorting data were analyzed using Kruskal's nonmetric MDS solution in SPSS 14.0 for windows (SPSS Inc., 2004). The category-description text was analyzed using Polyanlyst™ 5.0, advanced machine learning system, (Megaputer Intelligence, Inc., 2002), and the relation-classification data were analyzed using Ward's minimum variance hierarchical clustering method (Hartigan, 1967; Romesburg, 1984, p.30;

Ward, 1963) using again SPSS 14.0 for windows (SPSS Inc., 2004). The results of the analysis and associated findings are discussed in Chapter 4. By way of summary and conclusion, this Chapter recapitulates significant findings, limitations of the present study, and suggestions for future research.

Summary of Findings

Categorization and conceptualization as primary cognitive activities, an attempt was made to investigate the nexus between human cognition and knowledge structure in the UMLS. Both human cognition and the UMLS knowledge structure are qualified for a select and specific domain. The semantic type “Food” and the semantic relation types in the “associated_with” class are the specific domains. With the underlying rationale that understanding the human conceptual representation will offer a valuable insight on how to improve represent concepts in IR systems, four research questions guided the research work. Highlights of the significant findings are summarized below by each of these research questions.

The purpose of the first research question was to learn the bases of categorization judgment (coherence criteria). Analysis of the result of the food sorting task revealed that individuals in general can create consistent and coherent categories. Forty-nine participants partitioned the 60 food names into categories, ranging from 6 to 17 numbers of categories (see Table 4). The average number of partition for the entire participants is about 12, with an average of five food names per category. While this result offers insight to the lumpers-splitter continuum, it does not show the actual contents of the category members. Using the decision tree exploration engine, however, we managed to predict consistently and correctly the category members for

12 of the 17 categories, a 71% consistency and coherence rate. As shown by the p-value statistic (see Table 7), the predicted category members show strong dependence.

The coherence criteria individuals frequently use to base their category judgment was revealed by running text analysis. Textual patterns, their frequency, and extraction of the normal forms in the text revealed that humans largely use simplified, definition type, naïve explanations. For example, participants' descriptions for various categories resemble a pattern: for the "grain" category, "the blooms of grasses produce grains which are processed in many ways and eaten"; for the "dairy" category, "milk products, things that come from a cow," and for the "Condiment" category, "things whose primary role is to make other things taste better."

The second research question concerned the characterization issue of the categorization task. Because characterization is essentially explained by intension, and the intensions are obtained through the description/verbalization statements, an attempt was made to analyze and graphically present these intensions. Without reducing the intensions or theoretical explanations into strict word forms - phrases, conceptual units, and synsets were extracted from the text analysis in order to create a meaningful and coherent interrelationships among the nodes in the graph. The link analysis and link chart visualization tools in Polyanalyst helped present the web of semantic relationships between these conceptual units for each category. The biographical variables of the research participants were also represented in the graph and it clearly shows the remarkable difference between sexes and subject expertise. For the samples of graphs presented (see Figures 17-21), one can easily observe the

contribution from “Male” participants and participants whose major is “biology,” and “hospitality management” or from those participants who said “yes” for a food science background.

Research question 3 was about the correspondence between human conceptual map and the knowledge structure in the UMLS. Using the least restrictive notion of “distance” as a function of relatedness, this study used the nonmetric multidimensional scaling to represent individuals sorting data (as a reflection of their conceptual map) in a low-dimensional space. Kruskal’s stress coefficient measures of 0.12791 and RSQ of 0.93092 were obtained to explain the goodness-of-fit. These results indicate significant level of variance in the low-dimensional space being accounted for by the experimental data (individuals sorting data).

The fourth research question aimed to complete the task of the exploratory investigation by taking semantic relations as important constraints for a coherent knowledge structure. Guided by the psychological theory of semantic relations, the purpose of the fourth research question was to investigate the correspondence between human relation classification and existing relation hierarchy in the UMLS semantic network. Forty research participants performed the classification of semantic relations into five categories/classes and overall, they consistently classified 29 of the 47 semantic relations in their appropriate class (about 62% accuracy). The appropriateness, accuracy, consistency, or coherence, throughout this research are gauged by comparing results with the UMLS knowledge structure. The results differ significantly from one class to another, ranging from 80% consistency (in the functional class) to 23% consistency (in the conceptual class). This result of a pre-test

done during the proposal presentation of this study was corroborated by this final result. During the pre-test, involving 15 participants, we managed to obtain 26 of the 47 relations classified consistently. Given the size of participants, it is plausible to conclude that the current result further confirms the extent to which the experimental data can be used to recreate the relation hierarchy.

Limitations of the Study

As suggested in the beginning of this Chapter, the nature of this study is very complex to tackle in its entirety. The concepts selected from the UMLS are not in any way representative of the 1.3 million concepts available in the 2006AC version of the UMLS knowledge sources. The millions of concepts are abstracted into 135 broader categories, one of which is the semantic type “Food.” Concepts are purposefully selected from this semantic type, food, because a general assumption is being made from the beginning that native English speakers without any prior background in food science can understand and organize food concepts. While the selection of the 60 food names has a theoretical basis, the number is very small again compared to the total food and substance names that exist in the “food” category. A fine line is being drawn to balance the cognitive stress on the part of the participants that the cognitive task may impose and the adequacy of the stimulus materials. The extant literature considers 50-100 stimulus materials as a relatively large stimulus sets for the task of sorting (Kruskal & Wish, 1978).

In view of the limitations with the size and composition of the stimulus materials, the result of this study cannot, therefore, be generalized to the UMLS knowledge structure as it exists today. Moreover, the UMLS is a knowledge

representation scheme primarily for bio-medical domain and involving participants from the same domain would have provided a better result. Polyanalyst™ 5.0, the software used for the analysis of text and for discriminant analysis, requires large amount of data for a better prediction and knowledge discovery. The data collected, especially the description statements, are very small for a highly credible result.

Concluding Remarks

Cognitive theories offer numerous explanations on how humans represent, organize, and use concepts. The literature review Chapter discusses the strength and limitations of several of these cognitive theories. This study is, in some sense, a two-pronged approach: (a) to investigate the theory-based approach to categorization as an adequate theoretical view to concept representation, and (b) to investigate the connection between human conceptual representation system (the cognitive map-like structure that people have inside their head) and a knowledge representation scheme in IR systems.

The complexity and limitations of this study aside, its driving force has been the recognition of the idea that only a rich and complex network of representation based on semantic content would address the profoundly complex human conceptual system. And for this to happen, the logical route is to start with an understanding of the human conceptual system and to gain better theoretical views that go beyond similarity and feature/attribute/property correlation.

In a small but modest approach, this study carried out an exploratory investigation to reveal humans' conceptual organization for a select sample of concepts. The results and findings are indicative of the potential for using humans'

cognitive performance as input for knowledge structures. The fact that strong support is found in this study that there is a remarkable difference in the way novices and experts organize, represent, and describe concepts (although not new finding in this study), the methodological and data-analytic tools lend themselves to wider-scale investigation.

The classification of semantic relations also has important findings to consider after a re-test with medical expertise. For example, in both the pilot-test and the final study for the classification of semantic relations, subjects performed well in the “functional” and “spatial” classes. Moreover, in both studies, the relation type “occurs_in” is classified in the “temporal” class. Despite naïve and at times idiosyncratic nature of subjects’ theoretical explanations of the category judgment, it was possible to represent a coherent structure of concepts as illustrated by the web of relations in the link graphs. It is also worth emphasizing the important role played by language for there are about 22 participants out of 49 who created a category “Other/I do not know” for food names they said they do not know, such as “sprats, lard, ghee, & weetos.” This can be partly explained by language or it may partly has to do with cultural issues. The language issue is a reminder that both lexical and semantic relations are important for a coherent knowledge structure. This is a further testimony to the notion that the lexical relation is a reflection of the underlying semantic relation (Khoo & Na, 2006).

Implications of Research Findings

Consistency problems and structural issues are widely reported in the unified medical language system. At the Lister Hill National center for Biomedical

Communications (National Library of medicine, NIH), several initiatives, such as SPECIALIST natural language tools, semantic knowledge representation, and medical ontology research, have been made to develop automated algorithms to check the consistency and structural coherence of the UMLS knowledge structure.

However, there is no adequate solution to the problems in the UMLS knowledge sources (i.e., representing the content of the biomedical resources in explicit and coherent fashion to render human cognitive validity and desired reasoning by the machine).

The implication of this study is one that it may potentially offer an alternative to concept auditing task on a subject-by-subject basis. Through the theory-based approach to concept representation, domain expertise in a specific area can be involved to provide a richer explanation of the domain that can be reconstructed in the system. Concept auditing by humans is cumbersome and prone to error. Once initial human conceptual structure is obtained for a given domain of the scheme, automated natural language system tools can be implemented for the auditing task.

Further implications can be in the area of ontology research. The fact that a complex knowledge representation system needs to address structural, semantical, syntactical, and ontological considerations to maintain a coherent knowledge source, this study can be of a good start in terms of techniques and tools to address the semantical and ontological issues. Moreover, involving cognitive theories with actual domain expertise will offer a promise to a better visualization tools for concept auditing and even for presenting search results.

The theory-based approach to categorization and the set of data-analytic tools employed in this study are novel approaches and given further validation in future studies, there is a great potential to develop a standard concept auditing tool. In addition to knowledge representation framework, enterprise-wide information architecture, taxonomies, and directory services can be better organized by using the sets of tools and theories used in this study.

Suggestions for Future Research

The Holy Grail in this type of study is probably to design a detailed specification of a rich and complex knowledge representation system that is both valid in cognitive terms and adequate to understand content in IR systems. This study is far from achieving that noble cause. However, this study is an attempt in the right direction, for it emphasizes to model human conceptual system. Future research should explore to find a method for the formalization the theory-based approach to concept representation as a venue to a grand knowledge representation framework.

Together with an exhaustive knowledge representation system, future research also needs to explore the method for implementing a rich semantic relations framework that clearly and exhaustively establishes the link between concepts. The current semantic relations in the UMLS are organized under two root nodes, *isa* and *associated_with*. The *associated_with* relation category is further organized under five dimensions. Although this study did not focus on the *isa* type of semantic relations, research has already shown that the *isa* relation type is too overloaded (Burgun & Bodenreider, 2001). For example, the provision for two sibling categories as children of a higher level category may necessarily be incompatible, which is non-

existent in most relation hierarchy implementations today. If, for instance “mental state” and “physical state” are children of “state,” structurally they are compatible. However, in terms of ontological and semantic properties, both concepts need to be incompatible.

In the *associated_with* type of semantic relations alone, future research needs to reevaluate the relations in the “conceptual” class. Research participants performed poorly in the classification of the conceptual relation category and subjects largely interpret the conceptual dimension with the physical and functional dimension. The restructuring of a relational class will definitely have enormous effect because reasoning by the machine is based on how these relations link concepts and categories at different levels of instantiation.

Summary

It is always appropriate to ask what one has achieved as a result of a dissertation work like this. This Chapter attempts to provide answers to such questions by presenting the major findings by way of conclusions. Because detailed findings and discussions were made in the previous Chapter, the conclusion here is meant to provide the skeleton of the study. This Chapter also presents limitations of the study given the nature of the study itself.

The implications of this study are also discussed briefly with a view to provide insight on what this study might offer in terms of methodology and approach for similar studies. In the end, this Chapter offered suggestions for future research.

APPENDIX A
THE UMLS 135 SEMANTIC TYPES

- ▪ ▪ Regulation or Law
- ▪ Language
- ▪ Occupation or Discipline
- ▪ ▪ Biomedical Occupation or Discipline
- ▪ Organization
- ▪ ▪ Health Care Related organization
- ▪ ▪ Professional Society
- ▪ ▪ Self-help or Relief organization
- ▪ Group Attribute
- ▪ Group
- ▪ ▪ Professional or Occupational Group
- ▪ ▪ Population Group
- ▪ ▪ Family group
- ▪ ▪ Age Group
- ▪ ▪ Patient or Disabled Group

EVENT

- Activity
- Behavior
- ▪ Social Behavior
- ▪ Individual Behavior
- ▪ Daily or Recreational Activity
- ▪ Occupational Activity
- ▪ ▪ Health Care Activity
- ▪ ▪ Laboratory Procedure
- ▪ ▪ Diagnostic Procedure
- ▪ ▪ Therapeutic or Preventive Procedure

- ▪ ▪ Research Activity
- ▪ ▪ Molecular Biology Research Technique
- ▪ ▪ Government or Regulatory Activity
- ▪ ▪ Educational Activity
- ▪ Machine Activity
- Phenomenon or Process
- ▪ Human-caused Phenomenon or process
- ▪ ▪ Environmental Effect of Humans
- ▪ Natural Phenomenon or process
- ▪ ▪ Biologic Function
- ▪ ▪ ▪ Physiologic Function
- ▪ ▪ ▪ Organism Function
- ▪ ▪ ▪ ▪ Mental Process
- ▪ ▪ ▪ ▪ Organ or Tissue Function
- ▪ ▪ ▪ ▪ Cell Function
- ▪ ▪ ▪ ▪ Molecular Function
- ▪ ▪ ▪ ▪ Genetic Function
- ▪ ▪ ▪ Pathologic Function
- ▪ ▪ ▪ Disease or Syndrome
- ▪ ▪ ▪ Mental or Behavioral Dysfunction
- ▪ ▪ ▪ ▪ Neoplastic process
- ▪ ▪ ▪ Cell or Molecular Dysfunction
- ▪ ▪ ▪ Experimental Model of Disease
- ▪ Injury

APPENDIX B

THE UMLS 54 SEMANTIC RELATIONS

ASSOCIATED_WITH

▪ **Physically_related_to**

- ▪ part_of
- ▪ consists_of
- ▪ contains
- ▪ connected_to
- ▪ interconnects
- ▪ branch_of
- ▪ tributary_of
- ▪ ingredient_of

▪ **Spatially_related_to**

- ▪ location_of
- ▪ adjacent_to
- ▪ surrounds
- ▪ traverses

▪ **Functionally_related_to**

- ▪ *Affects*
- ▪ ▪ manages
- ▪ ▪ treats
- ▪ ▪ disrupts
- ▪ ▪ complicates
- ▪ ▪ interacts_with
- ▪ ▪ prevents
- ▪ *Brings_about*
- ▪ ▪ produces
- ▪ ▪ causes
- ▪ *Performs*

- ▪ ▪ carries_out

- ▪ ▪ exhibits

- ▪ ▪ practices

- ▪ *Occurs_in*

- ▪ ▪ process_of

- ▪ uses

- ▪ manifestation_of

- ▪ indicates

- ▪ result_of

▪ **Temporally_related_to**

- ▪ co-occurs_with

- ▪ precedes

▪ **Conceptually_related_to**

- ▪ evaluation_of

- ▪ degree_of

- ▪ *Analyzes*

- ▪ ▪ assesses_effect_of

- ▪ measurement_of

- ▪ measures

- ▪ diagnoses

- ▪ property_of

- ▪ derivative_of

- ▪ developmental_form_of

- ▪ method_of

- ▪ conceptual_part_of

- ▪ issue_in

APPENDIX C

THE EXPERIMENTAL 60 FOOD NAMES

1. Apricot
2. Avocado
3. Bacon
4. Banana
5. Beans
6. Beef
7. Berries
8. Bread
9. Broccoli
10. Butter
11. Cabbage
12. Cashews
13. Cheese
14. Chocolate
15. Corn
16. Cracker
17. Ghee
18. Grapes
19. Honey
20. Kit kat
21. Lager
22. Lard
23. Lentils
24. Lettuce
25. Margarine
26. Martini
27. Melons
28. Millet
29. Milk
30. Milky way
31. Muffin
32. Mustard
33. Noodles
34. Nougat
35. Nuts
36. Oats
37. Onions
38. Pasta
39. Peas
40. Pineapple
41. Popcorn
42. Potato
43. Rice
44. Salad
45. Sardine
46. Scampi
47. Shreddies
48. Snicker
49. Soybeans
50. Sprats
51. Stout
52. Tea
53. Tigernut
54. Tomatoes
55. Twix
56. Wheat
57. Wine
58. Weetos
59. Yogurt
60. Zucchini

APPENDIX D

LETTER OF INVITATION TO THE FIRST GROUP

Date: _____

Letter of Invitation

Shimelis Assefa
Ph.D Student, Interdisciplinary program in information sciences
School of Library and Information Sciences, University of North Texas.

Dear research participant,

As part of the partial requirement for the Ph.D. program in Information Sciences, I am conducting a research on how humans use, organize and represent concepts. The title of my dissertation is “Human Concept Cognition and Semantic Relations in the Unified Medical Language System: a Coherence Analysis.”

You are presented with 60 food names all taken from a bio-medical knowledge source known as ‘Unified Medical Language System’ or UMLS for short. Your task is to sort these food names into piles (no limit on the number of piles with each pile containing as many food names as you find it appropriate). The sorting task is based on your understanding of how food names relate to one another. Following the sorting, you are required to verbalize or describe the criteria or rules you used to form each pile.

The goal of this study is to offer an alternative on how concepts can be better organized and represented in the knowledge source in the bio-medical sciences. Your participation is vital to the success of this study because it offers knowledge representation systems the human element, i.e., how humans understand, use, and represent concepts, that is little understood thus far.

This research will only take about 20-30 minutes of your time. When you are ready, please visit www.websort.net/go/foodsorting to perform the task.

Please note that 1) Participation is voluntary and you may discontinue participation at any time without penalty, 2) Your personal information and answers will be kept confidential and results of this study will be reported only on a group basis, and 3) You may keep this notice for your records.

This research project has been reviewed and approved by the UNT Institutional Review Board (940) 565-3940. Contact the UNT IRB with any questions regarding your rights as a research subject.

Thank you very much for taking time to participate in this study and please don't hesitate to contact me or my major advisor (address below) if you have any questions.

Sincerely,

Principal Investigator

Shimelis Assefa

308 Bradley St., Apt. 35

SAssefa@lis.admin.unt.edu

(214) 507 7729 or (940) 565 2186

Faculty Major Advisor

Dr. Brian C. O'Connor

SLIS, UNT

BOConnor@lis.admin.unt.edu

(940) 206-1172

APPEDNIX E

LETTER OF INVITATION TO THE SECOND GROUP

Date: _____

Letter of Invitation

Shimelis Assefa
Ph.D Student, Interdisciplinary program in information sciences
School of Library and Information Sciences, University of North Texas.

Dear research participant,

As part of the partial requirement for the Ph.D. program in Information Sciences, I am conducting research on how humans use, organize and represent concepts. The title of my dissertation is “Human Concept Cognition and Semantic Relations in the Unified Medical Language System: a Coherence Analysis.”

You are presented with 47 relation types all taken from a bio-medical knowledge source known as ‘Unified Medical Language System’ or UMLS for short. Your task is to sort these relation types into 5 categories. The 5 categories are labeled and defined for you to facilitate the task of sorting. Your sorting is based on your understanding of which relation types belong to which of the 5 categories.

The goal of this study is to offer an alternative on how concepts can be better organized and represented in the vocabulary store in the bio-medical sciences. Your participation will help gain a better insight to incorporate the human element in knowledge organization practices, i.e., (how humans understand, use, and represent concepts), that is little understood thus far.

This research will only take about 20-30 minutes of your time. When you are ready, please visit www.websort.net/go/relationssort to perform the task.

Please note that 1) Participation is voluntary and you may discontinue participation at any time without penalty, 2) Your personal information and answers will be kept confidential and results of this study will be reported only on a group basis, and 3) You may keep this notice for your records.

This research project has been reviewed and approved by the UNT Institutional Review Board (940) 565-3940. Contact the UNT IRB with any questions regarding your rights as a research subject.

Thank you very much for taking time to participate in this study and please don't hesitate to contact me or my major advisor (address below) if you have any questions.

Sincerely,

Principal Investigator

Shimelis Assefa

308 Bradley St., Apt. 35

SAssefa@lis.admin.unt.edu

(214) 507 7729 or (940) 565 2186

Faculty Major Advisor


Dr. Brian C. O'Connor




SLIS, UNT

BOConnor@lis.admin.unt.edu

(940) 206-1172



APPENDIX F
FOOD SORTING TASK

Address  http://websort.net/go/foodsorting


websort  text size   [Instructions](#)
[Add Comments](#)

60 items remaining to sort

Drag items into these folders

 click to name 



- Avocado
- Milky way
- Zucchini
- Sardine
- Apricot
- Sprats
- Tomatoes
- Twix
- Wheat
- Cracker
- Pasta
- Chocolate
- Grapes
- Millet
- Lard
- Margarine
- Onions
- Lager
- Berries
- Cabbage
- ...






 Done


APPENDIX G

SEMANTIC RELATION CLASSIFICATION TASK

Address <http://websort.net/go/relationsort>

websort text size   [Instructions](#)
[Add Comments](#)

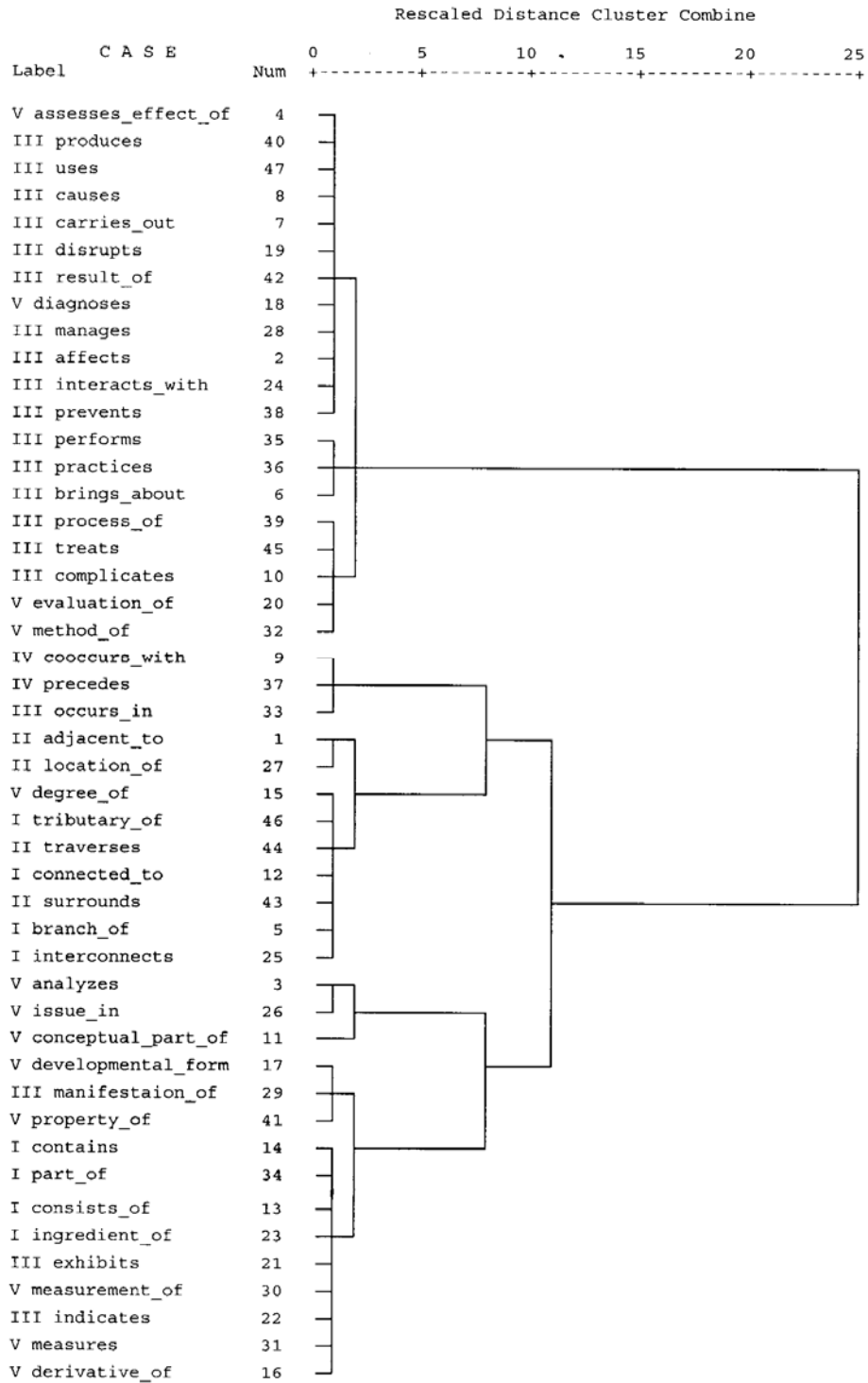
| | | |
|---|---|--|
| <p>47 items remaining to sort</p> <ul style="list-style-type: none"> <input type="text" value="process_of"/> <input type="text" value="uses"/> <input type="text" value="occurs_in"/> <input type="text" value="manages"/> <input type="text" value="tributary_of"/> <input type="text" value="location_of"/> <input type="text" value="evaluation_of"/> <input type="text" value="affects"/> <input type="text" value="conceptual_part_of"/> <input type="text" value="interconnects"/> <input type="text" value="interacts_with"/> <input type="text" value="treats"/> <input type="text" value="connected_to"/> <input type="text" value="measurement_of"/> <input type="text" value="co-occurs_with"/> <input type="text" value="ingredient_of"/> <input type="text" value="exhibits"/> <input type="text" value="assesses_effect_of"/> <input type="text" value="result_of"/> <input type="text" value="degree_of"/> <input type="text" value="derivative_of"/> <input type="text" value="property_of"/> | <p>Drag items into these folders</p> <ul style="list-style-type: none">  TEMPORAL RELATION TYPES  SPATIAL RELATION TYPES  PHYSICAL RELATION TYPES  FUNCTIONAL RELATION <li style="background-color: #ffff00;"> CONCEPTUAL RELATION ▶ | <p>Selected group's name: CONCEPTUAL RELATION TYPES (Related by some abstract concept, thought, or idea)</p> |
|---|---|--|

 Done

APPENDIX H
DENDROGRAM TREE ACCORDING TO PARTICIPANTS'
RELATION CLASSIFICATION

***** H I E R A R C H I C A L C L U S T E R A N A L Y S I S *****

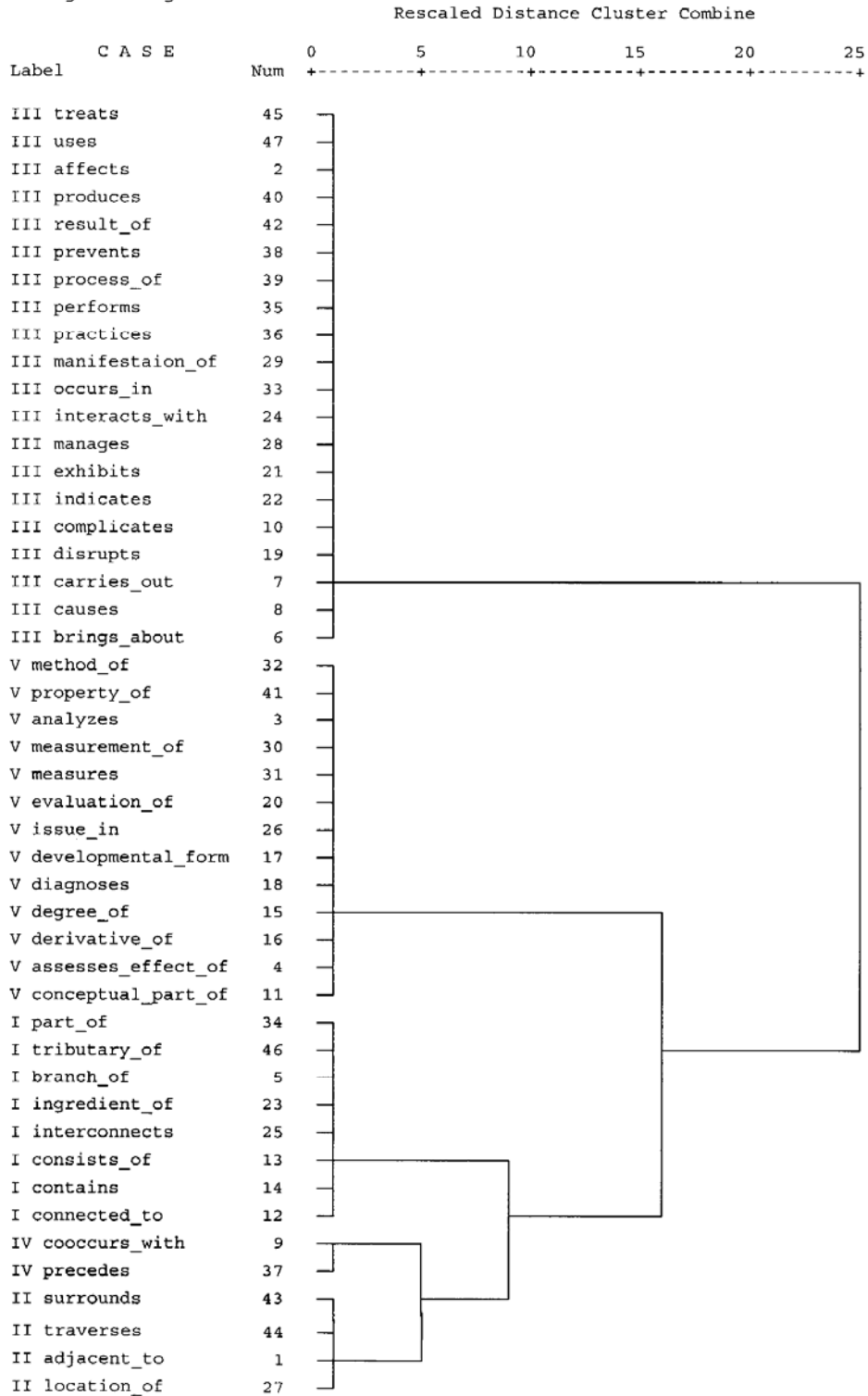
Dendrogram using Ward Method



APPENDIX I

DENDROGRAM TREE ACCORDING TO UMLS RELATION HIERARCHY

***** HIERARCHICAL CLUSTER ANALYSIS *****
 Dendrogram using Ward Method



APPENDIX J
THE ALSICAL PROCEDURE IN SPSS

Alscal Procedure Options

Data Options-

| | |
|---------------------------------------|---------------|
| Number of Rows (Observations/Matrix). | 60 |
| Number of Columns (Variables) . . . | 60 |
| Number of Matrices | 1 |
| Measurement Level | Ordinal |
| Data Matrix Shape | Symmetric |
| Type | Dissimilarity |
| Approach to Ties | Leave Tied |
| Conditionality | Matrix |
| Data Cutoff at | .000000 |

Model Options-

| | |
|----------------------------------|---------------|
| Model | Euclid |
| Maximum Dimensionality | 2 |
| Minimum Dimensionality | 2 |
| Negative Weights | Not Permitted |

Output Options-

| | |
|--|-------------|
| Job Option Header | Printed |
| Data Matrices | Printed |
| Configurations and Transformations | Plotted |
| Output Dataset | Not Created |
| Initial Stimulus Coordinates | Computed |

Algorithmic Options-

| | |
|-------------------------------------|----------|
| Maximum Iterations | 30 |
| Convergence Criterion | .00100 |
| Minimum S-stress | .00500 |
| Missing Data Estimated by | Ulbounds |
| Tiestore | 1000 |

APPENDIX K
IRB APPROVAL LETTER

March 30, 2007

Shimelis Assefa
School of Library and Information Science
University of North Texas

Institutional Review Board for the Protection of Human Subjects in Research (IRB)
RE: Human Subject Application #07-013

Dear Mr. Assefa:

The UNT IRB has received your request to modify your study titled "Human Concept Cognition and Semantic Relations in the Unified Medical Language System: A Coherence Analysis." As required by federal law and regulations governing the use of human subjects in research projects, the UNT IRB has examined the request to change the data collection instrument for this study. The modification to this study is hereby approved for the use of human subjects. **Approval for this project is March 7, 2007 through March 6, 2008.**

It is your responsibility according to U.S. Department of Health and Human Services regulations to submit annual and terminal progress reports to the IRB for this project. Please mark your calendar accordingly. The IRB must also review this project prior to any other modifications made. **Federal policy 21 CFR 56.109(e) stipulates that IRB approval is for one year only.**

Please contact Shelia Bourns, Research Compliance Administrator, at (940) 565-3940, or Boyd Herndon, Director of Research Compliance, at (940) 565-3941, if you wish to make changes or need additional information.

Sincerely,



Scott Simpkins, Ph.D.
Chair
Institutional Review Board

SS/sb

REFERENCES

- Allen, B.L. (1991). Cognitive research in information science: implications for design. In M.E. Williams (Ed.), *Annual review of information science and technology* (pp. 3-37). Medford, NJ: Learned Information, Inc.
- Allen, J.F., & Frisch, A.M. (1982). What's in a semantic network? *Proceedings of the 20th Annual Meeting on Association for Computational Linguistics* (pp. 19-27). Morristown, NJ: Association for Computational Linguistics.
- Arocha, J.F., Wang, D., & Patel, V.L. (2005). Identifying reasoning strategies in medical decision making: A methodological guide. *Journal of Biomedical Informatics*, 38(2), 154-171.
- Asher, R.E. (1994). *The encyclopedia of language and linguistics*. Oxford, UK: Pergamon Press.
- Bar-Hillel, Y. (1964). *Language and information: Selected essays on their theory and application*. Reading, MA: Addison-Wesley Co.
- Barsalou, L.W. (1992). *Cognitive psychology: An overview for cognitive scientists*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Barsalou, L.W. (2001). The human conceptual system. *Proceedings of the International Conference on Formal Ontology in Information Systems* (p. 186). New York, NY: ACM Press.
- Barsalou, L.W., & Hale, C.R. (1993). Components of conceptual representation: from feature lists to recursive frames. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 97-144). London: Academic Press.
- Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5), 407-424.
- Bean, C.A., & Green, R. (2001). *Relationships in the organization of knowledge*. Dordrecht: Kluwer Academic Publishers.
- Belkin, N.J., Oddy, J.R.N., & Brooks, H.M. (1982). ASK for information retrieval: Part I, background and theory. *The Journal of Documentation*, 38(2), 61-71.
- Belkin, N.J. (1990). The cognitive viewpoint in information science. *Journal of Information Science*, 16(1), 11-15.
- Blair, D.C. (1990). *Language and representation in information retrieval*. Amsterdam: Elsevier Science Publishers.
- Bodenreider, O., & McCray, A.T. (2003). Exploring semantic groups through visual

- approaches. *Journal of biomedical informatics*, 36(6), 414-432.
- Boorman, S.A., & Arabie, P. (1972). Structural measures and the method of sorting. In R.N. Shepard, A.K. Romney, & S.B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences. Vol 1: Theory* (pp. 225-249). New York, NY: Seminar Press.
- Bower, G.H., & Clapper, J.P. (1989). Experimental methods in cognitive science. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 245-300). Cambridge, MA: The MIT Press.
- Burgun, A., & Bodenreider, O. (2001). Aspects of the taxonomic relation in the biomedical domain. *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 222-233), New York, NY: ACM Press.
- Burton, M.L. (1975). Dissimilarity measures for unconstrained sorting data. *Multivariate behavioral research*, 10(4), 409-423.
- Carpineto, C., & Romano, G. (2004). *Concept data analysis: Theory and applications*. West Sussex: John Wiley & Sons Ltd.
- Carroll, J.D., & Corter, J.E. (1995). A graph-theoretic method for organizing overlapping clusters into trees. *Journal of classification*, 12(2), 283-313.
- Chaffin, R., & Herrmann, D.J. (1988). The nature of semantic relations: A comparison of two approaches. In M.W. Evens (Ed.), *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks* (pp. 289-334). Cambridge: Cambridge University Press.
- Chaffin, R., & Herrmann, D.J. (1987). Relation element theory: a new account of the representation and processing of semantic relations. In D.S. Gorfein & R.R. Hoffman (Eds.), *Memory and learning: The Ebbinghaus Centennial Conference* (pp. 221-245). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cimino, J.J., Min, H., & Perl, Y. (2003). Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *Journal of Biomedical Informatics*, 36(6), 450-461.
- Corter, J.E. (1996). *Tree models of similarity and association*. Thousand Oaks, CA: Sage Publications, Inc.
- Coxon, A.P.M. (1999). *Sorting data: Collection and analysis*. Thousand Oaks, CA: Sage Publications.
- Cruse, A. (2004). *Meaning in language: An introduction to semantics and pragmatics*. Oxford, NY: Oxford University Press.
- Davison, M.L. (1983). *Multidimensional scaling*. New York: John Wiley & Sons.

- De Mey, M. (1982). *The cognitive paradigm: cognitive science, a newly explored approach to the study of cognition applied in an analysis of science and scientific knowledge*. Dordrecht: D. Reidel Publishing Company.
- Dervin, B. (1999). On studying information seeking methodologically: the implications of connecting metatheory to method. *Information Processing & Management*, 35(6), 727-750.
- Edelbrock, C., & McLaughlin, B. (1980). Hierarchical cluster analysis using interclass correlations: A mixture model study. *Multivariate Behavioral Research*, 15(3), 299-318.
- Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation*, 45(3), 171-212.
- Ellis, D. (1992). The physical and cognitive paradigms in information retrieval research. *Journal of Documentation*, 48(1), 45-64.
- Feger, H., & De Boeck, P. (1993). Categories and concepts: introduction to data analysis. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 203-223). London: Academic press.
- Fodor, J.A. (1998). *Concepts: Where cognitive science went wrong*. New York, NY: Clarendon Press/Oxford University Press.
- Foskett, D.J. (1997). Thesaurus. In K.S. Jones & P. Willett (Eds.), *Readings in information retrieval* (pp. 111-134). San Francisco, CA: Morgan Kaufmann Publishers, Inc.
- Frumkina, R.M. & Mikhejev, A.V. (1996). *Meaning and categorization*. New York: Nova Science Publishers, Inc.
- Green, R. (2001). Relationships in the organization of knowledge: an overview. In C.A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 3-18). Dordrecht: Kluwer Academic Publishers.
- Griffith, R.L. (1982). Three principles of representation for semantic networks. *ACM Transactions on Database Systems*, 7(3), 417-442.
- Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2006). *Multivariate data analysis*. Upper Saddle River, N.J.: Prentice Hall.
- Hampton, J. (1993). Prototype models of concept representation. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 67-95). London: Academic Press.

- Hampton, J., & Dubois, D. (1993). Psychological models of concepts: introduction. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 11-33). London: Academic Press.
- Harloff, J. (2005). Multiple level weighted card sorting. *Methodology*, 1(4), 119-128.
- Harnad, S. (1987). Introduction: Psychophysical and cognitive aspects of categorical perception: a critical overview. In S. Harnad (Ed.), *Categorical perception: the groundwork of cognition* (pp. 1-25). Cambridge: Cambridge University Press.
- Hartigan, J.A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association*, 62(320), 1140-1158.
- Hautamaki, A. (1992). A conceptual space approach to semantic networks. In F. Lehmann (Ed.), *Semantic Networks in Artificial Intelligence* (pp. 517-525). Oxford: Pergamon Press.
- Hayes, P. (1999). Knowledge representation. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 432-434). Cambridge, MA: The MIT Press.
- Houston, A.L., Chen, H., Schatz, B.R., Hubbard, S.M., Sewell, R.R., & Ng, T.D. (2000). Exploring the use of concept spaces to improve medical information retrieval. *Decision Support Systems*, 30(2), 171-186.
- Houston, A.L. (1998). *Knowledge integration for medical informatics: An experiment on a cancer information system*. Unpublished doctoral dissertation, University of Arizona, Phoenix, AZ.
- Humphreys, B.L., & Lindberg, D.A. (1993). The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, 81(2), 170-177.
- Ingwersen, P. (1999). Cognitive information retrieval. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology* (pp. 3-52). Medford, NJ: Information today, Inc.
- Ingwersen, P. (1993). The cognitive viewpoint in IR. *Journal of Documentation*, 49(1), 60-64.
- Iris, M.A., Litowitz, B.E., & Evens, M. (1988). Problems of the part-whole relation. In M.W. Evens (Ed.), *Relational models of the lexicon: Representing knowledge in semantic networks* (pp. 261-288). Cambridge: Cambridge University Press.
- Keil, F.C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: The MIT Press.

- Khoo, C.S.G., & Na, J.-C. (2006). Semantic relations in Information science. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (pp. 157-228). Medford, NJ: Information Today, Inc.
- Koll, M.B. (1979). *The concept space in information retrieval systems as a model of human concept relations*. Unpublished doctoral dissertation, Graduate School of Syracuse University, Syracuse, NY.
- Koubek, R.J., & Mountjoy, D.N. (1991). *Toward a model of knowledge structure and a comparative analysis of knowledge structure measurement techniques*. Unpublished manuscript, Wright State University at Dayton, OH.
- Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27.
- Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2), 115-129.
- Kruskal, J.B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage Publications, Inc.
- Kuhn, T.S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Lehmann, F. (1992). Semantic networks. In F. Lehmann (Ed.), *Semantic Networks in Artificial Intelligence* (pp. 1-50). Oxford: Pergamon Press.
- Lin, J., & Demner-Fushman, D. (2006). The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 99-106). New York, NY: ACM Press.
- Loftus, E.F. (1975). Spreading activation within semantic categories: Comments on Rosch's "Cognitive Representations of Semantic Categories". *Journal of Experimental Psychology: General*, 104(3), 234-240.
- Lyons, J. (1977). *Semantics*. Cambridge, UK: Cambridge University Press.
- Markman, A.B., & Makin, V.S. (1998). Referential Communication and Category Acquisition. *Journal of Experimental Psychology: General*, 127(4), 331-354.
- Markowitz, J. (1988). An exploration into graded set membership. In M.W. Evens (Ed.), *Relational models of the lexicon: Representing knowledge in semantic networks* (pp. 239-260). Cambridge: Cambridge University Press.
- Markowitz, J.A., Nutter, J.T., & Evens, M.W. (1992). Beyond IS-A and part-whole: More semantic network links. In F. Lehmann (Ed.), *Semantic Networks in Artificial Intelligence* (pp. 377-390). Oxford: Pergamon Press.

- McCray, A.T., & Nelson, S.J. (1995). The representation of meaning in the UMLS. *Methods of Information in Medicine*, 34(1-2), 193-201.
- Mechelen, I.V., & Michalski, R.S. (1993). General introduction: Purpose, underlying ideas, and scope of the book. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 1-8). London: Academic Press.
- Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist*, 44(12), 1469-1481.
- Medin, D.L., & Aguilar, C. (1999). Categorization. In R.A. Wilson & F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 104-106). Cambridge, MA: The MIT Press.
- Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207-238.
- Medin, D.L., & Smith, E.E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113-38.
- Mervis, C.B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89-115.
- Michalski, R.S. (1993). Beyond prototypes and frames: the two-tiered representation. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and Concepts: Theoretical Views and Inductive Data Analysis* (pp. 145-172). London: Academic Press.
- Michalski, R.S. (1989). Two-tiered concept meaning, inferential matching, and conceptual cohesiveness. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 122-145). Cambridge: Cambridge University Press.
- Miller, G.A. (1969). A psychological method to investigate verbal concepts. *Journal of mathematical psychology*, 6, 169-191.
- Miller, G.A., & Johnson-Laird, P.N. (1976). *Language and perception*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Murphy, G.L. (1993). Theories and concept formation. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173-200). London: Academic Press.
- Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.
- Murtagh, F.D. (1993). Cluster analysis using proximities. In I.V. Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), *Categories and concepts:*

- theoretical views and inductive data analysis* (pp. 225-245). London: Academic press.
- Nelson, S.J., Powell, T., & Humphreys, B.L. (2002). The unified medical language system (UMLS) project. In A. Kent (Ed.), *Encyclopedia of Library and Information Science* (pp. 369-378). New York: Marcel Dekker.
- Ng, T.D. (2000). *A concept space approach to semantic exchange*. Unpublished dissertation, The University of Arizona, Phoenix, AZ.
- NISO. National Information Standards Organization (2005). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (ANSI/NISO Z39.19 - 2005)*. Bethesda, MD: NISO Press.
- O'Connor, B.C. (1996). *Explorations in indexing and abstracting: pointing, virtue, and power*. Englewood, Colorado: Libraries Unlimited, Inc.
- O'Connor, B.C., Copeland, J.H., & Kearns, J.L. (2003). *Hunting and gathering on the information savanna: Conversations on modeling human search abilities*. Lanham, MD: Scarecrow Press.
- Oroumchian, F. (1995). *Information retrieval by plausible inferences: An application of the theory of plausible reasoning of Collins and Michalski*. Unpublished doctoral dissertation, Syracuse University, New York, NY.
- Osherson, D.N., & Smith, E.E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1), 35-58.
- Palmer, S.E. (1978). Fundamental aspects of cognitive representation. In E. Rosch, & B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 259-303). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Perl, Y., & Geller, J. (2003). Research on structural issues of the UMLS—past, present, and future. *Journal of Biomedical Informatics*, 36(6), 409-413.
- Pirolli, P., & Card, S. (1998). Information foraging models of browsers for very large document spaces. *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 83-93). New York, NY: ACM Press.
- Megaputer Intelligence, Inc. (2002). PolyAnalyst (version5.0.) [Computer software]. Bloomington, IN. 1994-2002.
- Popper, K.R. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Clarendon Press.
- Pris, U. (2005). Linguistic applications of formal concept analysis. In B. Ganter, G. Stumme, & R. Wille (Eds.), *Formal concept analysis: foundations and applications* (pp. 149-160). Berlin: Springer-Verlag.

- Quillian, M.R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic Information processing* (pp. 216-270). Cambridge, MA: The MIT Press.
- Raphael, B. (1968). SIR: Semantic Information Retrieval. In M. Minsky (Ed.), *Semantic Information Processing* (pp. 33-134). Cambridge, MA: The MIT Press.
- Robertson, S. (2000). Salton Award Lecture on theoretical argument in information retrieval. *ACM SIGIR Forum*, 34(1), 1-10.
- Rodwan, A.S. (1964). An empirical validation of the concept of coherence. *Journal of Experimental Psychology*, 68(2), 167-170.
- Romesburg, H.C. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.
- Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3), 192-233.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Rosch, E., & Lloyd, B.B. (1978). *Cognition and categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4), 573-605.
- Rosenberg, S., & Kim, M.P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10(4), 489-502.
- Salton, G. (1997). A blueprint for automatic indexing. *ACM SIGIR Forum*, 31(1), 23-36.
- Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context. *Discourse Processes*, 24(1), 119-147.
- Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, 42(3), 319-345.
- Schauble, P. (1987). Thesaurus based concept spaces. *Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 254 – 262). New York, NY: ACM Press.
- Schulze-Kremer, S., Smith, B., & Kumar, A. (2004). *Revising the UMLS semantic network*. Unpublished manuscript, IFOMIS, University of Leipzig, Germany

- Schuyler, P.L., Hole, W.T., Tuttle, M.S., & Sherertz, D.D. (1993). The UMLS Metathesaurus: Representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2), 217-222.
- Schvaneveldt, R.W., Durso, F.T., & Mukherji, B.R. (1982). Semantic distance in categorization tasks. *Journal of Experimental Psychology: Learning, memory, and cognition*, 8(1), 1-15.
- Smith, Barry. (2004). Beyond concepts: ontology as reality representation. In A. Varzi, & L. Vieu (Eds.), *Proceedings of FOIS 2004. International conference on Formal Ontology and Information Systems*.
- Smith, E.E. (1989). Concepts and induction. In M.I. Posner (Ed.), *Foundations of cognitive science* (pp. 501-526). Cambridge, MA: The MIT press.
- Smith, E.E., & Medin, D.L. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.
- Sowa, J.F. (1984). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley Publishing Company.
- SPSS Inc. (2004). SPSS for Windows (version 14.0) [Computer Software]. Chicago, IL: SPSS Inc.
- Thagard, P. (1997). Coherent and creative conceptual combinations. In T.B. Ward, S.M. Smith, & J. Vaid (Eds.), *Creative thought: An investigation of conceptual structures and processes* (pp. 129-141). Washington, DC: American Psychological Association.
- Tullis, T. & Wood, L. (n.d). How many users are enough for a card-sorting study? Retrieved January 7, 2007 from the Comcast Web site:
<http://home.comcast.net/~tomtullis/publications/UPA2004CardSorting.pdf>
- Turkington, C., & Harris, J. (2001). *The encyclopedia of memory and memory disorders*. New York, NY: Facts on File, Inc.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.
- U.S. National Library of Medicine. (2005). UMLS knowledge source server (UMLSKS): 2006AC. Bethesda, MD: U.S. National Library of Medicine. Lister Hill National Center for Biomedical Communications. Retrieved December 21, 2006 from: <http://umlsks.nlm.nih.gov/kss/servlet/Turbine/action/KssLogin>
- U.S. National Library of Medicine. (2006). UMLS knowledge sources: July Release 2006AC Documentation. Retrieved September 27, 2006 from: <http://www.nlm.nih.gov/research/umls/documentation.html>

- van der Kloot, W.A., & van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research*, 26(4), 563-581.
- van Rijsbergen, C. J. (1983). Information retrieval: New directions: old solutions. *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 264-265). New York, NY: ACM Press.
- Wang, P. (1999). An empirical study of knowledge structures of research topics. In L. Woods (Ed.), *Proceedings of the 62nd Annual Meeting of the American Society for Information Science* (pp. 557-571). Medford, NJ: Information Today, Inc.
- Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *American Statistical Association Journal*, 58(301), 236-244.
- Weber, R.P. (1990). *Basic content analysis*. Newbury Park, CA: Sage Publications, Inc.
- Wille, R. (1992). Concept lattices and conceptual knowledge systems. In F. Lehmann (Ed.), *Semantic networks in artificial intelligence* (pp. 493-515). Oxford: Pergamon press.
- Wille, R. (2005). Formal concept analysis as mathematical theory of concepts and concept hierarchies. In B. Ganter, G. Stumme, & R. Wille (Eds.), *Formal Concept Analysis: Foundations and applications* (pp. 1-33). Berlin: Springer.
- Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control*. Berkeley: University of California Press.
- Yao, Y. Y. (2004). Concept formation and learning: A cognitive informatics perspective. *Proceedings of the Third IEEE International Conference on Cognitive Informatics* (pp. 42 – 51). Washington, DC: IEEE Computer Society.
- Zhang, Li. (2004). *Enriching and designing metaschemas for the UMLS semantic network*. Unpublished doctoral dissertation, New Jersey Institute of technology, USA.
- Zúñiga, Gloria L. (2001). Ontology: Its transformation from philosophy to information systems. *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 187-197). New York, NY: ACM Press.