DEVELOPING AN ONTOLOGY EVALUATION METHODOLOGY – COGNITIVE MEASURE OF QUALITY

by

Chia-wen (Jennifer) Fang

A thesis

submitted to the Victoria University of Wellington
in fulfilment of the requirement for the degree of
Master of Commerce and Administration in Information Systems

Victoria University of Wellington

October 2008

Abstract

Ontologies are formal specifications of shared conceptualizations of a domain. Important applications of ontologies include distributed knowledge based systems, such as the semantic web, and the evaluation of modelling languages, e.g. for business process or conceptual modelling. These applications require formal ontologies of good quality. In this thesis, we present a multi-method ontology evaluation methodology, which consists of two techniques (sentence verification task and recall) based on principles of cognitive psychology, to test how well a specification of a formal ontology corresponds to the ontology users' conceptualization of a domain. Two experiments were conducted, each evaluating the SUMO ontology and WordNet with an experimental technique, as demonstrations of the multi-method evaluation methodology. We also tested the applicability of the two evaluation techniques by conducting a replication study for each. The replication studies obtained findings that point towards the same direction as the original studies, although no significance was achieved. Overall, the evaluation using the multi-method methodology suggests that neither of the two ontologies we examined is a good specification of the conceptualization of the domain. Both the terminology and the structure of the ontologies, may benefit from improvement.

Contents

List of Tables

List of Figures

Acknowledgements

I would especially like to thank my supervisor, Dr. Joerg Evermann, for his invaluable support, guidance, and friendship, throughout the course of this study.
I would also like to thank the following people for their involvement - Dr. David Mason, my family and friends, for their support; and all the participants who contributed their time to the study.

Chapter 1

Introduction

Ontology is a new concept that is getting a lot of attention from many disciplines, such as information science, finance, medicine, and education sectors [39]. Many definitions are offered to describe this concept. However, the most often cited definition was given by Gruber [38] - an ontology is a formal explicit specification of a shared conceptualization. He describes conceptualization as an abstract model that depicts how people perceive things in the world, usually in a specific subject area. Explicit specification means that explicit terms and definitions are given to the concepts and relationships of the abstract model. Ontologies provide a common vocabulary to share information in a domain [61], and also provide concepts to structure and represent knowledge about a domain [25].

The importance of ontologies is becoming widely recognized. In the field of information systems (IS), understanding the real world domains that IS represents, and managing our knowledge about the domains, is important for developing effective IS. Recent software applications require a more complete set of precise concepts for enabling progress in electronic commerce and software integration [59]. Furthermore, applications of ontologies include knowledge description for intelligent reasoning - ontologies are considered the backbone to application development for the work in the Semantic Web, a variety of Semantic Web Services, Knowledge Management, medical informatics, electronic commerce and other areas [29, 89, 45]. Ontologies are also used for natural language processing, and reference standards for model and modelling language evaluation [95].

The recent interest in ontologies, and the availability of mature, industrial-strength software tools such as Protégé and JENA have led to widespread

efforts to develop and use formal ontologies in a variety of areas. As a result, several competing ontologies have been proposed to represent the same, or overlapping domains, and they often appear to have equal expressiveness and claim similar validity [25, 11, 30, 59]. Ontology evaluation is a keystone to ensure high-quality and representative ontologies [20, 31, 60], and thus, a necessary step to secure the success of the above applications.

There are few widely used techniques to evaluate and compare different ontologies, for example, the task based approach gold-standard ontology evaluation [64], and the data driven ontology evaluation [15]. However, ontologies are intended to conceptualize and reflect the empirically perceived reality [25]. No consensus in the literature has been developed to examine and evaluate the representativeness (the ability to represent the perceive reality) of ontologies. It has been pointed out that if an ontology does not capture the intended semantics of the user's terminology, it will be of little practical use even with great formal properties [39]. Hence, empirically observed structures of reality should be used as an instrument when assessing a particular ontology [25]. It is important to explore the possible methods in order to compare directly the specifications in a formal ontology with people's conceptualization of perceived reality. The conceptualization, that is, our mental model of the domain, is the product of cognitive processes such as perception and recognition. Psychological research on knowledge representation, which refers to the study of cognition and the use and access of pre-existing knowledge structures, is thus motivated to be seen as the most appropriate reference discipline for developing ontology evaluation methodologies. An initial development of such an evaluation approach is presented in Fang and Evermanns' work [28, 27, 26].

The focus of this study is to develop an ontology evaluation methodology based on cognitive concepts and psychology studies that can evaluate the quality (representativeness) of ontological structures - whether they reflect the perceived reality of the world. In this study we develop a multimethod evaluation methodology which consists of two independent evaluation techniques, sentence verification task and recall. We evaluate one ontology, SUMO, and a popular thesaurus, WordNet, which is often used as an ontology and will be referred to as an ontology in this study. The evaluation of these two ontologies is used as a demonstration of the application of this methodology by applying the two techniques and assessing the level of agreement between the results obtained from these techniques.

This thesis makes three contributions. First, we present a new notion of

the quality of ontologies, based on the idea that this quality concerns the relationship between a cognitive conceptualization and an explicit specification. Second, we present an experimental method to evaluate this new aspect of quality. Because we are concerned with the relationship between a cognitive, mental model and its explicit specification, this method is based primarily on research in cognitive psychology. Third, we provide a demonstration of an evaluation of two ontologies, using this notion of quality and the proposed method. This study will contribute to practitioners such as web developers and system integrators with a more adequate tool for the development and deployment of the applications, by ensuring the adoption of the ontologies that best represent the perceived world.

The remainder of this thesis is structured as follows. Chapter 2 positions this thesis within related work and motivates our proposed method. We also describe the cognitive theories underlying our method. This is followed in Chapter 3 by a description of the ontology selection and experimental techniques selection. In Chapter 4 we discuss the overall experimental design of the multi-method evaluation methodology. The next two chapters, 5 and 6, independently present and discuss the design, procedure and results of each experimental technique. The thesis closes with a general discussion of the overall results and their implications (Chapter 7), possible limitations that require further investigations (Chapter 8), and finally some interesting areas to be explored in future studies (Chapter 9).

Chapter 2

Literature review

2.1 Quality and evaluation of Ontologies

Formal ontologies have two main purposes. They represent a specific domain and they are used as computational entities for reasoning purposes. In this study we focus only on the representational aspect of an ontology.

To study the representational aspect of an ontology, we use existing quality frameworks as guidelines. Since conceptual models share the same purpose with ontologies that they are also representations of a domain, we can apply existing frameworks for model quality as in this study we view ontologies primarily as descriptions of a domain.

The Guidelines of Modelling (GoM) [83] list construction adequacy as one of six quality principles; along with the principle of language adequacy, economic efficiency, clarity, systematic design, and comparability. Construction adequacy concerns the relationship between the (explicated) model and the model developer's views of the domain and is described as a necessary condition for a good model. This principle explicitly recognizes the fact that the relationship between a model and the domain is mediated by the modeller and his/her mental model of the domain. The rationale for not considering the other five quality principles is explained later in this section. Furthermore, the framework in [46], an extension of [48] which recognizes only a direct relationship between the domain and the model as semantic quality, sees a human component to model quality - although only in model interpretation; not in model construction. Rather than examining the semantic quality of a model, the *perceived* semantic quality concerns the relationship

between a modeller and his/her mental model of the domain and the model. Thus, existing quality frameworks appear to recognize the relationship between an explicit description and mental models of a domain as important for model quality. However, these frameworks and in particular, this quality dimension, has received little attention in the ontology literature.

The need for ontology evaluation-methodologies has become prominent as ontology evaluation is a keystone to ensure high-quality and representative ontologies [20, 31]. However, quality is a judgment rather than a property of something, and different stakeholders have quite different views on quality. The quality characteristics generally involve recognition of design tradeoffs, in particular the interaction of adequacy for human cognition (principle of uncertainty, ability to distinguish alternatives, ease of learning and so on), with technical factors (principle of variety, control of redundancy, implementability, reusability and so on) [20].

The quality of ontologies has been examined from various aspects. Three main types of measures for ontology evaluation were identified: structural measures, that are typical of ontologies represented as graphs; functional measures, that are related to the intended use of an ontology and of its components, i.e. their function; usability-related measures, that depend on the level of annotation of the considered ontology [31]. The focus of ontology evaluation has been on the technical factors.

One aspect of quality examined is the usefulness and usability that is determined by the appropriateness of the description language and the availability of the software tools for its manipulation and use [50, 90]. Lozano-Tello and Gomez-Perez proposes OntoMetric, an adaptation of the Analytic Hierarchy Process, i.e. a mathematical method for scaling priorities in hierarchical structures [50]. The main goal of this method is to help choose the appropriate ontology for a new project. The functions supported by OntoMetric are the ordering by importance of project objectives, the qualitative analysis of candidate ontologies for the project, the quantitative measure of the suitability of each candidate. The application of OntoMetric can only follow ontology release. The method is meant for users types like Engineers or Project Managers who need to look for ontologies over the Web at the purpose of incorporating them into their systems. Therefore, OntoMetric makes itself useful as a support to the evaluation of the relative advantages and risks of choosing an ontology over others. The main drawback of OntoMetric is related to its usability - specifying the characteristics of an ontology is complicated and time consuming, and assessing its characteristics is quite

subjective. On top of this, the number of use cases is limited, which is an important obstacle to defining (inter-subjective or objective) parameters based on a large enough number of comparable cases.

Another aspect of quality is the metaphysical and logical properties of a good ontology which are specified in the OntoClean method [40]. This method is meant for application at the premodelling and modelling stages, i.e. during ontology development. The main goal is to detect both formal and semantic inconsistencies in the properties defined by an ontology. The main function of OntoClean is the formal evaluation of the properties defined in the ontology by means of a predefined ideal taxonomical structure of metaproperties such as rigidity, identity and unity.

Spyns presents EvaLexon which finds application at the premodelling/ modelling stage [88]. Its main goal is to evaluate at development time ontologies that are created by human beings from text. In sharp contrast with OntoClean, EvaLexon is meant for linguistic rather than conceptual evaluation. Its main function is the measurement of how appropriate the terms (to be) used in an ontology are. A term is judged more or less appropriate depending on its frequency both in the text from which the ontology is (being) derived and in a list of relevant domain specific terms. Regression allows for direct and indirect measurement of the ontology's recall, precision, coverage and accuracy.

Finally, a linguistics-based approach partly comparable to EvaLexon is proposed to evaluate ontologies with respect to three basic levels: vocabulary, taxonomy and (non-taxonomic) semantic relations [64]. The functions proposed are based on two key arguments: the task and the gold standard [64]. The task needs to be sufficiently complex to constitute a suitable benchmark for examining a given ontology. The gold standard is a perfectly annotated corpus of part-of-speech tags, word senses, tag ontological relations, given sets of answers (so-called keys) used to evaluate the performance of algorithms that are run on the ontology to perform the task.

One aspect that the ontology literature has not focused on is the representational capabilities of ontologies, which is said to be one of the main purposes of ontologies and an important quality characteristic. The focus of evaluation in our study is based on the principle that an ontology, as a formal description of a domain, must conform to the way in which the domain is perceived and understood by a human observer [38, 25, 34], that is, its conceptualization. We term this *cognitive ontology quality*. Our notion of cognitive quality is related to construction adequacy [83] or perceived se-

mantic quality [46]. We compare the domain understanding that is formally specified in an ontology with the conceptualization - the domain understanding that is held in human cognitive structures. Hence, we use techniques from cognitive psychology for the evaluation of ontologies, described in the following sections. An initial development of such an evaluation approach was presented in [28, 27, 26], and the sentence verification technique, which will be described in the latter sections, was used in these studies.

Note that we propose to evaluate only one aspect of quality among many; other aspects of quality discussed in this section, including the principles in GoM [83], are of course important as well. Providing an integrated account of quality is beyond the scope of this thesis.

2.2 The role of cognition in ontology construction

Ontology within IS research as described in Chapter 1 is "a set of concepts and their relationships" ([25], p.150) which is said to represent what is perceived in the world. In this study, the quality dimension that we evaluate the ontology on, is the agreement between the formal specification of a domain in an ontology and the conceptualization of the domain by human beings.

Evermann's review of previous literature suggests that ontology-based IS research has the assumption that the elements in the world are universally known or knowable [25]. However, the studies of perception, meaning and language, propose a different viewpoint - knowledge of the world is neither immediate nor universal. It was argued that our perception of reality is structured by the physiological sensory apparatus, such as auditory and visual, and is shaped within our brain [47, 37]. As an individual's experience of perceiving reality accumulates, knowledge structures specific to the individual are formed. In turn, the interpretation of perception is influenced by individuals' world views and shaped by their cognitive/knowledge structure [25]. Instead of immediate and universal knowledge of reality, cognitive concepts structure the way we perceive and interpret the reality in the world [47, 57], and the organization of concepts is knowledge-based and driven by theories about the world [16, 32, 33, 69]. Therefore, knowledge concepts of categories can be used to study ontologies. In support of this, classification, like that used in constructing ontologies, should be guided by cognitive principles of

using and accessing of pre-existing knowledge structures, rather than being evaluated systematically, since it is intended to represent human knowledge of a domain [63].

An evaluation methodology that appropriately and accurately examines the correctness of the classifications of the concepts in an ontology should examine the appropriateness of the knowledge representations of an ontology with respect to our cognitive structures.

In the development of the evaluation technique, one important source of empirical study is cognitive psychology research, which can be used to study representational aspects of ontological structures. In the following sections, we study the role knowledge access plays in examining ontological structures. We introduce two theories proposed for accessing knowledge structures, and subsequently propose evaluation techniques for ontological structures.

2.3 Accessing knowledge structures

Humans have an enormous amount of knowledge that needs to be quickly and efficiently searched. Much information has been stored about each of the concepts in the world, for example, its relation to other concepts, its syntactic class, and its phonological form. Psychologists typically use experimental approaches to understand how humans represent and access such vast knowledge bases. Two widely used frameworks of conceptualizing knowledge representation are the hierarchical structure of knowledge representation [18], and the spreading activation framework [17]. The existing cognitive research on categorization is based largely on these two frameworks.

2.3.1 Hierarchical structure

Quillian suggested that an economical way to store large amounts of information about concepts/categories would be to assume a hierarchical structure [67, 68]. Evidence that supports this view was later reported in Collins and Quillians' study [18]. Knowledge is a hierarchical network of stored information. The network describes each concept, such as "animal" and "fish", as "nodes", and describes how the concepts are connected. For each concept, the properties of a concept are stored with the concept. This model proposes that in our mind, concepts are organized in the form of hierarchies with the superordinate concepts at the top, the subordinates which represent more

specific concepts at the bottom, and basic level concepts categories in the middle (see Figure 2.1). Two important theoretical concepts, semantic distance effects and category size effect, are based on the hierarchical network model [18, 19].

Semantic distance effects Semantic distance effects look at the effect of the distance between two concepts, that is, the number of links from one concept to the other on people's conceptual processing. Collins and Quillian found that the decision times for verifying a sentence that consists of a concept and its superordinate concept, such as "A robin is a bird", vary directly with the number of levels separating the two nodes in the sentence [18]. For example, referring to Figure 2.1, it takes longer for subjects to verify the sentence "A robin is an animal" than to "A robin is a bird", because "animal" is semantically more distant to "robin" (two nodes), than "bird" is to "robin" (one node).

Category size effect The second theoretical concept, category size effect, refers to the relationship between the reaction time (RT) and the size of category. The results from the semantic verification experiments show that the larger the category, the longer the time required for search. For example, because the concept "animal" consists all instances of "bird", as well as all instances of "fish" and instances of other category members of "animal", it is a larger category than its category member 'bird'. Therefore, the cognitive space the subjects have to search through to determine if the word "robin" belongs to the category "bird" is less compared to the category "animal" [18] (refer to Figure 2.1).

2.3.2 Spreading activation

The spreading activation framework proposed by Collins and Loftus is a similar way of conceptualizing knowledge representation [17]. The framework is a fundamental memory retrieval mechanism developed within network theory [2, 17, 18], and has been widely used as an explanation for category structures and as a search mechanism [3, 49, 58, 52]. The authors proposed a network of interconnected nodes representing conceptual information. Retrieval of information or concepts from this network involves activation of a node, and this activation spreads along the pathways through the network to

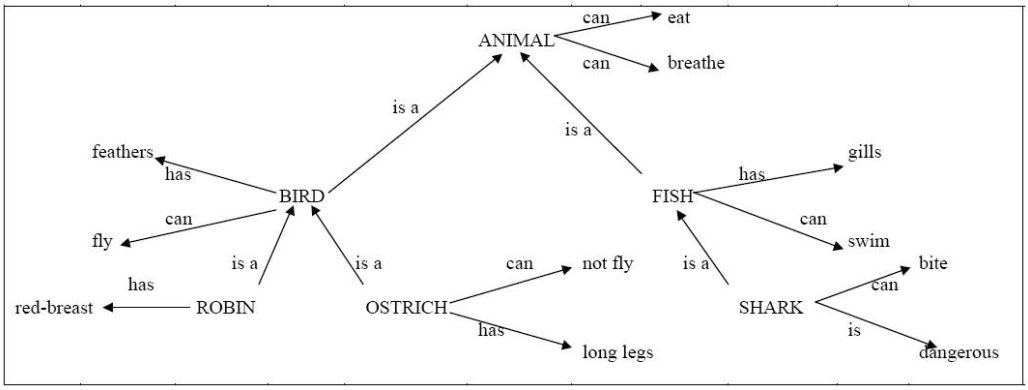

Figure 2.1: Collins and Quillians' cognitive model of semantic representation (from [44])

other related nodes (related areas in memory). For instance, based on Collins and Quillians' cognitive model of semantic representation (see Figure 2.1), it would be easier to recall "shark" after being given the word "fish" than after being given the word "ostrich". This is because shark and fish are semantically more closely related. This spread of activation allows the related areas of the memory network to be more quickly available for further cognitive processing.

The spreading of activation is automatic as opposed to being under conscious control [4, 6, 58]. People tend to define attributes or concepts by the already existing category system in a culture at a given time. For example, because we already have a cultural and linguistic category "bird", we do not perceive the attribute "wing" only as separate object, but rather, as a part of a bird's body. Thus, the concept of wings co-occurs with feathers more than with fur. It implies that an association made among concepts is not limited to their physical resemblance, but is also made on the concepts' semantic relation. This framework is thus applicable to ontologies that describe abstract concepts, such as upper level ontologies.

One important property of the activation process is that since concepts are assumed to be associated within a network of associations, activation may spread not only to directly related concepts but also from those concepts to concepts further in the memory network, that is, multiple steps within the network. The "multiple step" assumption of spreading activation theory [7] has been particularly important in accounting for category verification

response latency. For example, the time taken for subjects to recognize the word "shark" would be shorter after presenting to the subjects the word "fish" (a concept directly related to "shark") than with the word "animal" (a concept one step further in the network from "shark") (refer to Figure 2.1).

The concept of spreading activation is important for this study on ontological categorization and the development of evaluation techniques. The theory implies that people, when categorizing concepts, base their categorizing decisions on the degree of association (the number of nodes to traverse) with other available concepts. The degree of association can thus serve as a measure for the evaluation of ontological structures. Spreading activation is the underlying mechanism involved in tasks such as category exemplar production [49], semantic priming in lexical decisions [58, 7], sentence verification [49], episodic sentence and word recognition [3], and perceptual word recognition [52].

The two effects (semantic distance and category size) and the spreading activation framework can reinforce each other in certain situations. This is the case for category subsumption statements. Verifying that "A robin is an animal" is slower than verifying "A robin is a bird" because the animal concept is semantically more distant hierarchically and in terms of network theory from the robin concept than is the bird concept. This is also because the animal category contains many more prototypical instances to search than the bird category. We use these effects in our proposed method to determine cognitive ontology quality.

Chapter 3

Methodology

Fang made an initial attempt at developing an ontology evaluation methodology that tests how well a specification of an ontology corresponds to the ontology users' conceptualization of a domain, and whether the ontological structures reflect the perceived reality of the world [28, 27, 26]. A sentence verification technique adopted from cognitive psychology (described in Section 3.2.2) was applied to compare two upper level ontologies (the most general and abstract form of ontologies), SUMO and BWW [28, 27, 26]; and in Evermann and Fangs' study [26], SUMO and WordNet (described in Section 3.1.2) were also examined. These studies were exploratory. The findings suggested that none of the three ontologies properly represents the perceived world - It generally takes longer to verify a sentence statement associating concepts that are, according to the examined ontologies, less semantically distant than a sentence statement associating concepts that are more semantically distant. This is contrary to what is known from cognitive psychology studies.

It is important to examine whether this technique is a valid measurement of the applied ontologies. In this thesis, we aim to develop an evaluation method with a more rigorous design to ensure the validity of the measure, taking into account factors that may lead to skepticism of the applicability of the method on evaluating ontologies. Some limitations found in the initial studies [28, 27, 26] that need to be controlled are: First, the concepts selected for evaluation should not be too abstract since the technique was used on studies with more concrete concepts built around the basic level (the concept of basic level will be explained in Section 3.2.4). BWW, for example, might therefore not be a suitable choice of ontology for the purpose of this study

if sentence verification task is selected as an evaluation technique. This is because the concepts contained within it are all quite high on the abstraction level. Second, compound terms were included as stimuli. The reason against including compound terms is discussed in Section 4.1. In this study, the development of this evaluation method ensures comparability of the domain focus and abstraction level of concepts selected for evaluation. Other criteria will be discussed in Section 4.1.

This study is intended to develop a multi-method ontology evaluation methodology, based on principles of cognitive psychology, in order to evaluate ontologies through human cognition. The motivation of the multi-method approach is that it allows us to have more confidence in the findings, assuming the results from both evaluation techniques converge.

To ensure the validity of the evaluation methodology, we examine the applicability of the two evaluation techniques. The first step is to select two appropriate ontologies and two appropriate techniques suitable for adoption (to be discussed in this chapter). The second step is to replicate two of the psychology studies where the techniques were used, for validation of each of the selected techniques. We then evaluate the SUMO ontology and WordNet with the two experimental techniques conducted independently, and together as a demonstration of the multi-method evaluation methodology (see Chapter 5 and 6 for the replication and test trials of these techniques).

3.1 Selection of ontology

The ontologies selected should:

- Be well accepted and widely known. This is so we can assume correct ontological structures and specifications.
- Include concepts with overlapping domains of interest and common levels of abstraction, if not the same. This is so we can have concepts selected from two ontologies that are comparable for evaluation.
- Represent concepts with clear subset (categorical) relationships. To ensure that the sets of concepts selected for evaluation are not subject to the bias of being related by other types of relationships, the sample frame of sets of concepts should be limited to those that are related only by subset relationships, and not by other types of relationships, such as object-property relationships.

In this study, the SUMO ontology and WordNet are selected for evaluation. In Fang's study [28, 27, 26], the BWW and SUMO ontologies were used, and it was found that BWW contains only concepts that are on the high level of abstraction, which has been criticized in [79]. BWW was also disadvantaged by the large proportion of concepts interlinked by other types of relations. WordNet is not designed as a formal ontology, it is a thesaurus that builds primarily on hypernym/hyponym relationships between terms (not concepts). WordNet provides a level of abstraction that is more concrete than the BWW concepts, and arguably more comparable to the SUMO ontology in terms of the level of abstraction and domain focus [26].

3.1.1 Suggested Upper Merged Ontology (SUMO)

The SUMO is an upper-level ontology created by Teknowledge Corporation as a starter document for The Standard Upper Ontology Working Group with the input from the SUO email list. The working group is IEEE-approved and consists of experts from the field of engineering, philosophy, and information science. SUMO is written in the SUO-KIF language. It merges existing ontologies such as John Sowa's upper-level ontology [86] and Russell and Norvigs' upper-level ontology [81] into a single, comprehensive, and cohesive structure that is now considered the largest formal public ontology in existence. For example, it is widely accepted by the knowledge representation community [35]. The ontology is still growing; it contained 654 terms and 2351 assertions by June 2001 [59]. It is mainly used for research and applications in search, linguistics and reasoning.

3.1.2 WordNet

WordNet is a semantic lexicon for the English language. The goal of WordNet was to develop a system that would be consistent with the knowledge acquired over the years about how human beings process language. It groups English words into sets of synonyms called synsets; providing short, general definitions, and it records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of a dictionary and a thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. The database and software tools have been released under a BSD style license.

WordNet was created and is being maintained at the Cognitive Science Laboratory of Princeton University under the direction of psychology professor George A. Miller. Development of WordNet began in 1985, and over the years, the project received about 3 million dollars of funding, mostly from government agencies interested in machine translation. As of 2006, the database contains about 150,000 words organized in over 115,000 synsets for a total of 207,000 word-sense pairs; in compressed form, it is about 12 megabytes in size.

3.2 Selection of experimental techniques

This section examines five techniques that can potentially be used for the development of the evaluation technique. For each technique, the appropriateness for our study is evaluated.

The sentence verification technique and the priming technique have been used to study knowledge representation, which is the purpose of ontologies. The sentence verification technique was used in studies for both the hierarchical network model [18] and spreading activation [49]. The priming technique was often used in the study of spreading activation [58, 7]. They are two of the most frequently used techniques for psychological research in the study of categorization. Categorization is a fundamental aspect of knowledge representation [33] which concerns the grouping of like items based on people's pre-existing knowledge about the perceived world [73]. The categorization studies that utilized these two techniques provide us with templates on the use of the techniques, thus allowing us to derive well constructed ontology evaluation methodologies. The only concern for the adoption of these two techniques is that there are no previous studies that evaluate ontologies using the theories of knowledge representation. The studies of knowledge representations (e.g. [18, 7]) use objects of natural and basic level categories (e.g. bird, animal) within hypothesized memory network models, rather than the more abstract ontological concepts (e.g. objects, world). However, if, as they claim to do, ontologies represent real concepts, this should not affect applicability of these techniques.

Priming and sentence verification techniques assume automatic processing of activation which test an implicit and subconscious mental representation of one's perception of a concept, and rely on implicit and reflexive measures such as reaction times. We further examine three techniques that

utilize attentional processing over which we exert conscious control involving testing conscious, explicit knowledge and the memory of individual concepts.

The following two evaluation techniques that utilize attentional processes are based on categorization concepts. Since ontologies claim to describe how people see the world, or at least an application domain, the theories that apply to cognitive category structures can be assumed to be applicable to ontological structures. The two frequently used techniques in the studies of categorization that appear suitable for the purpose of ontology evaluation are learning, which is based on the concept of prototypicality, and shared attributes which is based on the principle of inclusiveness. Descriptions of these categorization concepts will be discussed in Section 3.2.3 and 3.2.4. Finally, the recall technique which will be discussed in Section 3.2.5 is based on the concept that the strength of the semantic relationships between concepts can be used to construct elaborations about the to-be-remembered material [14] and these elaborations can mediate later recall [51, 66].

3.2.1 Priming

Priming, after a few modifications since it was first used by Beller [10], is effective as a means of "investigating the nature of mental representations" ([75], p.304). The method of priming is based on testing whether advance information about a category name (prime) facilitates or inhibits responses to the matched target category name. The logic of this technique is that "the prime can only facilitate a match when it makes possible the generation of a mental code which contains within it some of the information needed to make the match" ([75], p.304). This technique is based on the idea of spreading activation [17].

Beller, as the first to use the priming technique, primed the subjects with a letter, and two seconds later presented a pair of letters [10]. There were two conditions: subjects were to select the response "same" if the pairs of letters are physically identical (e.g. A – A); in the other condition, when the letters possess the same name (e.g. A – a). The author found that facilitation by the prime letter was found in both conditions. This suggests that subjects were not simply retaining a literal representation of the presented letter but were generating an abstract expectation or representation which did not depend on case [75]. Therefore, the priming technique should be applicable to our study, since priming takes place where abstract concepts are concerned.

Building on Beller's study [10], Rosch found that the technique of ad-

vance priming in a matching task has consistently yielded the result that the selection of the response "same" to pairs which belonged to the same superordinate category (e.g. robin – sparrow), is facilitated by advance presentation of the category name (bird). Whereas response to poor examples (e.g. penguin – turkey) of the category is inhibited by advance presentation of the category name (bird) [74]. Similar results were found in Rosch's study of colour [75]. Priming is of direct relevance both to the nature of colour categories and to the nature of cognitive representations; the facilitation and inhibition effects were observed in both cases.

We can base our methodology on the research procedures in Rosch's studies [74, 75]. Bad category members (ontological concepts) can be detected by a slower response time (inhibition) as proposed by Antos [5] and Rosch [74]; or when "different" was selected as a response to the question whether pairs of items are members of the same category. This technique can also test the order of the level of category structure based on category size effect, semantic distance effects [18], and degree of association [7]. For example, a longer RT (response time) in giving the "same" or "different" response to the presentation of "robin" can be observed when the prime belongs to a superordinate category on the more distant level (e.g. animal) than on a less distant level (e.g. bird) (refer to Figure 2.1).

There is, however, a limitation in this technique. Each of the selected ontological concepts requires at least two sub-ordinate category members for the measurement of distance effects; this adds limitations to the selection of concepts from the ontologies to be evaluated.

3.2.2 Sentence verification tasks (SVT)

Rosch noted that the speed with which subjects can judge statements about category membership is one of the most widely used measures of processing in semantic memory research within the human information-processing framework [78]. Subjects are typically asked to verify a large number of true and false category statements, for example, "A robin is a bird" (true), and the response time to each of the statements are measured. These are referred to as sentence verification tasks. This technique has mostly been used to determine how words are stored in semantic memory, particularly the memory about word meaning. Studies that use natural language categories (e.g. [70, 77]), modify the task to study the typicality effect by changing some of the statements to include members that are considered typical to

the category. They found that the time taken was shorter for the statements with items that were rated more typical to the category than those rated less typical. Thus, experiments using this technique imply that more typical members of a category are more accessible. This means a good category would be a typical member of its superordinate category, and statements that exhibit short verification time can be identified as an indication of a good category structure.

SVT was also used in the studies of the hierarchical network model. The verification process starts when a node in the semantic network that corresponds to the subject term (e.g. "robin" in the statement "A robin is a bird") is triggered, and then the searching for the predicate term "bird" starts, from the next higher hierarchical level of the subject term's level. As soon as the predicate term is found, the search stops and the response "true" is selected. If it is not found, the search moves up the hierarchy to the next level [44]. Each move to the higher level consumes verification time, thus the "semantic distance effects". Collins and Quillians' study shows a linear relationship between verification time and number of levels of approximately 75 milliseconds [18]. This semantic distance between the subject term and the predicate term is also an indication of the "degree of association". Moreover, the higher the level a category is in the hierarchy, the more instances there are in the category to search; thus taking a longer time (category size effect). Both, semantic distance effects and category size effects, were supported by this particular measuring technique in the studies of [18, 19].

Barres and Johnson-Laird [8] and Keenan [44] emphasize that these predictions only apply to the true statements. Collins and Quillian have explored the possibilities for the falsifying statements, but concluded that there are so many different ways to achieve falsification so they preferred not to offer any predictions for false statements [18]. Barres and Johnson-Laird [8] explain that there is no direct "route" in our mental model to falsification.

In this study, we can examine the two ontological structures by making statements about category memberships of any two concepts in each ontology, and ask subjects whether each statement is true or false. Only response time data from the true statements are used. For statements where the true category membership of two concepts was responded to as "false", the particular membership of the subject term may be problematic, signalling an error in the ontological structure. Based on the concepts of the category size effect, semantic distance effects, and degree of association, the correctness of the hierarchical order among three concept words can be examined by

comparing the RT of two statements constructed from these three words (construction of sentence statements will be described in Section 4.1).

This technique allows experimenters to examine several aspects concerning the evaluation of ontologies, and provides empirical evidence and explanations for each finding. For example, error responses about a statement imply a problematic categorization in the structure. Moreover, data such as response error rate and verification time can be generated from each statement asked; this data can be used to examine the assumptions made in the study, and be analyzed for different purposes. For example, we can present the overall quality of ontologies, examine the correctness of structure order of concepts from each statement. However, the subjects in the experiments will have to go through many trials because of the inclusion of the false statements - although they do not have much analytical power for evaluating ontological structures, they are needed to ensure that participants will verify each statement carefully. The RT of participants' responses might be biased if participants have the impression that the majority or all of the statements are true statements.

3.2.3 Learning

The rate of learning new information and the order in which new information is learned are considered two of the most pervasive measuring techniques in psychological research.

This measuring technique is based on the concept of "prototypicality". It is suggested that there are no rules for the inclusion of instances in categories, nor is there any clear cut boundary to categories [78]. Instead, people's judgment about a category membership is often determined by the prototypicality of that particular member. Rosch describes the prototypes of a category as the clearest set of members that are judged by people to "most reflect the redundancy structure of the category as a whole" ([78], p.37). That is, prototypes of categories appear to have the maximum category resemblance of the attributes of categories within categories. The more prototypical a category member is, the more common its attributes are with other members of the category, and fewer common attributes it has with members of different categories [76]. Thus, a commonly agreed typical fruit, apple, is similar to many other fruits; and very dissimilar to exemplars of other categories such as vegetable; whereas an atypical exemplar of fruit such as tomato, shares similarities with both concepts [9].

The concept of prototypicality is not limited to the physical attributes of objects [72], but also applies to the semantic attributes of objects [74, 75]. The concept is therefore applicable to this study of ontology evaluation.

In the experiments conducted by Rosch and Mervis [76], artificial categories were constructed with a differing degree of prototypicality of the items in the categories, or the amount of attribute overlap between categories. The formation of a category structure affects the rate of learning of category items: the more prototypical the items in a category are, the faster it is for the subjects to learn the category. It also affects the RT in judging category membership once the categories were learned. The same technique was also used in another work by Rosch [72] in which prototypicality was found to be a predictor of learning of the categories. Alternatively, the quality of category membership can be examined by testing the order in which category items are learned - the items that are rated as more prototypical of a category tend to be learned before less prototypical items [74, 76, 77, 78].

As proposed above, examination of the rate of learning, RT in judging category membership, and the order in which category items are learned, are all useful in judging the quality of the category membership of each concept in each of the ontological structures. There are, however, some limitations to this technique that experimenters have to take into consideration. First, the study of rating or order of learning is often coupled with other techniques such as rating of prototypicality of items [78] and verification task [76]. It appears that learning by itself does not explain much of what happens when examining a category structure. Second, it is difficult to control for possible mortality effect[1], history effect[2], and maturation effect[3] which could occur between the time of learning and that of recalling conducting this measuring technique. Third, many ontologies offer only category structures, rather than instances or members of categories, which are required for the application of the learning technique. Fourth, the technique has been used in studying the features (prototypicality) of stimuli. There is no empirical evidence to support the applicability of this technique on stimuli that are not concrete objects.

[1]The loss of subjects from comparison groups.

[2]The events occurring between the first and second measurements in addition to the experimental variable.

[3]The process of maturing in the subject during the duration of the experiment which is not a result of specific events, for example, subjects growing more experienced in the experiment.

3.2.4 Shared attributes

Rosch argues that categories within taxonomies are structured such that there is one level of abstraction at which the most basic categories are placed [73]. In general, the basic level of abstraction in a taxonomy is the most inclusive level at which there are attributes common to all or most members of the category (e.g. chair, car). The more abstract the level is, the less attributes its members share [78]. For example, categories one level more abstract will be superordinate categories (furniture, vehicle) whose members share only a few attributes among each other. In addition, categories below the basic level are subordinate categories (e.g. kitchen chair, sports car) which also share many attributes, but contain many attributes which overlap with other categories. For example, a kitchen chair shares most of its attributes with other kinds of chairs.

We can utilize this finding in this study to examine the correctness of the level of abstraction of each category in each of the two ontological structures, by asking subjects to write their perceived attributes for category members of each concept that is situated at different levels. The greater the attribute overlaps, the more likely the category belongs to the lower levels of abstraction.

This technique allows us to directly examine whether each concept is placed on the correct level of an ontological structure. However, there are several limitations to this technique.

First, this technique examines the correctness of the structural level of each concept by counting the number of common attributes among its category members, and the number of common attributes among the concepts one level up the hierarchy. Fewer shared attributes in concepts of the more abstract level (higher level) indicates the concept is situated on the correct structural level. However, this evaluation can only be achieved with the assumption that every category member listed under a concept is indeed its correct categorical sub-ordinate. This is because one incorrect category member would generate attributes that are different from other category members and have fewer common attributes with other category members. This would thus lower the value of the number of shared attributes for the concept examined, resulting in an incorrect evaluation of the structural level of the concepts. Incorrect category members can potentially be identified by finding members that have significantly fewer attributes in common with other category members. However, there is not yet a standardized measure

that allows us to determine the number of shared attributes appropriate for different levels of abstractions.

Second, this technique appears to be more applicable to lower level ontologies that deal with concrete concepts and those with clear category structures. This is because concepts in upper level ontologies are often abstract. It is possible that some concepts, especially those on the top levels of these ontologies, may not be examinable, as these top level concepts can be too abstract to have any shared attributes.

Third, this technique is only applicable for concepts above the basic level. Even though it was said that the greater the attribute overlaps, the more likely the category belongs to the lower levels of abstraction, there is no empirical evidence that this rule applies when the examined category is positioned below the basic level. Thus, the correctness of the level structure of concepts below basic level should not be evaluated using this technique. Moreover, to control this potential bias, an additional examination is needed - the basic level of abstraction of the ontologies has to be identified prior to the evaluation.

Finally, this experimental design requires a lot of writing for the participants, and therefore runs a high risk of subject fatigue effect[4].

3.2.5 Recall

Semantic relationships among to-be-remembered items facilitate memory for those items. The finding that strongly related materials are better remembered than weakly related materials is ubiquitous. This result has been found in paired-associate and word-list learning [23, 55, 94, 99]; cued recall, free recall, and recognition testing [41, 51, 54, 93]; and when nonsense syllables or words are the to-be recalled items [66, 71]. Further, these results occur with many different definitions of semantic relatedness. These measures include associative strength as determined in free-association norms, number of associates given to an individual item, frequency of items in the language, and experimenter-defined relations such as taxonomic category and congruity [22, 41, 82, 94]. In this thesis, we study the information retrieval of memory by focusing on the subset relationship aspect of semantic relatedness (see Section 4.1 for description) supported by the theories of spreading activation, semantic distance effect and category size effect.

[4]Change in subject's response behaviour due the fatigue after a series of trials.

The theoretical explanations of the results center on the idea that semantic knowledge transfers to episodic tasks. Strongly related materials permit more positive transfer from previous learning than do weakly related materials; this knowledge can be used at study to construct elaborations about the to-be-remembered material [14] and these elaborations can mediate later recall [51, 66].

Despite this persuasive theoretical rationale, there are reasons to expect that weakly related materials will be better remembered than strongly related materials. Both humans and animals respond very strongly to unexpected or novel stimuli [87, 98] and weakly related materials often represent unexpected or novel semantic combinations.

If subjects respond to the unexpected semantic combinations represented by weakly related materials as they do to other unexpected stimuli, weakly related materials may enjoy a memory advantage over strongly related materials. There are two plausible ways in which this could occur. First, the response to unexpected stimuli may encourage a more elaborate encoding of weakly related materials, and, second, a memory of the response to unexpected stimuli may mediate later recall of the weakly related materials.

When subjects attempt to encode pairs of words for later recall, they attempt to semantically relate the items in the pair. This can be conceived of as an attempt to search the attributes of both items in semantic memory to find attributes common to both items [18, 13]. When such attributes are found, these constitute a relationship between the two items, and this relationship is stored as part of an episodic memory. When subjects fail to find such a relationship, a process hereafter referred to as a blind-alley search, a response akin to a surprise or novelty response occurs; expectations are violated.

The theoretical explanation of the expectation-violation effect claims that a failure to understand the relationship between the items in a word pair can improve memory performance on that word pair. These failures, which are called blind-alley searches, occur when the items in word pairs represent unexpected or novel semantic combinations. The blind-alley search results in a memory representation, a blind-alley search cue, which can mediate the later retrieval of the word pairs on which the blind-alley search is committed. The expectation-violation effect occurs because weakly related pairs are more likely to represent unexpected or novel semantic combinations than are strongly related pairs. Subjects are thus more likely to commit blind-alley searches, with their attendant memory benefits, on weakly related pairs than

on strongly related pairs.

Hirshman and Bjork reported an experiment in which weakly related word pairs were better remembered than strongly related word pairs [43]. The authors presented subjects with lists of word pairs. These lists contained both strongly related and weakly related word pairs where strength of relationship was defined by the number of times a response was given to a stimulus in a free-association task. They found that responses from weakly related pairs were better recalled than responses from strongly related pairs. This result occurred when subjects read or generated the response terms at study. It did not occur in cued recall; in cued recall, responses from strongly related pairs were better recalled than were responses from weakly related pairs. Hirshman and Bjorks' study was later replicated [42] and its findings were supported. In this study, we can construct a free-recall condition and the semantically distant ontological concepts should be more frequently recalled than the semantically more closely related concepts. We can also examine the quality of the ontological structure by constructing a cued-recall condition in which the expectation is the opposite of that in the free-recall condition.

Even though the semantic relationship tested for in this study is limited to the subset relationship, the effect of this type of relationship is well established [18] and therefore the theories above (the assumption we based on) should be highly applicable.

3.2.6 Evaluation technique summary and selection

An important purpose of developing a multi-method approach is to better ensure the validity of the findings. The evaluation techniques to be adopted into the multi-method methodology should be reasonably different. This is to reduce the risk of having the converging findings interpreted as being due to the similarity between the two techniques adopted. For example, although both the priming technique and the SVT technique are feasible options for adoption in this study as there are no unmanageable limitations to the techniques, they are similar approaches. They both measure the RT based on the strength of associations between concepts; both have similar experimental designs (computer based); both are based on the same theories and principles; both involve implicit and reflexive mental processes.

It has been suggested that there is an operational distinction between automatic processes and attentional processes [65, 91]. Since the ontological concepts we evaluate in this study are different from the concept stimuli used

in the previous psychology studies on many aspects (e.g. level of abstraction and domain focus), we cannot ignore the possibility that different processing pathways are used for the evaluation of the ontological concepts. Thus, we select two techniques that utilize different mental processes.

We choose the SVT over the priming technique to be our first evaluation technique because it has been used and studied in [28, 27, 26], and is therefore a more mature technique for this study. The SVT assumes an automatic process of the verification task. We therefore choose the second evaluation technique from one of the following - the learning technique which examines whether the ontologies effectively represent people's perception of the world by studying the influence of the level of familiarity to prior existing knowledge on conscious encoding, or the recall technique which examines conscious retrieval of qualitative event information. We disregard the shared attributes technique as it has several apparent and significant limitations that are likely to hamper the validity of the findings of this study if it were to be adopted.

We choose the recall technique over the learning technique because the recall technique has been used to study words and association strength of word pairs, whereas stimuli used in the learning technique are often constructed artificial categories [76]. Furthermore, ontology evaluation examines people's existing knowledge representations. Studies that utilized the learning technique often study subjects' abilities to learn new information during the study phase of the experiment, rather than study the effect of level of acquired prior knowledge on abilities to memorize and recall like that in the recall technique.

We adopt SVT and recall as our evaluation techniques for the multi-method methodology. Although there are a few limitations in the adoption of these two techniques, they are manageable with careful design.

To ensure the validity of the two evaluation techniques selected, for each technique, we conduct a replication of a psychology study that utilized the technique. Research designs for the two evaluation techniques in this study are based on the psychology studies chosen [70, 42]. In Appendix A and Appendix B, we outline the experimental designs and the findings of these psychology studies. The experimental designs and results of the two replication studies are presented in Chapter 5 and Section 6.2.

The design and development of the two evaluation techniques selected are discussed in Chapter 4 and then individually in Chapter 5 and 6.

Chapter 4

Research design

In this chapter, we describe the overall experimental design for the multi-method evaluation methodology. Methodological aspects specific to the two experimental techniques are described in Chapter 5 and 6.

4.1 Stimuli selection and design

This section describes the general stimulus selection criteria for the multi-method study.

For a valid comparison of the quality of ontologies, it is important that the domain focus and the abstraction level of the concepts selected from each ontology for examination are comparable - that they are similar and compatible.

The SVT and recall techniques are adopted from psychology studies. The application of these techniques has been limited to the examination of natural categories which are generally concepts of and close to the basic level of abstraction. To ensure the validity of the multi-method methodology, the concepts we select for evaluation should be within both of the techniques' application domain - concepts/categories of, and close to the basic level. The replication studies presented in Chapter 5 and Section 6.2 verify the applicability of the techniques to the concepts/categories of, and close to the basic level. If the findings of the replication studies are significant and consistent with previous studies, the techniques are assumed to be applicable to ontologies with concepts on the lower level of abstraction (natural semantic categories), and are valid techniques to be adopted in this study.

It is also important to select ontological concepts with comparable domain focus. This is because people are likely to process concepts from different domains of interests differently, hence different evaluation outcomes. For instance, the verification time for the truth of statements about abstract concepts is likely to be longer than that for the truth of statements about physical concepts. Therefore, comparisons of ontological quality using this evaluation method require that the ontologies are competing ontologies with overlapping domains of interests and cover concepts around the basic level of abstraction.

The multi-method methodology acts as a cross-reference tool. Since the evaluation is based on the level of convergence of the findings of two techniques, we use the same sets of concept stimuli for both the SVT experiment and the recall experiment.

For the recall experiment, it is important that we present our stimuli in the form of sentences rather than word pairs as used in previous studies [43, 42]. This is because in this study, we examine only the subset relationship between concepts. Using sentence stimuli such as "A robin is a bird" minimizes the risk of subjects making other inferences about "robin" and "bird". Thus, the same stimulus list is used in the SVT and recall experiments.

In this study, we examine sentences that state subset relations (S sentences). We use *S1* to label sentences that involve two concepts separated by a single hierarchical level in the formal ontology and *S2* to label sentences that involve two concepts separated by two hierarchical levels. Our methodology requires concepts on at least three hierarchical levels in order to compare S1-type sentences against S2-type sentences on verification time and on recall performance. An example of a set might be "animal – bird – robin", "Every robin is a bird" would be a S1 sentence, and "Every robin is an animal" would be a S2 sentence (see Figure 2.1). In previous studies (e.g. [18, 70]), all sentences are presented in the form of *A(n) S is a(n) P*, where S is the subject term and P is the predicate term. For this study, the sentences are modified and presented in the form of *Every S is a(n) P*. This change is necessary because a category membership implies that a category member belongs to the category regardless of the variations of the form. The use of the word "every" ensures the inclusion of all forms when subjects verify the truth of the category membership and when remembering the statements. From each ontology to be evaluated, we select an equal number of sets with concepts around the basic level (for the reason stated earlier) and construct S1 and S2 sentences that are true according to the formal ontology.

SUMO True Sentences			
Set#	Set	S2	S1
1	object-agent-nation	every Nation is an Object	every Nation is an Agent
2	contest-game-sport	every Sport is a Contest	every Sport is a Game
3	process-motion-walking (Walk)	every Walk is a Process	every Walk is a Motion
4	motion-radiating (radiation)-music	every Music is a Motion	every Music is a Radiation
5	process-creation-cooking	every Cooking is a Process	every Cooking is a Creation
6	artifact-text-book	every book is an artifact	every Book is a Text
WordNet True Sentences			
Set#	Set	S2	S1
1	organization-unit-union	every Union is an Organization	every Union is a Unit
2	activity-diversion-escape	every Escape is an Activity	every Escape is a Diversion
3	change-move-step	every Step is a Change	every Step is a Move
4	perception-sensation-noise	every Noise is a Perception	every Noise is a Sensation
5	act-change-reversal	every Reversal is an Act	every Reversal is a Change
6	work-publication-magazine	every Magazine is a Work	every Magazine is a Publication

Table 4.1: True sentence statements constructed from SUMO and WordNet

4.2 Constructing the stimulus list

In total we constructed 12 true sentences (critical stimuli of which the results will be used for analysis) in 6 pairs (six sentences of type S1 matched with six sentences of type S2) from each of the two ontologies. The concepts used in each set are shown in the second column of Table 4.1. The third and forth columns present the S2 and S1 sentences constructed from each set.

In addition to the general selection criteria presented in Section 4.1, there are a few other factors we controlled for when constructing the stimulus list:

- The selection of the concept sets for sentence pair construction for this study specifically avoided compound terms, which was made easier by our selection of WordNet rather than BWW [26]. This was due to the concern that some sentences use compound terms, such as "mutual property", while others do not. This clearly indicates a specialization due to attributes, for example "property – mutual property", and might therefore enable subjects to be more quick to make correct judgments. Given a hierarchically organized set of concepts "A – B – C", we generate sentences of the form "Every B is an A" (S1) and "Every C is an A" (S2). Thus, if the B term is an attribute–noun form specialization of the A term while the C term is not, this might cause the S1 sentence to be verified quicker, leading to a negative S2–S1–TD value (false negative) (See Section 5.1 for description of "S2–S1–TD"). Some BWW and SUMO sets in [28, 27] were of this form, and that could be an explanation for the observed negative S2–S1–TD value in those studies. On the other hand, when the C term is an attribute–noun form specialization of A, but the B term is not, this might lead to inflated positive S2–S1–TD values (false positives). Thus, when a compound term is encountered during the search for a superordinate term with a basic level concept selected, we ignore the compound term and search one level up until we find a single-word concept term. If we take the set "object – agent – nation" as an example, "nation" was randomly selected among the basic level terms. Instead of using its direct superordinate "geopolitical area", "agent" which is 2 levels above "nation" was used. "Object", one level above "agent", was used as the first level of the three concepts of this set.

- The difference in the sentence length (number of letters) between S1 and S2 sentences (S2-S2 letter count difference) should not deviate too much from 0. The length of the sentence had obvious effects on the verification time in [28, 27], and hence impeded the result/judgment of the correctness of the set's structural order. We thus use the sets that will yield sentences with similar S1 and S2 lengths (which vary only on the length of the predicate terms). The difference in sentence length between the S1 and S2 sentence of each set for one ontology should also match that (or be at least in the same positive or negative direction) of the corresponding set in the other ontology.

- Ideally the S2-S1 word frequency value should be similar between sets in the different ontologies, and we should take this variable into account when selecting sets for evaluation [27, 26]. However, in this study we decided not to make word frequency one of the stimulus selection criteria; instead we use it as a covariate in analysis. This is because matching word frequency adds more constraints on stimulus selection. Also, since the quality of an ontology involves the specifications of the concepts, the difficulties of the specifications used in the ontologies (such as the familiarity of the words used) should be included in the evaluation assessment for the quality of the ontology evaluation. These differences should not be controlled since the complexity of the words used can be a component that determines the expressiveness of an ontology.

In selecting stimuli, we excluded the abstract SUMO concepts because WordNet deals only with physical entities. We randomly selected ten concept terms in SUMO ontology that are considered the basic level concepts. The selection frame of basic level concepts in SUMO were decided upon by Jennifer Fang and Joerg Evermann, based on the definition of "basic level" stated in Section 3.2.4, and compared with the basic level concepts used in previous studies (e.g. [18, 70, 42, 73, 78]). For each basic level concept selected, we found its superordinates hierarchically above it and included them as the level 1 and level 2 concepts of the set. We then calculated the S2-S1 letter count difference for each set, and selected out of these ten sets, six with the smallest S2-S1 letter count difference value as evaluation stimuli. In this study, a basic level concept is used as the third level rather than the second level concept of a set such as that in the previous studies (e.g. [18, 70, 42, 73, 78]) because there are very few concept words in SUMO below the basic level.

Once all six SUMO sets have been selected, we then selected a corresponding set from WordNet for each of the SUMO sets. This was to ensure that the sets between ontologies are comparable in terms of their domain focus, level of abstractness, and sentence length.

We use a selected SUMO set, "artifact – text – book", as an example to demonstrate how its corresponding set "work - publication - magazine" was selected. First, we searched the word "book" using WordNet Search 3.0[1], 15

[1] `http://wordnet.princeton.edu/perl/webwn`

hits with different meanings of "book" resulted. We chose one that matches SUMO's meaning of "book"; for example, book - a written work or composition that has been published (printed on pages bound together). We then randomly selected one concept term from the sister terms[2] of "book" which are non-compound concept words (e.g. volume, read, impression/printing, collection, magazine). We used the sister term selected (magazine) as the basic level concept (third level of the set) of the WordNet set corresponding to the SUMO set "artifact – text – book". This was to ensure that the two sets are comparable in terms of the domain focus. We then found the direct hypernym of "magazine"; very often there were multiple hypernyms (e.g. press, public press, publication). In these situations, we chose randomly from the available non-compound word concepts. We then selected the first level concept (work) by finding the hypernym of the second level concept selected (publication), again, avoiding the compound words. In the situations where there were multiple hypernym terms that were non-compound terms, we selected the concept term that would allow similar S2-S1 letter count difference with that of the corresponding SUMO set.

Furthermore, whenever we encountered the situation where all hypernym concepts were compound words, in selecting SUMO sets or WordNet sets, we skipped the level and found a hypernym from two levels above. In this case, we made adjustments in the corresponding ontological set to make the level gaps between concepts consistent between the two ontological sets. This was to ensure that the level of abstraction of S1 and S2 sentences were comparable between ontologies.

We also modified some concept terms so the sentence stimuli would make sense. For example, "Every walking is a motion" was changed to "Every walk is a motion". The sentence stimuli selected for evaluation are presented in Table 4.1. Set 1–6 in the SUMO stimulus list are the corresponding sets of the Set 1–6 in the WordNet stimulus list respectively. The modified terms are presented in the brackets.

Because of a possible influence of word familiarity on the memorization, reading and verification speed, word frequency (as a proxy for familiarity) is computed for each set. First, for each concept, we looked up the Kucera-Francis word frequency information in the MRC Psycholinguistics Database

[2]Sister (coordinate) terms: concept terms that share the same hypernym (superordinate).

at the University of Western Australia[3]. For each sentence, containing two terms (subject term and predicate term), the average term frequency of the two terms was computed. For each sentence pair (S1 and S2), we calculated the difference in average term frequencies. We included this characteristic as a covariate in our analysis (Sections 5.5 and 6.3.4). Take the first SUMO set for example (see Table 4.1), the terms "object", "agent", and "nation" have word frequencies of 65, 44, and 139 respectively, and the average word frequency for the S1 sentence "Every nation is an agent" is 91.5. For the S2 sentence "Every nation is an object", the average word frequency is 102. The difference in the average frequencies is 10.5. This suggests that the S2 sentence may be more familiar to participants (if frequency is a good proxy for familiarity) and this needs to be controlled for.

4.3 Experimental procedures

We conducted two experiments with two separate groups of subjects, each group was used for one evaluation technique. A between subject design for the two techniques was necessary because the same set of stimuli selected from SUMO and WordNet was used for evaluation (See Appendix C for a more complete analysis for choosing this subject design). Experiment 1 was comprised of three parts:

1. A replication of Hirshman's study [42] for the recall technique
2. A replication of Rips et al.'s study [70] for the SVT technique
3. The SVT test trials for the evaluation of SUMO and WordNet

Experiment 2 consists of only the test trials of the recall technique for the evaluation of SUMO and WordNet. Experiment 1 Part 2 and 3 will be reported in Chapter 5; Experiment 1 Part 1 and Experiment 2 will be reported in Chapter 6.

It is an economical design incorporating three parts into Experiment 1. Having the replication of Rips et al.'s study [70] and the SVT test trials in one experiment saved us from recruiting two sets of participants. Also, with this design we could minimize possible subject effect - that differences found in the findings of the replication study and the test trials are due to

[3] `http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm`

differences in subject groups. The replication of Rips et al.'s study [70] serves in Experiment 1 as the practice trials for the SVT test trials.

The replication of Hirshman's study [42] for the recall technique is included in Experiment 1 because, different from the SVT technique, the recall technique should not have practice trials. This is because a prior exposure to the experimental procedure may bring biases to the study of the recall test trials in several ways. First, subjects are likely to learn and work out the purpose and measurements for the experiment and thus in the test trials change their strategy for memorization in order to achieve a higher recall score. Also, subjects may become better at the task after practice. These maturation and pre-testing[4] effects would have lowered the internal validity of the study. Second, during the test trials when recalling the stimuli studied, subjects might have confused the stimuli in the test trials with those seen in the replication study. This would greatly affect subjects' performances in the test trials.

All experimental instructions were given in written form to avoid experimenter effects [80] (Appendix M). Participants had no knowledge of the nature and purpose of this study until the de-briefing after the experiment. An information sheet was given to participants and a brief verbal introduction repeating the content of the information sheet. Two different sets of information sheets and de-briefing sheets were used for Experiment 1 and Experiment 2 (See Appendices F, G, H, I). Subjects had no explicit knowledge of the two ontologies. As we are trying to elicit their (implicit) conceptualization, any explicit knowledge of the formal ontologies would be a confound.

4.4 Pilot tests

4.4.1 Experiment 1

It was considered that having three tests/tasks in one experiment might result in fatigue effects in subjects. To examine this, we ran a pilot test with three

[4]The effect created on the second measurement by having a measurement before the experiment. This issue might be created by the practice trials as participants learn from the practice trials and might change their answers to what they think are more acceptable or more appropriate. Individuals who were pretested might be more sensitive to the experimental variable or might have learned from the pre-test making them unrepresentative of the population who had not been pre-tested.

subjects (these subjects were not used in Experiment 1 or 2) who were tested individually. The total duration of Experiment 1 was around 20 minutes on average. I asked for feedback at the end of the experiment with each subject. To the question "What do you think about the length of the experiment? Too short, appropriate, or too long?", all three subjects reported "appropriate". To the question "Was your concentration level throughout the experiment fairly consistent?", subjects reported that because they were doing three different tasks, they were able to concentrate fairly well on each of them. Furthermore, originally we structured Experiment 1 to have the replication of Hirshman's study [42] at the end, as the last part of the experiment. After the pilot test, we moved it to the first part of the experiment. This is because when asked for other comments, a subject reported that he confused the stimuli seen in the replication of Rips et al.'s study [70] with the stimuli to be recalled in the replication of Hirshman's study [42] when doing the recall task.

4.4.2 Experiment 2

We conducted a pilot test with 3 subjects (excluded from subjects used in the Experiment 1 and 2) using the material, design and procedure described in Section 6.3 to test the adequacy of the instructions provided in Experiment 2. This also enables us to gain an estimate for the duration of the experiment. We found that the subjects were clear on what they had to do after reading the instructions (see Appendix M) presented as the first PowerPoint slide, and the total duration of the experiment (including briefing and debriefing) was on average 13 minutes. The only modification made from the pilot test was to include 2 filler sentence statements at the end of the stimulus list for study, instead of only 1 which was designed to be consistent with Hirshman's study [42]. This was because we found that all three subjects successfully recalled the last as well as the second last stimulus, and these two stimuli were often recalled quite early on in the recall phase. To avoid the possibility of the second last stimulus being biased by the recency effect, we added one more sentence statement to make two fillers at the end of the study list (refer to Section 6.3.2).

4.5 Subjects

There were 30 university students participating in Experiment 1 and 60 students participating in Experiment 2.

The 30 subjects in Experiment 1 were recruited via recruitment posters displayed in different schools/faculties of Victoria University of Wellington (see Appendix E). Experiments were run individually to avoid competitive behaviour.

Ensuring a sufficient number of participants is important as it plays an important role in the generalizability of the findings [21]. Based on the number of observations included in Collins and Quillians' [18] and Rips et al.'s [70] studies, there should be at least 10 observations in each set. This is because statistical analysis is used to compare S1 and S2 also on the individual set level (refer to Section 5.1). This means that for each set, we required at least 10 observations from subjects with correct responses in both S1 and S2 sentence verification. With 30 subjects in this experiment, this criterion was easily achieved.

There were 86 subjects recruited in Experiment 2 (among these subjects, 26 were later excluded for not meeting the subject selection criteria listed in the following paragraphs). These subjects are second year students majoring in Information Systems. They were approached and tested at the beginning of their tutorial sessions. Participation was completely voluntary (see Appendix H). Data collection was completed in 5 tutorial sessions with around 19 students in each tutorial class. With the experimental design of the recall technique, each sentence stimulus was presented to 30 subjects to memorize and recall (this will be further described in Section 6.3)

We used university students because both ontologies, SUMO and WordNet, claim to represent general and generic human categories and there should consequently be no need for specialist knowledge.

We recruited participants who are native English speakers. This is because

1. Both ontologies use English terms for their concepts

2. There is a strong relationship between language and cognitive structure [84]

3. We chose to increase internal validity of a homogeneous sample over generalizability in this study

Subjects recruited for Experiment 2 were asked to indicate in the consent form whether English is their first language (see Appendix J), among the 86 students who participated, only the data from 60 native English speakers were included for evaluation.

Gender difference has not been raised as an issue in the study of category organization and semantic memory. Thus, it is not considered in the design of this experiment.

An award of $50 for the best performance was used as a motivation for each experimental technique. Subjects in Experiment 1 were told during briefing (Appendix F) that the award will be given to the best performer in the experimental trials, and during de-briefing (Appendix G) informed that performance would be measured through both speed and correctness of their answer in the test trials (Part 3 of Experiment 1). We reserved the details until debriefing because we needed the subjects to also take the replication studies seriously (Part 1 and 2 of Experiment 1). Subjects in Experiment 2 were told during briefing (Appendix H) that the award will be given to the best performer in the recall task, and during debriefing (Appendix I), specified that the number of correct recalls of stimuli presented is the factor of evaluation. The measurements were not made known to the subjects until debriefing, this was to avoid having subjects speculating the purpose and nature of the study and be biased with the pre-testing effect.

Chapter 5

Sentence verification task

We base the development of the SVT evaluation technique on Rips et al.'s study [70] for the following reasons: First, it provides us with a full list of stimuli used and the result (mean RT) obtained for each stimulus (See Table A.1). This data is important because it allows us to do a full replication of the study and compare the findings in detail. Second, it brings up an interesting finding about the SVT technique which is important to investigate further and be taken into account for - that memory structure does not necessarily mirror logical structure (this will be further discussed in Chapter 8). Rips et al.'s study [70] is feasible for adoption as it is a study which was developed based on Collins and Quillians' study [18]. These two studies have very similar experimental designs and the same assumptions attached to them.

In the study by Rips et al. [70], reaction time was used to measure the time taken for people to retrieve information from memory. In this study, we replicated the first experiment of Rips et al.'s study [70] (see Appendix A).

The experimental method follows the description in Chapter 4. This chapter reports on issues specific to the SVT technique. It presents the designs and findings of the replication of Rips et al.'s study [70] and the SVT test trials. The replication study was designed and used as the practice trials for our SVT test trials. Subjects were informed that their responses in the practice trials were recorded for analysis. The SVT test trials followed immediately after the replication study, and the same 30 participants were used for these two parts of the study.

5.1 Evaluation assumptions

Based on the previous studies [18, 70], the assumption was made that the time taken to respond "true" is longer when verifying S2 sentences than when verifying S1 sentences, if the order structure of concepts in the ontologies is correct.

For each set, if both S1 and S2 are responded to correctly, and the assumption (S2 takes longer to verify than S1) is held true, the category order of the three concepts of the set can be assumed correct. Based on the assumption above, we evaluate the two ontologies on three levels.

5.1.1 Ontologies

This analysis design compares the overall quality of the two ontologies. For each ontology, we calculate the mean difference in the time taken to verify S2 sentences and the time taken to verify S1 sentences (S2–S1–TD). We then compare the overall correctness of the ontological structures by comparing the mean measures of S2–S1–TD of the two ontologies.

This analysis provides us with a summarized comparison on the quality of the conceptualizations of the two ontological structures, as an indication of how well these ontologies represent human's knowledge structure. This analysis illustrates the overall quality of the two ontological structures. The ontology that obtains a larger positive mean S2–S1–TD value (longer verification time for S2 than for S1) is assumed to have a better ontological structure. On the other hand, if the ontologies obtain negative means (longer verification time for S1 than for S2), an ontology with a larger negative mean would suggest a more problematic ontological structure in general. The ontology is the independent variable, and the mean S2–S1–TD of the ontology is the dependent variable.

5.1.2 Sets

This analysis design examines the quality of each set included in the evaluation. This is done by identifying the sets with significant positive S2–S1–TD values and the sets with significant negative S2–S1–TD values, in both the SUMO ontology and in WordNet.

Each set is tested individually in this analysis design. Based on the earlier assumption, it is assumed that the sets with significant positive S2–S1–TD

values have concepts ordered in the correct structure; while sets with significant negative S2–S1–TD values have incorrectly ordered concepts, indicating erroneous structures. The independent variable is the nodes of separation (semantic distance between the two related concepts in a sentence stimulus, as recorded on the ontology) that is manipulated in a within-subject design, and the dependent variable is the S2–S1–TD value which is a proxy for quality.

This analysis allows us to examine the correctness of individual sets and therefore gives us a more detailed understanding of the quality of the two ontologies. This analysis is important in evaluating ontologies, especially if the quality of the ontological sets varies. This is because the mean S2–S1–TD value of an ontology obtained in Section 5.1.1 averages across the S2–S1–TD values of the individual sets, and we might thus overlook other important results that indicates the quality of ontologies. For instance, assuming a significant positive mean S2–S1–TD value is found in one ontology during the analysis described in Section 5.1.1, it is possible that the significant result is contributed to by only very few sets that have highly significant positive S2–S1–TD values.

5.1.3 Error rates

We identify and compare the error rates in the responses when verifying sentences constructed from SUMO and that constructed from WordNet.

Error rate serves as an indication of the quality of the ontologies. Based on the previous psychology studies (e.g. [7]), we assume that the ontology with a higher error rate is less representative of the perceived reality. Low representativeness can mean that the categorization of the concepts is incorrect or not representative to the participants.

The main purpose of including the analysis designs in Section 5.1.2 and 5.1.3 is to demonstrate methods of assessing ontologies on the level of individual concepts. These analyses provide valuable information for the ontologies evaluated, as the quality (correctness of structure order and error response rate) of each set is examined. These analysis designs provide the evaluators with a means to identify the problematic concepts and what the problem might be. For instance, if the S1 and S2 sentences in a set both have low error rates in their responses, but the set appears to have a significant negative S2–S1–TD value, we can assume that the concepts in this set are generally correctly categorized but the level 1 concept and the level 2 concept of the

particular set are in the incorrect order. Only after identifying problems can the quality of ontologies be improved.

5.2 Material

To avoid response bias, the SVT technique typically involves presenting subjects with both true and false statements. Verification times are only meaningful for the true statements, as "there is no direct route in our mental model to falsification" [8]. Moreover, true sentences that subjects recognize as false and vice versa indicate further problems with the relative hierarchical level of the concepts in the formal specification. We therefore randomly selected elements of each ontology to be evaluated and constructed as many false sentences as true ones. For example, from Figure 2.1, we might construct "Every robin is a fish" or "Every shark is a bird". To avoid possible priming bias [85], each of the concepts in the subject position of the false statements was randomly selected from concepts that were not used for the true sentences, and the subject terms were paired with the predicate terms used to construct true sentences. This was to be consistent with the design of Rips et al.'s study [70] where words used as the predicate terms appeared equally often in both true and false statements. Thus, to construct false sentences, we listed all 24 true sentences, and for each sentence (e.g. "Every nation is an object") we swapped the subject term (nation) with a randomly selected basic level concept that had not been used in the true sentences, was not a member of the predicate term (object), and was not a compound term. Using the example above, the concept term "reasoning" was selected to construct the false sentence "Every reasoning is an object".

In total we constructed 12 true sentences in 6 pairs (six sentences of type S1 matched with six sentences of type S2) from each of the two ontologies. The concepts used in each set are shown in the second column of Table 4.1. These sentences were presented in random order to subjects, interspersed randomly with 12 false sentences from each ontology to remove response bias. Thus, each subject was presented with 48 sentences (12 true sentences plus 12 false sentences for each of the two ontologies). The false sentence stimuli are shown in Table 5.1.

Set#	S2	S1
SUMO 1	every Reasoning is an Object	every Reasoning is an Agent
SUMO 2	every Fish is a Contest	every Fish is a Game
SUMO 3	every Virus is a Process	every Virus is a Motion
SUMO 4	every Government is a Motion	every Government is a Radiation
SUMO 5	every Vitamin is a Process	every Vitamin is a Creation
SUMO 6	every City is an Artifact	every City is a Text
WordNet 1	every Socialism is an Organization	every Socialism is a Unit
WordNet 2	every Background is an Activity	every Background is a Diversion
WordNet 3	every Election is a Move	every Election is a Change
WordNet 4	every Institution is a Perception	every Institution is a Sensation
WordNet 5	every Date is an Act	every Date is a Change
WordNet 6	every Necklace is a Work	every Necklace is a Publication

Table 5.1: False sentence statements constructed from SUMO and WordNet for the SVT test trials

Practice/Replication trial sentences

Since the replication study is only one of the three parts of Experiment 1, we were concerned about fatigue effects and therefore only tested a portion of the concept sets in Rips et al.'s [70] stimulus list (see Table A.1). We selected six sets from the stimulus list, two sets from each of the bird, mammal, and car categories. The sets from the bird and mammal categories were selected randomly. Only the cars that are familiar to New Zealanders were included in the selection frame of the car category because this study was based in New Zealand, and "Porsche" and "Toyota" were picked. See Table 5.2 for the replication trial sentence stimuli used. The first column of Table 5.2 shows the classification of types of sentences used in Rips et al.'s study [70]. The construction of the 12 true sentences and 12 false sentences shown in the second column of Table 5.2 was based on Rips et al.'s material design [70] (see Appendix A).

Type of sentences used in [70]	Sentence used in our study
TRUE SENTENCES	
An S(B) is a bird	a Parrot is a Bird a Sparrow is a Bird
An S(B) is an animal	a Parrot is an Animal a Sparrow is an Animal
An S(M) is a mammal	a Rabbit is a Mammal a Cow is a Mammal
An S(M) is an animal	a Rabbit is an Animal a Cow is an Animal
An S(C) is a car	a Porsche is a Car a Toyota is a Car
An S(C) is a vehicle	a Porsche is a Vehicle a Toyota is a Vehicle
FALSE SENTENCES	
An S(B) is a mammal	a Chicken is a Mammal a Robin is a Mammal
An S(B) is a car	a Pigeon is a Car
An S(B) is a vehicle	a Parakeet is a Vehicle
An S(M) is a bird	a Pig is a Bird a Dog is a Bird
An S(M) is a car	a Deer is a Car
An S(M) is a vehicle	a Sheep is a Vehicle
An S(C) is a bird	a Volkswagen is a Bird
An S(C) is a mammal	a Honda is a Mammal
An S(C) is an animal	a Mitsubishi is an Animal a Nissan is an Animal

Table 5.2: Sentence stimuli constructed for the SVT practice/replication trials

5.3 Experimental design and procedure

We chose a within-subject experimental design for comparing the two ontologies for efficiencies of sample size and control over subject-specific influences.

In this study, sentence presentation and data collection (including timing) were controlled by the FLXLab software[1], and the sentences were displayed on an LCD computer screen. An introduction to the application of the

[1] `http://flxlab.sourceforge.net`

experimental program is presented in Appendix D. Subjects made true/false responses by pressing either the V or B key on the keyboard. Each key was assigned "true" for half of the participants, and "false" for the other half in order to avoid bias of dominant hands [70]. A sticker with the letter "T" was applied on the true response key and another sticker with the letter "F" on the false response key.

Two random presentation orders of the test sentences were used, with half the participants receiving each order. The four sets of experimental programs (2 presentation orders $\times$ 2 key assignments) were counterbalanced over the participants. Two programs for the replication of Rips et al.'s study [70] differed only in the assignment of the true and false response keys. For every participant, the key assignment remained the same from the replication study to the test trials.

The experimenter prepared the software for the subject by entering the subject's ID (explained in Appendix D). An instruction page then appeared on the screen (see Appendix M). After reading the instructions, the subject pressed a key as instructed on the instruction page, a black "+" sign appeared at the center of the screen for 2 seconds, followed by the test sentence. For each sentence presentation, the timer started as each sentence appeared on the screen, and the subject's response terminated the sentence presentation and stopped the timer. This was repeated until all replication trials were completed. A two-second interval was chosen in agreement with [18]. After the participants responded to the last replication sentence, the software terminated and the experimenter checked with the participant that he/she understood what he/she was asked to do. The experimenter then started the test program, and again entered the subject's ID. The experimenter did not face or observe the participants to avoid experimenter bias [80]. The procedure for the test sentences was the same as that for the replication trials. Appendix M contains the on-screen instructions for subjects.

We analyzed only the RTs for correct responses from the true sentences for further analysis; the false sentences are not designed to be analyzed. Moreover, only those pairs of true sentences for which subjects recognized both the S1-type and the S2-type sentence as true, were considered in the analysis.

Our replication study followed Rips et al.'s study [70] as closely as possible. However, modifications to the design were made:

1. A different apparatus was used, the design of the replication study al-

lowed more control and was more standardized and consistent. For instance, in this study the sentence presentation, timing and data recording were monitored/controlled by the FLXLab program. Instead of a signal light, a "+" was used to signal for the onset of the presentation of each sentence; instead of having sentence typed on a white card and have an exposure device that consisted of a half-silvered mirror illuminated from behind, each sentence (also the instructions and the signal "+") was presented on an LCD screen. One aspect of the experimental program procedure was also modified - we decided against informing the subject during a 7-second intertrial interval if he/she had made an error in the sentence verification. This change was made to avoid possible frustration or emotional distress which might influence participant's latter performance. It might also minimize the risk of possible maturation effect, pre-testing effect and experimenter effect [100], as was shown with Rips et al.'s design, where participants might have learned from the previous trials. Thus, instead of a 7-second intertrial interval, a 2-second interval was chosen in agreement with [18].

2. The presentation programs in the replication study differed only in the assignment of the true and false response keys, but not in the presentation order. This was because the replication study was used as the practice trials for the SVT test trials. To be consistent with Rips et al.'s study [70], we only needed one presentation order for the practice/replication trials.

The experimental design and procedure for the SVT test trials were consistent with those of the replication study, except for some details in the material design, listed as follows:

1. Four experimental programs (2 presentation orders $\times$ 2 key assignments) were counterbalanced over the participants, this was in order to be consistent with Rips et al.'s study [70].

2. The instances (subject terms of the sets) in Rips et al.'s study [70] were drawn by free association (most commonly listed instances were used) so that subjects were all familiar with them. Whereas in the test trials of this study, the sets were constructed from formal ontologies; hence subjects' familiarity to the instances and concept terms used might vary. We controlled for this possible biasing factor by including the word frequency scale as a covariate in analysis.

3. In constructing the test stimuli from ontologies, we had the basic level concepts at the lowest level of the set hierarchy, and the level gaps were not as standardized as those in Rips et al.'s study [70] (Refer to Section 4.2).

4. The sentences were constructed in the form of "Every S is a P" rather than "A S is a P" (see Section 4.1).

There are two issues with the software design used in the trials. First, the software does not recognize the specified response keys V and B. Any key press terminates the sentence presentation. The keys pressed during the trials are recorded into the output data. Data with incorrect (not V or B key) response keys were discarded. Such incidents did not occur in this study. Second, the program is not able to distinguish double responses given to one sentence from normal responses. The second key press for the same sentence generates a response time for the next sentence. In this study, we were not able to identify and determine double answers from the result data. The occurrence of this event was reduced by emphasizing this instruction: It is important that you give only one key response to each sentence (see Appendix M).

5.4 Result of the replication study

The overall error rate was 9 percent, which was higher than the 4 percent error rate found in Rips et al.'s study [70] (see Appendix A). We analyzed the RT of correct true responses. Table 5.3 shows the sets selected from Rips et al.'s study [70] for the replication evaluation. The figures on the second and third columns are the mean RTs of the associated sentence statement as a function of the level of the predicate noun. The final column indicates whether a subset effect (Level 1 (S2) RT greater than Level 2 (S1) RT) was obtained for each set/instance. For each column, the results presented on the left hand side are Rips et al.'s [70] findings and the right hand side shows the findings of the replication study, for comparison.

We performed an ANOVA on an $atan()$ transformation of the S2–S1–TD scores. A transformation was needed because an ANOVA procedure assumes normality of the data which was not satisfied. We first standardized our data to $\sigma = 1, \mu = 0$. A histogram was plotted using the standardized data (see Figure 5.1) before transformation. It showed a peak with high kurtosis

Subject noun	Level 2 Predicate noun		Level 1 Predicate noun		Difference Level2-Level1	
	Rips	Repl.	Rips	Repl.	Rips	Repl.
	Bird		*Animal*			
Parrot	1284	1686	1342	1943	58	256
Sparrow	1339	1542	1477	4684	138	142
	Mammal		*Animal*			
Rabbit	1418	2868	1290	1693	-128	-1175
Cow	1258	1888	1322	1665	64	-223
	Car		*Vehicle*			
Porsche	1348	1355	1395	1391	47	36
Toyota	1320	1281	1136	1564	-184	283

Table 5.3: Comparison of verification time (msec) between Rips et al.'s study (Rips) [70] and the replication study (repl.)

and appeared to be non-normal (kurtosis $K = 21.89595$, Shapiro-Wilks test $p_{SW} < 2.2e{-}16$, Kolmogorov-Smirnoff test $p_{KS} = 2.977e{-}07$). The transformation significantly reduced the kurtosis ($K = -0.004446034$), the Shapiro-Wilks test showed no significant departures from normality ($p_{SW} = 0.0671$) (although the KS test still showed non-normality ($p_{KS} = 3.108e - 06$)), and a qq-normal plot also indicated normality (See Figure 5.2). The ANOVA analysis indicated that there is a significant category effect ($F(2, 115) = 9.8614, p < 0.001$) (was not significant in Rips et al.'s study) which suggests that the mean S2–S1–TDs are significantly different between sets of different categories (Figure 5.3). There is also a significant difference observed between sets ($F(3, 115) = 3.6234, p = 0.015$)(Figure 5.4).

In Rips et al.'s study [70], (positive) subset effect has been obtained for the instances in the bird and car categories, whereas the opposite effect was found in the mammal instances. The findings in the replication study point towards the same direction as the findings in [70], however, the (positive) subset effects found in the bird and car categories and the opposite subset effect found in the mammal category in this replication study are mostly non-significant. To evaluate whether in each set the difference due to levels (S2 vs. S1) is significant, we performed the non-parametric Wilcoxon-Rank-Sum test (instead of a t-test) because of the non-normality of the data (indicated by the KS test). A significant result was found only for the set "Vehicle – Car – Toyota" ($p = 0.017$), which states a significant subset effect. Although the

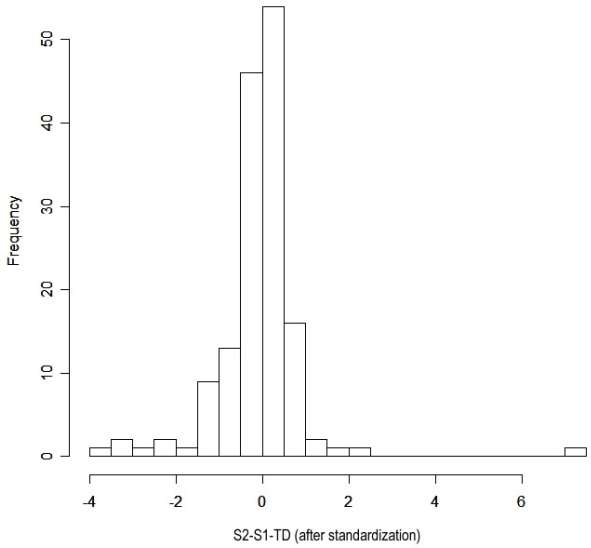

Figure 5.1: Histogram of time differences between the S1-type and S2-Type sentences (standardized data)

finding of this set differs from the opposite subset effect found in Rips et al.'s study [70], it supports our findings [70] on the category level, that instances in the "car" category obtains (positive) subset effect. The difference in the results of this set found between Rips et al.'s study [70] and the replication study is possibly due to the differences in characteristics of the subjects used; specifically, New Zealanders (our subjects) might associate this brand of the car (Toyota) more closely with the concept "car" than with "vehicle", and Americans (Rips et al.'s subjects) might associate "Toyota" more closely with "vehicle" than with "car".

It was suggested that the findings of the opposite effects in the mammal category are indications that our memory structures do not necessarily mirror the logical structures [70] (will be discussed further in Chapter 8). Moreover, the opposite subset effect obtained for the "mammal" instances might be due to the fact that "mammal" is not a common word (Kucera-Francis word frequency of 1); especially when compared to the Level 1 predicate term of this category, "animal" which has the word frequency value of 68. Therefore it may take subjects longer to verify the S1 sentence (e.g. "every rabbit is a mammal") than the S2 sentence ("every rabbit is a animal") for the

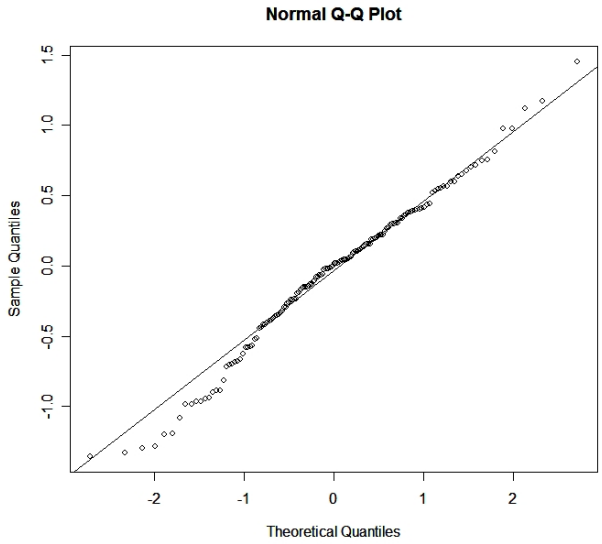

Figure 5.2: Q-Q plot of time differences between the S1-type and S2-Type sentences (standardized and transformed data)

"mammal" instances. The word frequency difference between the level 1 and level 2 predicate terms (S2-S1) for the "bird" instances is smaller (animal(68)-bird(31)= 37) than that for the "mammal" instances (68-1=67), and for the "car" instances, the level 2 predicate term has a higher word frequency value than the level 1 predicate term (word frequency difference is -239). With these figures, it appears that the differences in word frequencies between S2 and S1 sentences (S2-S1) have impacts on the S2–S1–TD obtained. It is thus important to control for the word frequencies of the concepts used for evaluation in the ontology study.

Comparing these results to Rips et al.'s findings [70] which show that within each category the difference due to levels is significant in all cases ($p < .01$), our findings for most sets in the replication study failed to reach the significance level needed to suggest significant differences, even though they point in the same direction as the findings in [70]. The non-significant results in this replication study, however, can be explained by the low power value (the power to detect an effect if it is there) of the study. The power values for Set 1 – 6 are 0.1756, 0.2210, 0.9939, 0.1520, 0.0091, and 0.4479 respectively. Set 6 "Vehicle – Car – Toyota" was found significant in the

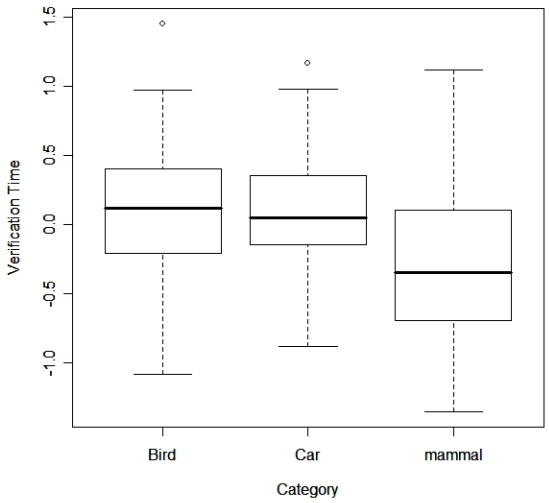

Figure 5.3: Boxplot of time differences between the S1-type and S2-Type sentences, by category (standardized and transformed data)

Wilcoxon-Rank-Sum test. Set 3 "Animal – Mammal – Rabbit", although was found not significant in the Wilcoxon-Rank-Sum test, it was significant in a t-test ($t(23) = -3.398, p = 0.002$). The non-significant results for the other four sets may become significant in the same study with a larger sample size, which is important for achieving higher power. Since similar findings to Rips et al.'s study [70] were obtained in this replication study, as shown in Figure 5.3 (positive subset effect in the "bird" and "car" categories, and opposite subset effect in the "mammal" category); the results may achieve significance with a higher-powered test; therefore we are confident in the validity of the SVT technique for ontology evaluation.

5.5 Result of the test trials

5.5.1 Ontologies

To compare the ontologies with respect to the differences in verification times between the S1 and S2 sentences (S2–S1–TD) we can use the ANOVA procedure. This procedure assumes normality of the data which was not satisfied.

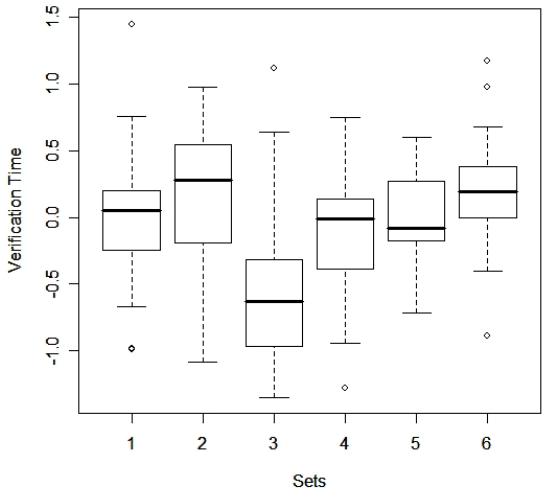

Figure 5.4: Boxplot of time differences between the S1-type and S2-Type sentences, by set (standardized and transformed data). 1 = the set of "Animal – Bird – Parrot", and 6 = the set of "Vehicle – Car – Toyota"

We first standardized our data to $\sigma = 1, \mu = 0$. Figure 5.5 shows a histogram plotted using the standardized data; it showed a high kurtosis (peaked) and was clearly non-normal (kurtosis $K = 21.89595$, Shapiro-Wilks test $p_{SW} < 2.2e-16$, Kolmogorov-Smirnoff test $p_{KS} = 2.977e-07$). To achieve normality, an $atan()$ transformation was applied to the data. The transformed data showed a significantly reduced kurtosis ($K = -0.004446034$) and appeared to be normally distributed (see Figure 5.6 for the qq-normal plot). The Shapiro-Wilks test also showed that there was no significant departure from normality ($p_{SW} = 0.0671$). Thus, we proceeded with this analysis using the $atan()$ transformed data which, according to the results of the Shapiro-Wilks test, achieved normality. However, the KS test still showed significant non-normality ($p_{KS} = 3.108e-06$) with the transformed data. Therefore, we also analyzed the data using a non-parametric analogue of ANOVA (Kruskal-Wallis test) which does not make assumptions about normality.

The S2–S1–TDs were compared using a within-subject ANOVA with factors "Ontology" (2 levels) and "Set" (12 levels, 6 nested within each "On-

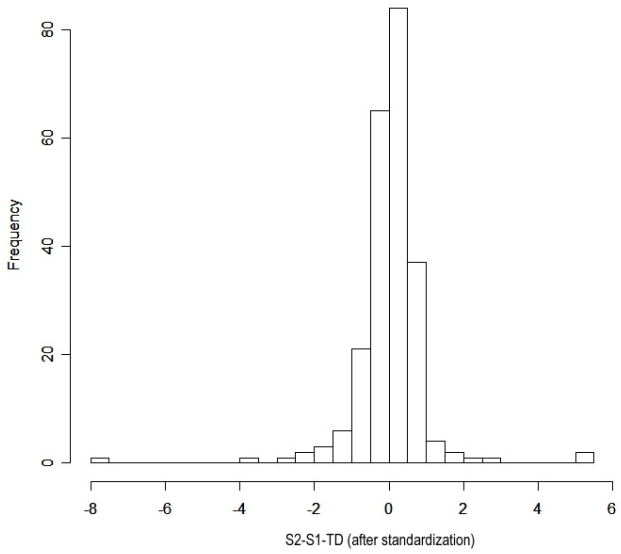

Figure 5.5: Histogram of time differences between the S1-type and S2-Type sentences (standardized data)

tology")(Figure 5.7). We included the differences in average word frequency (AVG–Diff) as a covariate in this ANOVA analysis, presented in Table 5.4.

We found no significant differences between the two ontologies ($F(1, 190) = 0.1558, p = 0.6935$) (Table 5.4). Thus, there is no overall difference in quality of the two ontologies. Figure 5.7 shows the mean S2–S1–TD between the two ontologies visually.

There is also no significant effect of the set (sentence pair) on the response time ($F(9, 190) = 0.6586, p = 0.7455$). This indicates that the ontological sets within the ontologies have consistent quality, in terms of the correctness of the structural order of concepts within each set. In Figure 5.8, Sets 1–6 represent SUMO ontological sets, and sets 7–12 represent WordNet sets. Furthermore, the average word frequency difference (AVG–Diff) has no significant effect on the response time difference ($F(1, 190) = 0.7031, p = 0.4028$).

We performed the Kruskal-Wallis non-parametric test twice; first to look for differences between the two ontologies, and second to examine the differences between the pairs of sentences within each ontology. This was because the Kruskal-Wallis non-parametric test does not permit modelling of nested factors. The results of this test indicate that there is no significant difference

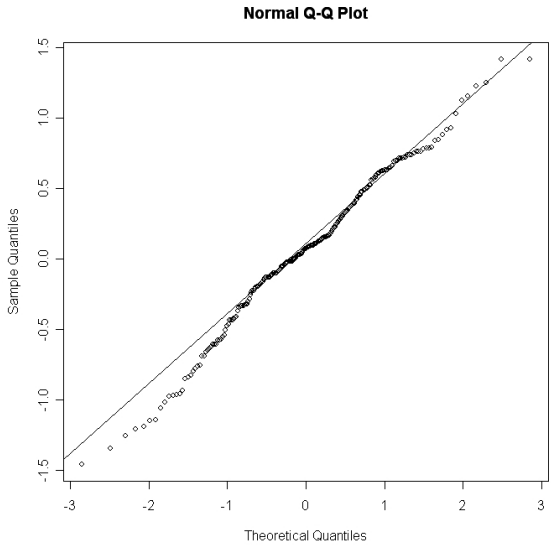

Figure 5.6: Q-Q plot of time differences between the S1-type and S2-Type sentences (standardized and transformed data)

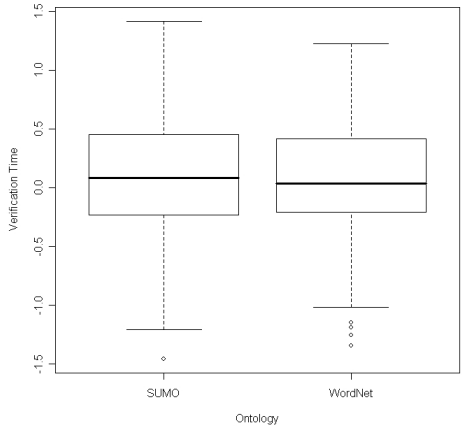

Figure 5.7: Boxplot of time differences between the S1-Type and S2-Type sentences, by ontology (standardized and transformed data)

between ontologies (the Kruskal-Wallis χ^2 statistic was 0.0522 with $df = 1$, not significant at $p = 0.8193$). There is also no significant difference be-

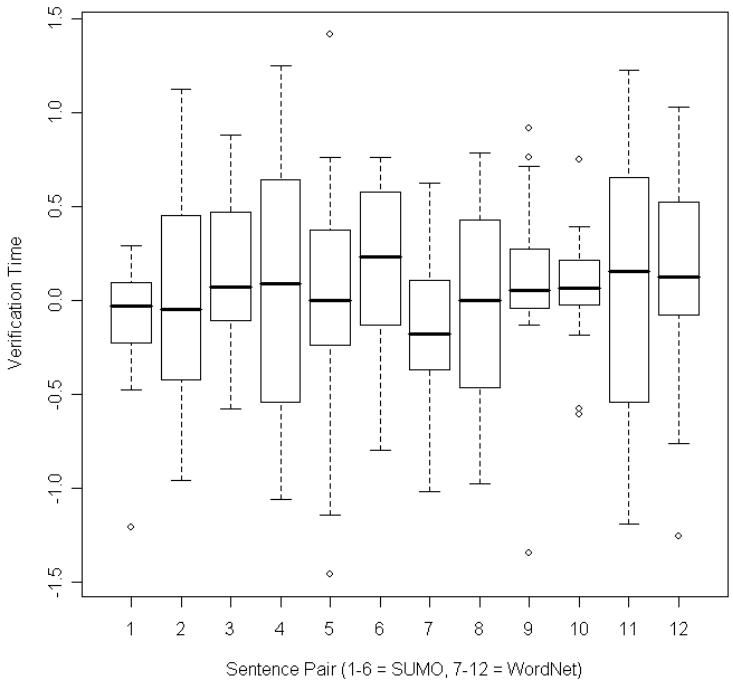

Figure 5.8: Boxplot of time differences between the S1-type and S2-Type sentences, by set (standardized and transformed data)

tween sets (the Kruskal-Wallis $\chi^2 = 10.008$ with $df = 11$, not significant at $p = 0.5297$). These findings are consistent with the analysis using ANOVA presented above. The Kruskal-Wallis test does not permit modelling a within-subject design, nor does it permit covariates such as the differences in word frequency we included in the ANOVA analysis.

5.5.2 Sets

To examine the quality of the ontologies in more detail, we compared the mean S2–S1–TD of the S1-Type sentence and that of the S2-Type sentence for each of the 12 sets of concepts (6 sets in each ontology). Because of the non-normality of the data, we could not use a t-test and instead performed the non-parametric Wilcoxon-Rank-Sum test. The results are shown in Table 5.5.

Error: Subject						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Sig
Ontology	1	0.0828	0.0828	0.2150	0.6485	
AVG–Diff	1	0.0592	0.0592	0.1536	0.6997	
Ontology:Pair	9	1.3987	0.1554	0.4035	0.9170	
Residuals	18	6.9325	0.3851			
Error: Within						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Sig
Ontology	1	0.047	0.047	0.1558	0.6935	
AVG–Diff	1	0.212	0.212	0.7031	0.4028	
Ontology:Pair	9	1.786	0.198	0.6586	0.7455	
Residuals	190	57.235	0.301			

Table 5.4: ANOVA Results (SVT). AVG–Diff is the difference in average word frequency

We used a one-sided test, as we expected a positive mean S2–S1–TD value if the concepts of the sets were correctly ordered, that is, the response time for S2 should be greater than the response time for S1. The results showed that only 1 of 6 SUMO sets (Process – Motion – Walk) and 1 of 6 WordNet sets (Change – Move – Step) had S2–S1–TD as significantly greater than zero ($\alpha = .05$), as highlighted in Table 5.5.

5.5.3 Error rates

The error rates for both sentences in a set are shown in the second column of Table 5.5. The sentence "Every nation is an object" from SUMO was responded to incorrectly more than 50% of the time, which implies a problematic ontological structure as perceived by our participants. An average of only 23 of the 30 subjects made correct responses on each of the SUMO sentences (SD = 6.14), and an average of only 24 of 30 subjects made correct responses on each of the WordNet sentences (SD = 4.96). The difference in the number of correct response made between the WordNet and in SUMO ontologies is not significant ($t(11) = -0.467, p = 0.649$).

Concepts	correct resp.	S2-S1 Mean	SD	SEM	Sig
SUMO Sets					
Object – Agent – Nation	10	-734.30	2096.84	663.08	0.8207
Contest – Game – Sport	16	100.81	2072.68	518.17	0.55
Process – Motion – Walk	27	465.96	1149.62	221.24	**0.03861**
Motion – Radiation – Music	13	721.08	3202.02	888.08	0.2709
Process – Creation – Cooking	28	84.00	6203.48	1172.35	0.4777
Artifact – Text – Book	22	475.73	1401.15	298.73	0.05273
WordNet Sets					
Organization – Unit – Union	21	-479.86	1502.96	327.97	0.9105
Activity – Diversion – Escape	15	-154.33	1939.15	500.69	0.533
Change – Move – Step	22	100.14	2607.46	555.91	**0.04587**
Perception – Sensation – Noise	14	139.79	1007.87	269.37	0.2131
Act – Change – Reversal	20	1.90	2877.85	643.51	0.3643
Work – Publication – Magazine	23	282.78	2222.20	463.36	0.06707

Table 5.5: Concept sets for sentence pairs used in this study, number of correct responses, differences in verification times (S2-S1 Mean), standard deviation (SD) and standard error of the mean (SEM) differences in verification times, and significance level of the one-sided Wilcoxon Rank-Sum non-parametric test for equality to 0 (Sig)

5.6 Discussion

The results of the replication study are consistent with the findings in Rips et al.'s study [70] - Figure 5.3 shows a positive subset effect in the sets of the "bird" and "car" categories, and an opposite subset effect in the sets of the "mammal" category. Although these results did not achieve significance in the replication study, we believe they would have with a higher-powered test which can be done with a larger sample size. We are therefore confident in the validity of the SVT technique, and that it is an appropriate tool for ontology evaluation.

The power calculated from the post-hoc power test is low because the effect size is small. In other words, the effect is too small for us to have sufficient power to detect it, hence the low power value. With a larger effect size, we would have had sufficient power. We expected the effect size to

be larger since we based our replication study on Rips et al.'s study and significant results were found in [70]. Moreover, we had a larger sample size (30) than that in Rips et al.'s study [70] (12), and therefore expected to obtain a large enough power in our replication study. We could not calculate Rips et al.'s [70] effect size and power to compare with those found in the replication study because the standard deviations were not provided.

Since the replication study is only one of the three parts of Experiment 1, we were concerned about fatigue effects and therefore only tested a portion of the concept sets in Rips et al.'s [70] stimulus list (see Table A.1).

It was suggested that the opposite subset effect obtained in the "mammal" instances could be explained by unbalanced word frequency values between the Level 1 predicate noun used in the S2 sentence ("animal" with the word frequency of 68) and the Level 2 predicate noun used in the S1 sentence ("mammal" with the word frequency of 1). We thus included word frequency as a covariate in the analysis. This was to ensure that the S2-S1 word frequency value was controlled for during analysis (refer to Section 4.2).

The findings of the test trial experiment suggested that neither of the two ontologies properly represents the perceived world. It was found that the ontologies do not differ in the quality of their formal specification of the domain conceptualization, after controlling for possible word frequency effects. It appeared that the word frequencies themselves have no effect on the sentence verification response times. There was also no variation between different sets.

As our test results in Table 5.5 show, and the error rates confirm, most of these sets are more or less bad: Only one of the six sets of each ontology showed the behaviour we would expect of an ontology that is a good specification of a conceptualization; the others either do not have strong enough explanatory power to suggest robust correct structural/hierarchical orders, or indicate incorrect ordering of concepts within the set (although the findings are not significant either). The two "good" sets have a reasonably good proportion of correct responses also, being recognized as correct by 27 and 22 of 30 subjects.

Chapter 6

Recall

The experimental method follows the description in Chapter 4. This chapter reports on issues specific to the recall experiments. It presents the designs and findings of the replication of Hirshman's study [42] and the recall test trials. The replication study was conducted as Part 1 of Experiment 1, the same 30 participants as in the SVT experiments were used here. The recall test trials used a separate group of 60 subjects and was referred to, in Section 4.3, as Experiment 2.

6.1 Evaluation assumptions

Based on Hirshman's study [42](Appendix B), we identified two effects which may occur under different study conditions. For each effect, we can modify the study conditions to make theoretical assumptions, to examine the correctness of ontological structure: A ← B ← C. We hereby introduce the two effects and their study conditions.

Familiarity effect When observing for the *familiarity effect*, strongly related word pairs[1] have significantly better recall performance than weakly related word pairs. This effect is evident when using a cued recall design, in both the pure list and mixed list conditions. Although a significant effect was found under both the pure list and the mixed

[1]Strength of association in Hirshman's study [42] was determined from published norms [12, 62]. Response words in strongly related pairs are the most frequently given associates of the stimulus word. Response words in weakly related pairs are generally very infrequently given associates of the stimulus word.

list conditions, a stronger familiarity effect was observed when under the pure list condition.

Expectation-violation effect When observing for the *expectation-violation effect*, weakly related word pairs have significantly better recall performance than strongly related word pairs. This effect is evident when using a free recall design, in the mixed list condition.

Descriptions of the study conditions are presented as follows.

Cued recall vs. free recall Cued recall: During the recall test, the stimulus term of each word pair in the stimulus list was given. Subjects wrote the response word in a space provided to the right of the stimulus term. Free recall: During the recall test, subjects were each given a piece of blank sheet without cues [42, 43].

Pure list vs. Mixed list Pure list: Subjects are assigned to study either the list of strongly related pairs or the list of weakly related pairs [42]. Mixed list: Subjects are assigned to either a stimulus list with the odd-numbered response words (in terms of list position) being strong associates and the even-numbered words being weak associates, or the other stimulus list with the even-numbered response words being strong associates, and the odd-numbered words being weak associates. The mixed list design will be described more in Section 6.3.2 and also in Appendix B.

Hirshman provided a theoretical explanation for the differing findings of the two stimulus list designs. The author stated that retrieval interference occurs when the retrieval of some items at test makes it impossible to retrieve other items that otherwise would have been retrieved [42] (see [92] for empirical evidence of retrieval interference). This explanation claims that the retrieval of weakly related pairs in the mixed list design blocks the retrieval of strongly related pairs that otherwise would have been retrieved. When a pure list design is used, and the blockage is removed, performance on strongly-related pairs increases. Theoretical explanations of retrieval interference claim that retrieval interference occurs because general contextual memory cues (i.e. environmental and temporal cues) become less effective, or overloaded, as retrieval proceeds [36, 96].

> Some studies on word frequency effect[2] (e.g. [97, 56]) found that in free recall tasks, when low- and high-frequency items are mixed within the to-be-remembered lists, the usual recall advantage found for high-frequency words is eliminated or reversed.

In this study, we chose to test for the expectation-violation effect only. This is because the familiarity effect obtained in Hirshman's study [42] was tested in a cued recall experiment design carried out immediately after the free recall test (where expectation-violation effect was found). Subjects' cued recall performances might therefore be affected by the prior free recall test. This is further discussed in Section 9.

In Hirshman's study [42], results that support the expectation-violation effect were found in the free-recall mixed-list condition. With this experimental design, we can assume that in an ontological structure $A \leftarrow B \leftarrow C$, statements with the subset relationship of weakly related concepts (S2-Type sentence - "every C is an A") would be better recalled (more frequently recalled in the free recall test) than statements with the subset relationship of strongly related concepts (S1-Type sentence - "every B is an A").

The strength of the subset relationship between the two concepts in a statement, which we call "associative strength", would be the independent variable; while the number of correct recalls (out of the total number of participants) of each stimulus (sentence statement studied) would be the dependent variable we measure.

In the next two subsections, we describe two analysis designs for evaluating the quality of the two ontologies using the recall technique.

6.1.1 Ontologies

We compare the overall quality of the two ontologies by conducting a within-subject ANOVA with factors "Ontology" (2 levels: SUMO and WordNet) and "level" (2 levels: S1 and S2). This is achieved by comparing, between ontologies, the difference between the mean proportion of the correct responses of S1-Type sentences and the mean proportion of the correct responses of

[2]The word frequency effect (WFE) suggests a frequency paradox, in which common words are more easily recalled than rare words, but rare words are more easily recognized [1, 53]. It shows how familiarity of prior knowledge influences recall performance (thus can support this study).

S2-Type sentences (S2–S1–CR). This can be achieved by examining the interaction effect of the two factors above. The analysis will provide us with results of the two main effects. The first being whether one ontology is easier to recall than the other ontology, regardless of the levels. The second will be whether S2 is better recalled than S1, regardless of ontology. An interaction effect between these two factors will also be examined. We included the differences in average word frequency (AVG–Diff) between S1 and S2 sentences as a covariate.

This analysis provides us with a summarized comparison on the quality of the conceptualizations of the two ontological structures; and thus an overview of the quality of the ontologies about how well these ontologies represent a human's knowledge structure. Also, this analysis illustrates the overall correctness of the two ontological structures.

Under the design of the expectation-violation effect, a higher mean proportion of correctly recalled S2-Type stimuli than S1-Type stimuli (positive S2–S1–CR) implies a good ontological structure. We thus compare the quality of ontologies by examining the interaction effect which allows us to test whether the level difference is significant between the ontologies. The ontology with a more positive S2–S1–CR value is the better ontology.

6.1.2 Sets

We examine the quality of each set included in the evaluation. This is done by identifying in both ontologies, sets that have significant positive S2–S1–CR values (indicating the concepts are ordered in the correct structure) and sets that have significant negative S2–S1–CR values (indicating incorrectly ordered concepts, hence erroneous structures).

In this analysis design, the independent variable is the nodes of separation that is manipulated in a within-subject design. While the dependent variable is the number of correct recalls for each sentence stimulus.

This analysis allows us to examine the correctness of the order structure of the concepts in each set, and therefore gives us a more detailed understanding of the quality of the two ontologies. It is important in evaluating ontologies especially if the quality of the ontological sets varies. This is because the mean S2–S1–CR value of an ontology obtained in the analysis as described in Section 6.1.1 averages across the S2–S1–CR values of its individual sets.

6.2 Replication study

Hirshman presented subjects with pairs of words that were either strongly related or weakly related (Experiment 1 in [42]). After a 5-minute retention interval, subjects freely recalled the response terms from the word pairs in a free recall test. Following this, they were given a cued recall test where the stimulus terms from the word pairs were given as cues for the responses. Hirshman's study [42] attempted to replicate the results of Hirshman and Bjorks' study [43] with new materials. See Section 3.2.5 for an introduction on [43]. Hirshman's study [42] was used because the materials and findings are available for us to effectively and reliably replicate the study. The experimental design of the Experiment 1 of Hirshman's study [42] is presented in Appendix B.

The design and procedure of this replication study were based on Hirshman's study [42], but modified as listed below:

1. Microsoft PowerPoint was the software program used for the experiment because it allows more control and more standardization from one sentence presentation to another. First, PowerPoint slides were used for the sentence presentations, as it allowed us to pre-set the length of time each slide/sentence was shown (10 seconds). Second, instructions were presented to subjects in written form as the first PowerPoint slide. See Appendix M for the instructions for the test trials, the instruction for the replication study was exactly the same except that there were 19 rather than 16 sentences in the replication study. Third, instead of a "turn" command spoken by the experimenter, a sound was pre-set to ring at the beginning of each slide presentation as an indication of the presentation of a new sentence.

2. We eliminated or altered some parts of Hirshman's study [42] in an attempt to shorten the length of the replication study in order to avoid possible fatigue effects in the SVT experiments, as it was only one of the three parts in our Experiment 1. First, we felt that practice trials were not necessary because the task was not complicated and practice trials may give away the purpose and measure of the study. Second, we omitted the cued recall test at the end of Hirshman's study [42] from our study design since in this study we tested only for the free recall condition. Third, the distraction task in Hirshman's study [42] was 5 minutes of word-search-task; whereas in this study, subjects were

asked to play 3 minutes of Sudoku. We reduced the time to 3 minutes to shorten the length of the replication study. Also, this was to be consistent with the experimental procedure of the test trials. Another reason for having a shorter distraction task was that the stimuli in this study might be more difficult to recall as they were in sentence form, with concept terms that were on higher levels of abstraction in comparison to those in [42].

3. Instead of word pair presentation, each stimulus was presented as a sentence. For example, "Grass and Green are related". Also, instead of asking the subjects to recall the response word members of the word pairs they have studied; subjects in this study were asked to recall whole sentences. These modifications were necessary because stimuli in the test trials were in the form of sentences. It was important we showed that the recall technique used in [42] was transferable in studying sentence-form stimuli, as subjects in our Experiment 2 (test trials recall technique) were asked to recall whole sentences that state subset relationship.

4. Consistent with Hirshman's study [42], we took primacy and recency effects into account. However, whereas Hirshman [42] omitted the first two and the last one studied word pairs from analysis, we, in both replication study and in the test trials, excluded the first two and the last two sentences. It was observed in our pilot test (Section 4.4.2), that the second last stimulus was still prone to recency effect. Thus the sentence stimuli constructed from the following word pairs were excluded from analysis: "Colour–Green", "Grass–Green", "Patch–lettuce", "Cabbage–Lettuce", "Worm–Bug", "Insect–Bug", "Room–Home", "House–Home" (refer to Table B.1).

6.2.1 Result

Table 6.1 shows the mean proportions of correct responses (in the free-recall condition) for strongly related and weakly related pairs in Hirshman's study [42] and in the replication study. Consistent with Hirshman's findings, the recall of responses from weakly related pairs is superior to recall of responses from strongly related pairs in free recall. However, the superiority was found significant in Hirshman's study [42] ($F(1, 23) = 5.28, p < .05$) but not in our replication study ($t(15) = 1.503, p = 0.155$).

Study	Strongly related pairs	Weakly related pairs
Hirshman [42]	.23	.34
Replication study	.33	.395

Table 6.1: Proportion of correct responses recalled as a function of associative strength - Hirshman's study [42] vs. replication study

The non-significant results in this replication study, however, can be explained by the low power value (0.2686) in the power test. These results may become significant in a higher-powered test, which can be achieved by having a larger sample size. Since similar findings to Hirshman's study [42] were obtained in this replication study, and the results are likely to be significant with a higher-power test, we are confident in the validity of the recall technique for ontology evaluation.

One interesting finding is that the replication study obtained higher mean proportions of correct responses than in Hirshman's study. This might be due to the shortened distraction task, or maybe the other differences in the experimental designs such as having the $50 prize as an incentive to perform better.

6.3 Test trials

6.3.1 Subject design

As found in Hirshman's study [42], the expectation-violation effect is strongest when using a mixed list design. In this study, we included both statements with strongly related concepts (S1) and statements with weakly related concepts (S2) in the stimulus list for all participants.

In the replication study, subjects were tested individually, and in the test trials subjects were run during tutorial sessions in groups of 17-22. It should not matter if the test was conducted individually or as a group. Since the time for study and the time for recall were pre-set and were standardized, we do not face the problem observed in Fang's pilot tests [28, 27] where subjects who took longer time to complete the experiment were intimidated by others who finished faster and often rushed through the rest of the trials. Also, all instructions were in written form and therefore minimized the difference in experimenter effect running experiments individually or in groups. No

distraction by other group members were observed or reported during the pilot test.

6.3.2 Material

The stimuli used in Hirshman's study [42] and those used in this study might be processed differently. Hirshman tested the degree of relation of the concepts in the word pairs, and subjects were not limited in the types of association when processing the word pairs. Whereas in this study we tested the degree of relation of two ontological concepts in sentence form and the semantic association between concepts was limited to the subset relationship.

We base our work on Hirshman's study [42], even though the stimuli are different. This is because the familiarity effect implies that more familiar items can generally be remembered better (and recalled better). This effect occurs in other types of to-be-remembered items. For example, the word frequency effect shows how familiarity of prior knowledge of individual words influences recall performance. It is a well established finding that high frequency words are better recalled than low frequency words [24]. Familiarity represents a generalized feeling of prior occurrence [24], and we can assume that this applies to our stimuli too. The same applies to the expectation-violation effect, where the effect arises when expectation (based on people's prior knowledge) is violated, regardless of the forms of stimulus presentation (sentence or word pair).

The same 24 critical sentence stimuli studied in the SVT test trials were used in this experiment (6 sets selected from each of the two ontologies, one S1 sentence and one S2 sentence were constructed from each set). We also selected 2 additional ontological sets from each ontology as fillers. They were used to construct the first two and last two sentence stimuli presented which were to be omitted in the effort to control for the primacy and recency effects. The four filler sets are "Communication – Disseminating – Advertising", "Commerce – Business – Tourism", "Substance – Mixture – Blood", and "Part – Organ – Muscle".

Four alternative lists (2 random presentation order $\times$ 2 alternation of strength association, see Table 6.2), each comprised of 16 sentence statements, served as the to-be-remembered materials; that is, 12 critical sentences and 4 filler sentences.

It is important to modify Hirshman's study [42] and include two different sets of presentation orders, because it is a well established finding that there

	Strength Association Alternation 1	Strength Association Alternation 2
Presentation Order 1	List A	List B
Presentation Order 2	List C	List D

Table 6.2: Design of recall stimulus list - presentation orders × strength association assignment

are recency and primacy effects in the studies of serial recall. This means that the first few stimuli and the last few stimuli presented to subjects are usually recalled better than stimuli in the middle of the list, even after excluding the first two and the last two stimuli in analysis. Also, according to the expectancy-violation effect, if sentence stimulus x is constantly placed immediately before or after sentence statement y - which states two concepts with the weakest subset relationship in the stimulus list; the power of stimulus x might be under estimated due the comparison influence of stimulus y.

The 16 subject terms ("every S is a P") in each list are the same words but of two different presentation orders (List A & B vs. List C & D in the Table 6.2). In List A and C, the odd-number stimuli (in terms of list position) are strong associates (S1 of the ontological set) and the even-numbered words are weak associates (S2 of the ontological set) whereas in List B and D the odd-numbered response words are weak associates and the even-numbered words are strong associates. If the S1 statement of a set is on one list (e.g. List A), the S2 statement of that set would be on the other list of the same presentation order (List B). Also, it is important to have corresponding strength associations on each presentation order; for example, if the fifth stimulus on List A is S1 of the ontological set X, List B would have S2 of the ontological set X also as the fifth stimulus; this is so the recall performance of S1 and recall performance of S2 of the same ontological set are not biased by the difference in presentation order (recency and primacy effects). Each of the four alternative lists is presented to 1/4 the subjects. The four sets of the stimulus list (2 presentation orders × strength association assignment) are counterbalanced over the participants. See List A and B in Appendix L.

6.3.3 Procedure

Subjects were given an information sheet (see Appendix H) and a consent form (see Appendix J). The experimenter gave a brief verbal introduction to the study and the nature of the experiment, but in no more detail than what is in the content of the information sheet. Special emphasis was given to the voluntary and confidential nature of the study. The experimental procedure for this study was largely the same as the procedure of the replication study (Section 6.2)

Study phase Participants were each given a note pad. The subjects were then shown 16 sentence statements one after the other on a computer screen, and were instructed to write each statement on a sheet in their note pad during the 10 seconds during which each sentence statement was shown. A sound would be made at the end of the 10 seconds which was also the beginning of the presentation of the next sentence, indicating that the subjects should turn the page in their note pads and write the new sentence on the next page. All instructions were in written form on the first slide (see Appendix M) before the sentence presentation. Written instructions are said to be more precise and consistent among experiments, and can also minimize the experimenter effect. We did not specify to the subjects the type of statements in the study - that is, the subset relationship of two concepts. This was to avoid having subjects guessing the nature of the study as they might notice that some subset relationships were stronger than others.

Distraction phase After the last sentence presentation, the experimenter took the note pads from the subjects and gave each of them a sheet with two Sudoku exercises (See Appendix K) in order for each of the subjects to be distracted for 3 minutes. Instructions on how to play Sudoku were printed at the top of the sheet.

The distraction task should ensure that every subject is fully and equally distracted for the same amount of time, minimizing the risk of subjects having different levels of rehearsal before recalling. Sudoku is an appropriate exercise that serves the distraction purpose.

Recall Phase Following the distraction task, each participant received a blank sheet and were asked to recall the presented sentences they had studied; after which they were collected after 3 minutes. It was found in Hirshman's

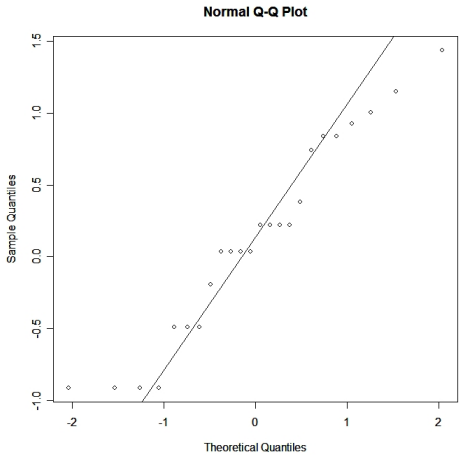

Figure 6.1: Q-Q plot of the number of correct responses (standardized and transformed data)

subsequent experiment that the findings did not change when recalling a response word and when recalling both words in a word pair. Therefore we should be able to ask for the whole sentence to be recalled instead of recalling only one of the two concepts in the sentence like that in Hirshman's experimental design.

6.3.4 Result

Because the examination of the data showed that the data was distinctly skewed and non-normal ($Skewness = 1.075515, p_{SW} = 0.00486, p_{KS} = 1.229e{-}05$), a logarithmic transformation was applied which provided a substantial improvement ($Skewness = 0.029937, p_{SW} = 0.2201, p_{KS} = 0.4122$). The QQ-Normal plot also showed normality (Figure 6.1).

Using a within-subjects ANOVA with factors "Ontology" (2 levels) and "level" (2 levels: S1 and S2), we compared the proportion of correct responses between S1 sentences and S2 sentences (S2–S1–CR), then analyzed whether the level difference (S2-S1) was significant between the ontologies. We included the average word frequency (FREQ–AVG) as a covariate. The ANOVA table is shown in Table 6.3.

We found a significant difference between the two ontologies on the number of correct responses of the sentence statements (Figure 6.2). The mean

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Sig
Ontology	1	3.3260	3.3260	8.4791	0.008942	**
Level	1	0.2284	0.2284	0.5822	0.454815	
FREQ–AVG	1	0.2352	0.2352	0.5997	0.448224	
Ontology×Level	1	0.0512	0.0512	0.1305	0.721946	
Residuals	19	7.4529	0.3923			

Table 6.3: ANOVA Results

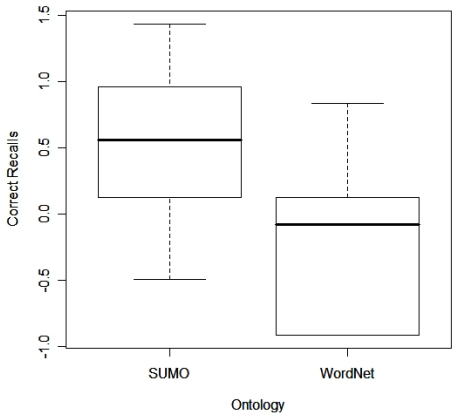

Figure 6.2: Boxplot of the number of correct responses, by ontology (standardized and transformed data)

proportion of the correct responses of the SUMO sentence stimuli was 0.2417 (SD=0.1730), and 0.0806 (SD=0.0870) for WordNet stimuli. This indicates that the sentence statements constructed from WordNet are overall harder to recall than those constructed from the SUMO ontology. According to the expectation-violation effect, WordNet (regardless of the levels) has more correct and close relations between the related concepts specified. Therefore, WordNet has a better quality of its formal specification of the domain conceptualization, after controlling for the word frequency effect.

There is no significant difference between the proportion of correct responses between the two levels (S1 and S2). The mean proportions of correct responses from the 12 S2 and 12 S1 sentence statements were 0.13 (SD=0.13) and 0.19 (SD=0.18) respectively. Although the finding is not significant, the

Ontological sets	Proportion of correct S2 resp.	Proportion of correct S1 resp.	S2–S1–CR
SUMO Sets			
Object – Agent – Nation	0.1	0.033	0.067
Contest – Game – Sport	0.033	0.6	-0.267
Process – Motion – Walk	0.033	0.133	-0.1
Motion – Radiation – Music	0.367	0.133	0.233
Process – Creation – Cooking	0.267	0.3	-0.033
Artifact – Text – Book	0.167	0.433	-0.267
WordNet Sets			
Organization – Unit – Union	0.1	0.3	-0.2
Activity – Diversion – Escape	0	0	0
Change – Move – Step	0.133	0.133	0
Perception – Sensation – Noise	0	0.1	-0.1
Act – Change – Reversal	0.033	0	0.033
Work – Publication – Magazine	0.067	0.1	-0.033

Table 6.4: Concept sets for sentence pairs used in this study, proportion of correct S2 responses, proportion of correct S1 responses, differences in proportion of correct responses (S2–S1–CR)

means suggest a negative S2–S1–CR value, indicating an overall opposite subset effect (level 1 and level 2 concepts are in a reversed order) in the sets constructed from the two ontologies. The assumption was that a correctly structured set should have a positive S2–S1–CR value. The non-significant level effect indicates that none of the 12 sets included for examination had significantly different S2 and S1 scores. Table 6.4 shows that only 2 out of six sets (Object – Agent – Nation, Motion – Radiation – Music) from SUMO and 1 out of 6 sets from WordNet (Act – Change – Reversal) have positive S2–S1–CR, however none were significant.

In our analysis, the Ontology×Set interaction effect was also found to be not significant. This means that the difference between S1 scores and S2 scores (not significant, as analyzed previously) does not differ between ontologies. Furthermore, the difference between ontologies, which was significant, was not significant on the two levels (S1 and S2) individually.

The average word frequency (FREQ–AVG) does not contribute variability to the number of correct responses.

6.4 Discussion

The findings of the replication study and those of Hirshman's study [42] are consistent - there is a higher proportion of correct recalls from weakly related pairs than the strongly related pairs. Although the subset effects (positive S2–S1–CR) found did not achieve significance in the replication study, we are confident in the validity of the recall technique, and that it is an appropriate technique for ontology evaluation. This is because the power value found in this replication study is low (0.2686), and the results are likely to be significant if it was in a higher-powered test, which can be achieved with a larger sample size. Similar to the SVT replication study, the power is low because the effect size is small. With a larger effect size, the power value would have been sufficiently high. Again, we expected a larger effect because we based our study on Hirshman's study [42]. Also, we have more subjects in our replication study (30, 15 in each group) than Hirshman's study [42] (24, 12 in each group) and therefore expected to obtain a large enough power for the replication study. We could not calculate Hirshman's [42] effect size and power to compare with those found in the replication study, because the standard deviations were not provided.

The findings of the recall test trials suggest that there is no significant difference between the proportions of correct responses for S2 stimuli and S1 stimuli overall; both for each ontology individually, and for each individual set examined, after controlling for possible word frequency effect. This implies that the expected subset effect was not obtained in any of the sets selected for evaluation; that the quality of the two ontologies' formal specification of the domain conceptualization was poor; and also that the quality was poor for each set examined and was indifferent. The word frequencies themselves were found to contribute no variability to the number of correct responses received.

There is a considerable variation between ontologies on the overall proportion of correct responses, regardless of the levels. It appears that WordNet has better specifications in general (see Figure 6.2), as the expectation-violation effect states that strongly related stimuli have weaker recall performance than weakly related stimuli. However, this superiority in quality does not include the specifications of conceptualization of structural order because there was no level difference (S2-S1) found.

Chapter 7

General discussion

In this study, we replicated one study for each of the two selected techniques (SVT and recall) in order to validate the studies and theories we based our experimental design and procedure on. Neither of the replication studies' results achieved significance. However, since our replication studies for both techniques obtained findings that pointed towards the same direction as the original studies, and that the power tests presented low power values which were likely to be the reasons for the non-significant results found in the replication studies, we are confident that both the SVT and the recall techniques are valid techniques for ontology evaluations. We argued that the low power levels found in the replication studies are due to the small effect size. The power would have been sufficient with a larger effect size. Moreover, both replication studies have larger sample size than the original studies [70, 42], and therefore expected a large enough power.

Nevertheless, we cannot discard the possibility that the non-significant results found in our replication studies indicate that the previous studies we based the techniques on have flaws, and significant results were found when there should not have been. Another interpretation of this finding is that the modifications made in the experimental designs and procedures in the replication studies (refer to Section 5.2 and 5.3 for the SVT technique, and Section 6.2 for the recall technique) resulted in the differences in the findings of the previous studies and our replication studies. For example, the stimuli used in the previous studies for both techniques were in the form of word pairs, whereas stimuli used in the replication studies were in the form of sentences. This change might have resulted in stricter mental processes when subjects were processing the stimuli, because it allows inferences to be

made on the subset relationship only, and the processing of sentence stimuli might have somehow reduced the perceived level difference (between S2 and S1). Another example might be that we modified from Rips et al.'s study [70] to exclude the 7-second intertrial interval where subjects were informed if he/she has made an error in the sentence verification. This might have also somehow reduced the perceived level difference. Whether the differing results between the replication studies and the original studies were due to the modifications made, needs to be further investigated in future studies.

Future studies can do two experiments with two groups of subjects. For each technique, one group could be tested with the non-modified replication study and the other group with the modified replication study. If the results in the non-modified replication study are still not significant or are significantly different from those presented in the original study, then exploration of the first possibility (the original study is flawed) is needed. In this case, we will need to examine other studies that have used the technique and base the evaluation technique on the design and procedure of another more appropriate study; otherwise conduct research for another evaluation technique. If the results in the non-modified replication study are the same or largely similar to the original study, we will need to explore modifications that might be able to make, as those will be important indications as to how the test study should be designed.

In our analysis, both techniques had "the difference between S2 and S1" as our point of analysis, and the correctness of structural order between the three concepts in a set were our measure of Quality. Our assumption was that if the findings of the SVT technique and the recall technique converge, then the analysis of the quality of ontologies is robust. This is an exploratory study and it verifies the usability of the multi-method methodology for evaluating ontologies. The experimental evaluation procedures and analysis designs presented in this study also serve as templates for future researchers and ontology evaluators to assess ontologies using the multi-method methodology.

We now compare the findings of the two techniques, and the converging findings are as follows. Neither of the two ontologies that we examined is a good specification of the conceptualization of the domain, as was suggested by our findings from both techniques. Both the terminology and the structure of the ontologies, may benefit from improvement.

While our method is primarily intended to investigate overall differences between two ontologies, it is dependant on the sampling of the concepts to test. Our method is also able to identify those sets of concepts that are struc-

tured correctly and those that are not. The measures of the two techniques, verification time differences, error rates, and the differences in proportions of correctly recalled responses, can all be used by an ontology creator or ontology maintainer in order to identify problematic concept sets and to improve their structure. For example, the sets with significant negative S2–S1–TD and S2–S1–CRs should be considered for improvement in future versions of the ontologies. In our study, no such set was discovered. However, not one set with a significant positive S2–S1–TD and S2–S1–CR was found either, which means that none of the 12 sets (6 from each ontology) robustly represents correct structural order. The only two sets with positive significant values obtained were "Process – Motion – Walk" and "Change – Move – Step" in the SVT experiment; however this finding was not supported in the recall experiment. Negative and zero S2–S1–CR values were found for these two sets in the recall experiment, which presented opposite and contradicting effects, although the effects were not significant.

We also suspected that the level of word familiarity of the concept terms included in the statements for evaluation may have affected the response time in the SVT and the proportion of correct recalls in the recall experiment. In both experiments we included word frequency as a covariate, and the findings from both experiments suggested that word frequency did not have a significant effect on our findings. This allows us to confidently say that negative values of the S2-S1 mean indicate problems with the hierarchy and positive values of the S2-S1 mean indicate quality sets with the correct structural order.

There is an interesting finding to note however. According to the recall technique's result, the mean proportion of correct recalls was significantly smaller for WordNet sentences than SUMO sentences. This finding indicates that in general the association between two related WordNet concepts appears to be stronger than two related SUMO concepts that have the equivalent abstractness level and level gap (number of levels apart). This suggests that WordNet in general represents a better formal specification of the domain conceptualization, however this superiority only applies to the association between the two concepts, and does not include the correctness of the structural order, as no significant difference was found between S1 and S2 in either ontology. This analysis was not available in the SVT technique and therefore we can not fully verify this finding using the multi-method methodology. However, this is an interesting area to explore further.

We compared our findings with those in Fang and Evermanns' work

[28, 27, 26]. In these studies, the SVT was the only evaluation technique used. Initially, the BWW and SUMO ontologies were the two ontologies evaluated. The concepts selected for evaluation were on the highest levels of abstraction for these upper level ontologies, and some concepts that were included were compound terms [28, 27, 26]. A follow up study was conducted in [26] to justify the possible issues identified in the earlier studies [28, 27]. Specifically, the inclusion of concepts from higher level of abstraction and the inclusion of compound terms. As a result, compound terms were excluded from stimuli selection, and SUMO and WordNet were adopted as the ontologies evaluated because in comparison to BWW, WordNet provides concepts on the lower (more concrete) levels of abstraction. It was found in both the earlier studies and the follow up study that these ontologies do not differ in terms of cognitive quality, and none of the three ontologies is a good specification of the conceptualization of the domain. Moreover, the follow up study did not find any evidence that a different level of abstraction has an effect on the study results, and the effect that inclusion of compound terms has on the results is also vague. However, we suspected that this might be due to the fact that the follow up study has a low power.

In this study, we followed the designs of the follow up study, but made stricter the process of stimuli selection. The concepts selected for evaluation were of and close to the basic level, to be more consistent with the original study [70] from which the technique was adopted from. The results we found were consistent with those in the follow up study. This supports the finding that neither of the two ontologies (SUMO and WordNet) properly represents perceived reality. Also, we again found no evidence that a different level of abstraction affects the results, which suggests that the SVT technique is robust across different abstraction levels. We rely on future studies to further examine whether the SVT technique has a wide application domain, across different abstraction levels (see Section 9.1).

Chapter 8

Limitations

In this chapter, we consider the possible biases and limitations in this study and suggest further studies where more investigation is needed.

8.1 Multi-method

The multi-method methodology has the advantage of cross checking the findings from multiple techniques in order to increase the validity of the findings obtained. Most of the findings are consistent between techniques. For example, overall, neither of the two ontologies has a good formal specification of the domain conceptualization.

However, as we make the benchmark of quality higher and the criteria more strict, the evaluation methodology may run into the threat of discounting and overlooking some positive findings such as good ontological sets. For example, the two sets that were found correct in the SVT experiment were disproved as the findings in the recall experiment did not support those positive findings. It was thus concluded that the quality/correctness of these two sets were not high enough to pass the test of both techniques.

There are findings that lead to a possible implication that one or more of the two techniques used are not suitable to be adopted as ontological evaluation tools. First, there are inconsistencies in our findings from the two techniques, such as the difference in the good sets found. One possible explanation might be that the two techniques are based on two different cognitive mechanisms as discussed in Section 3.2. It is possible that processing ontological concepts (which might still be on higher abstraction levels) uses

only one mechanism but not the other as had been presumed.

Another possible implication that one or more of the two techniques used was not suitable is that the non-significant results were found in the replication studies for both techniques. Although we believe that this was due to the lack of power in the analyses and that both techniques have high validity, we cannot discount the possibility that significant findings in the replication studies for one or both techniques might not be achieved even with higher-powered tests. Future studies should conduct the replication studies and the SVT and recall experiments with a larger sample size to better ensure the validity of these techniques and therefore the feasibility for their adoption.

8.2 Memory structure

Rips et al. [70] brought up an interesting finding about the SVT technique that is important to be investigated further and be taken into account - that memory structure does not necessarily mirror logical structure (refer to Appendix A). The findings in our replication study point in the same direction even though they are not significant - the opposite subset effect found in the Mammal category. These findings suggest that although the SVT technique has been used extensively in the studies of knowledge representation and categorization and the outcomes in those studies have been consistently proving the SVT as a valid measurement (e.g. [18, 17]); the knowledge representation the SVT is suitable to test is the representation in our memory structure but not necessarily a structure that represents logically valid relations such as the specifications of a formal ontology. However, we argue that a good ontology should represent the world we perceive and conceptualize, thus the quality of ontologies we tested for should really be based on the memory structure of the world mapped in our mind, and not a logical structure. The re-establishment of the quality of ontology leads to the argument that some of the negative or non-significant findings in our SVT experiment might not suggest "incorrect" sets or structural order, but rather, sets that mirror logical structure but not our memory structure. We hereby propose two possible approaches to differentiate the incorrect sets from the sets that are logically correct but fail to mirror that of our memory structure.

First, the results obtained from the second technique, which in our study is the recall technique, can be used to help us distinguish sets that are logically correct but fail to mirror our memory structure, and sets that are

incorrect both in terms of the logical structure and our memory structure. If a negative finding is obtained in both SVT and recall techniques, the set can be assumed to be an incorrect set. If a negative finding obtained in the SVT is found to be positive in the recall technique, the set is likely to be a logically correct set that does not mirror our memory structure. In our study, no such set was found since none of the sets were found to be significant in the S2–S1–CR scores.

Second, in the subsequent experiment (Experiment 2 in [70]), Rips et al. obtained ratings of semantic distance. Rated distance could be conceptualized as the distance derived from a semantic space which mirrors more closely to people's memory structures. They revealed that the previously obtained subset effect only occurred when the rated semantic distance between the instance and its immediate superordinate was less than that between the instance and its higher level superordinate. This suggests that the subset effect was mediated by variations in rated distance. We were unable to include the rated distance into our evaluation process in this study due to time and subject number constraints. The rated distance can be another tool, other than a second technique as proposed above, that can be used to differentiate the incorrect sets from the sets that are logically correct but do not mirror our memory structure.

8.3 Stimuli

The scope of the modern ontologies is too large for us to examine and evaluate every possible combination of concepts. Thus, to appropriately evaluate the cognitive quality of an ontology, the sampling of concept sets from an ontology is critical. For our methodology, the sampled concept sets needed to satisfy certain criteria with respect to length of terms, compound terms, related by only subset relationship, level of abstraction, domain focus and so on. This reduces the number of possible concept sets that we can select from, and this is especially a problem for smaller ontologies. Also, with these criteria, the stimuli selection process can be a lengthy one. These criteria are in place to ensure the validity of the evaluation methodology and the comparability of two ontologies for the evaluation. They also set the guidelines for the ontologies that the multi-method methodology is applicable to.

At this point, our method is designed to examine only the subset re-

lationship between concepts, and not other types of relationships, such as object-property relationships. While the semantic network model of human cognitive concepts makes predictions with respect to properties of concepts, the ontology and thesaurus examined here make little use of properties or other relationships. However, other ontologies may extensively use such features, and future research in this direction is therefore important.

Furthermore, although the concept terms were selected to be as close to the basic level as possible, some of the concept terms included for evaluation are still not concrete terms, and each of these terms may have multiple inferences that can be made. This ambiguity in the interpretation of these terms may become a limitation as subjects' responses may vary depending on which meaning of a term was activated at the time of the testing. This may potentially affect the RT, correctness, and the recall of participants. Such a limitation can be addressed by using other methods based on, for example, pictorial representations of concept instances. Initial ideas for such methods are briefly presented in [25].

8.4 Semantic distances between two levels

The construction of stimuli was carefully designed in this study so we have comparable sets between ontologies, and this includes having the same number of level gaps between concepts in a SUMO set and concepts in its corresponding WordNet set. However, there is one other factor that we should also take into account - the difference in semantic distance between two levels, which can be different between ontologies. For example, two related concepts (Animal and dog) can be found one level apart in one ontology (Ontology B), and two levels apart in the other ontology (Ontology A) which has smaller semantic distances between levels (refer to Figure 8.1).

It was found in the recall experiment that sentence stimuli constructed from WordNet have a significantly smaller number of correct recalls than stimuli constructed from SUMO, and it was concluded that WordNet in general has better specifications (refer to Section 7). However, this finding can also be an indication that WordNet has a small semantic distance between the two levels. In order to address this issue, future studies should control for this factor by calculating the relative semantic distance between two levels in Ontology A, using Ontology B as a standard, and make the relative semantic distance value a covariate in the analysis. This can be done by randomly

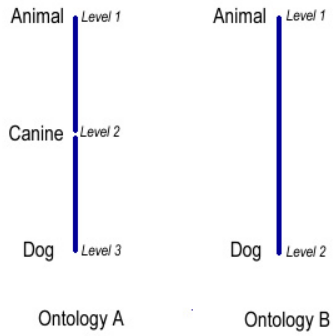

Figure 8.1: Demonstration of relative semantic distance between levels

selecting 10 sets of 2 related concepts that exist in both Ontology A and Ontology B. For each concept pair (Animal and Dog) we set the semantic distance between 2 levels in Ontology B as 1, and since Animal and Dog are two levels apart in Ontology A, the relative semantic distance between 2 levels in Ontology A is 0.5. As a covariate in the analysis, the relative semantic distance value for Ontology B would be 1, and the relative semantic distance value for Ontology A would then be the calculated mean of relative semantic distance for the 10 sets.

8.5 Test subjects with specific knowledge areas

The results obtained for some sets might allow alternative interpretations other than being a reflection of the correctness of categorization structures in the ontology. Some might argue that the results of S1 and S2 difference obtained for some sets is skewed by a different choice (or more consistent choice) of test subjects with knowledge within a specific domain from which the sentences are derived, For example, for a person who is familiar with the Business Process Management field, an S2 sentence such as "every Cooking is a Process" takes minimal time to assess, while he/she might take longer to assess its S1 counterpart "every Cooking is a Creation". This does not necessarily mean that the ontology structure is incorrectly designed, but it reflects the subjects' own personal knowledge as well as their personal

beliefs. The future study should take into account this possible sampling bias that we might not have detected. This possible bias can be mitigated by including a larger sample size. This reduces the influence one test subject with knowledge within a specific domain from which the sentences are derived have on the overall result. Another control for this possible sampling bias can be to enforce the balance of word frequency between S1 and S2 sentences for each sample set, and to avoid the use of low frequency concept words. This is because it is likely that words with higher frequency are more general and words with lower frequency are more knowledge/domain specific. A third approach to address the possible bias is to sample more sets from the ontology for evaluation. This should reduce the overall effect of sampling bias, although not the effect for individual sentence pairs.

Chapter 9

Future studies

Other than the areas that need to be further investigated as suggested in the previous section, there are also other interesting areas we can explore.

9.1 Explore the boundaries of the methodology's application domain

To examine the applicability of the multi-method methodology, we replicated two psychology studies to verify the validity of the two selected techniques. The next step should be to explore the boundaries of the application domain of the methodology by identifying the common application domain that the two techniques share, focusing on the level of abstraction of the ontologies. It is important to know which types of ontologies, or concepts on which levels of abstraction the methodology is applicable for. In this study, we evaluated the ontologies by selecting the concepts around the basic level which we assume the techniques are applicable for, as the techniques were used on concepts of that level in the previous studies.

It is important to examine the application domain of this methodology because ontologies come at many different levels of generality. For example, upper-level ontologies such as SUMO and BWW are on a high level of abstraction, are not domain specific, and concepts included are mostly basic concepts for human understanding of the world. For this reason, we cannot afford to be limited to evaluating only the basic level concepts. Identifying the application domain is also important so we do not apply the evaluation methodology on concepts of the non-applicable abstraction level, nor

ontologies that do not include the applicable abstraction levels.

We can categorize concepts by their levels of abstraction. This can be done by calculating the number of levels from the psychologically basic level of categorization [73].

In order to explore the methodology's application domain, we shall examine the applicability of these techniques on concepts on two different levels of abstractions (higher level and basic level concepts). Two approaches for extracting concepts of different levels of abstractness are possible:

Test on different ontologies of different abstraction levels

We can categorize existing ontologies by their levels of abstraction, and then explore the methodology's application domain by applying the selected techniques on selected ontologies of each different level of abstraction.

This can be done by first, identifying where the basic level concepts are in an ontology. Second, draw a chart with the level of abstraction as a function, and then position the ontology on the appropriate levels, using the basic level as the baseline. Third, select two ontologies positioned on and around the basic level and two ontologies from the higher levels of abstraction. Finally, we can apply the two evaluation techniques to compare the quality of the two ontologies on each level of the abstraction examined.

There are a few disadvantages with this option which future studies should take into account:

- The process of positioning ontologies onto their applicable levels of abstraction can be time consuming.

- Since two different sets of ontologies are used for the exploration of the basic level and higher level application domain, the results obtained from the examination of the methodology's application domains cannot be attributed solely to the factor of abstraction level since there are alternative explanations derived from other differing attributes in different ontologies, such as different domain focuses, inclusion of other associative relationships.

Test on different abstraction levels of the same two ontologies

Use only two ontologies and select comparable concepts from these two ontologies on the higher level and the basic level of abstractness.

This can be done by first, determining the range of levels of abstraction that each of these two ontologies present. Second, identify common concepts in the two ontologies for each of the two abstraction levels to be examined (one higher level and ideally one closer to the basic level). This is to ensure that in each level to be examined, we have comparable concept sets between ontologies. Finally, apply the selected evaluation techniques in order to compare the quality of the two ontologies on each level of abstractness examined.

This is possibly a better method to explore the application domain of the multi-method methodology as the concerns we have with the first option will have been removed. Moreover, the main purpose of this study is to evaluate and demonstrate the application of the methodology which is to compare the quality of ontologies. This approach allows us to focus the evaluation on two ontologies, which is also more similar to the actual application of this methodology.

9.2 Recall technique - observe for both the familiarity and the expectation-violation effects

Future studies should try to observe both the familiarity effect using pure-list cued-recall design and the expectation-violation effect using mixed-list free-recall design (refer to Section 6.1). The advantage of testing for both effects in the replication study is that it verifies Hirshman's findings in more detail and therefore allows the validity of the recall technique to be better examined. The advantage of testing for both effects in the test trials is that the results obtained using one design, such as the familiarity effect, can be a cross-reference tool for researchers to better verify the results obtained using the other design (expectation-violation effect). The results can be cross referenced as these two designs should give opposite results for each of the ontological sets examined. This would therefore add more credibility to our study results.

Experimental design

Different from Hirshman's experimental design (see Appendix B), different groups of subjects should be used for the pure-list cued-recall condition than for the mixed-list free-recall condition. This will avoid the potential bias of maturation effect (such as fatigue), and pre-testing effect, since participants might learn from the first experimental trials (free recall [42]) and as such change their answers to what they think are more acceptable or more appropriate in the second experimental trials (cued recall in [42]).

The design and procedure for the mixed-list free-recall condition (observe for the expectation-violation effect) would be the same as those presented in this study (refer to Section 6.3). A different group of subjects should be used in the pure-list cued-recall study condition (observe for the familiarity effect). Half of these subjects are assigned statements with strongly related concepts (S1) and the other half is assigned statements with weakly related concepts (S2). However, the subject effect is likely to occur with the pure-list design. Since in this study, the proportions of correct S1 recall and S2 recall - measures used to evaluate the quality of ontologies - are obtained from two different groups of subjects. For this reason, future studies can also choose to use the mixed-list design for the observation of the familiarity effect, as the familiarity effect is also evident in the mixed-list cued-recall design.

The pure-list cued-recall condition would have mostly the same design and procedure as those in the mixed-list free-recall condition, except for the recall phase. After the study phase and the distraction task in the pure-list cued-recall study condition, subjects were each given a sheet with the subject term of every sentence statement presented, and then asked to complete the sentence next to the subject term given. The subject terms are chosen to be the cue because the predicate terms are the same in both stimulus lists and thus should be the recall target.

Chapter 10

Conclusion

We have developed an ontology evaluation methodology that assesses the cognitive quality of ontologies - whether an ontology (an explicit conceptualization) represents a formal specification of the domain conceptualization (perceived reality). Multi-method methodology consists of two established psychology techniques, SVT and recall. We demonstrated our methodology by applying these two techniques independently in order to assess and compare the quality of an ontology, SUMO, and a popular thesaurus, WordNet, which is often used as an ontology.

Even though our findings suggest that neither of the two ontologies has good formal specification of the domain conceptualization, this does not necessarily diminish the value of the two ontologies. This is because cognitive quality is one aspect of quality among many others. The quality of these two ontologies may be high in other aspects, such as functionality and usability [64], and the others described in Section 2.1.

The multi-method methodology is developed to be an evaluation tool, it is not immediately applicable to the construction of ontologies. Nevertheless, it signals an importance of the cognitive quality in ontology construction [25]. Also, since our methodology allows examinations of individual ontological sets, we can identify the problematic sets that need improvement. We believe that ontology construction and evaluation should be done concurrently, much like that in software design and knowledge engineering, where building and testing are done concurrently.

This exploratory study contributes to both researchers and practitioners by presenting a new and important aspect of ontology quality, developing a methodology for its assessment and showing its application. For researchers,

it is an initial study which demonstrates the need for further research and improvements to the methodology. For practitioners, it provides a template of ontology evaluation procedure to evaluate ontologies on this important quality.

Bibliography

[1] J. R. Anderson and G. H. Bower. Recognition and retrieval processes in free recall. *Psychological Review*, 79:97–123, 1972.

[2] J. R. Anderson and G. H. Bower. *Human associative memory*. Winston, Washington, DC, 1973.

[3] John R. Anderson. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA, 1983.

[4] John Robert Anderson. *Cognitive Psychology and its implications*. W.H. Freeman and Company, New York, NY., fourth edition, 1995.

[5] S.J. Antos. Processing facilitation in a lexical decision task. *Journal of Experimental Psychology: Human Perception and Performance*, 5:527–545, 1979.

[6] D. A. Balota. Automatic semantic activation and episodic memory encoding. *Journal of Verbal Learning and Verbal Behaviour*, 22:88–104, 1983.

[7] D. A. Balota and R. E. Lorch. Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and cognitions*, 12(3):336–345, 1986.

[8] P. E. Barres and P. N. Johnson-Laird. Why is it hard to imagine what is false? In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates*, page 859, 1997.

[9] W. F. Battig and W. E. Montague. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 80:1–46, 1969.

[10] H. K. Beller. Priming: Effects of advance information on matching. *Journal of Experimental Psychology*, 87:176–182, 1971.

[11] Brandon Bennett. Space, time, matter and things. In *Proceedings of the 2001 International Conference on Formal Ontologies in Information Systems FOIS, Ogunquit, Maine*, pages 105–116, 2001.

[12] E. A. Bilodea and D. C. Howell. Government Printing Office, Washington, D.C., 1953.

[13] G. L. Bradshaw and J. R. Anderson. High-speed scanning in human memory. *Science*, 153:652–654, 1966.

[14] G. L. Bradshaw and J. R. Anderson. Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, 21:165–174, 1982.

[15] Brewster. C., Alani. H., Dasmahapatra. S., and Wilks. Y. Data-driven ontology evaluation. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2004.

[16] S. Carey. *Conceptual change in childhood*. Massachusetts Institute of Technology Press, Cambridge, MA, 1985.

[17] A.M. Collins and E. F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

[18] A.M. Collins and M.R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–248, 1969.

[19] A.M. Collins and M.R. Quillian. Facilitating retrieval from semantic memory: The effect of repeating part of an inference. *Acta Psychologica*, 33:304–314, 1970.

[20] R. M. Colomb. Quality of ontologies in interoperating information systems. Technical report, National Research Council Institute of Biomedical Engineering, Corso Stati Uniti, November 2002.

[21] T. D. Cook, D. T. Campbell, and L. Peracchio. Quasi experimentation. In M. D. Dunnette and M. H Leaetta, editors, *Handbook of Industrial and Organisational Psychology*, pages 491–516. Consulting Psychologists Press, Palo Alto, CA, 1990.

[22] F. I. M. Craik and E. Tulving. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104:268–294, 1975.

[23] J Deese. Influence of inter-item associative strength upon immediate free recall. *Psycholgical Reports*, 5:305–312, 1959.

[24] Ian G. Dobbins, Neal E. A. Kroll, Andrew P. Yonelinas, and Qiang Liu. Distinctiveness in recognition and free recall: The role of recollection in the rejection of the familiar. *Journal of Memory and Language*, 38:381–400, 1998.

[25] J. Evermann. Towards a cognitive foundation for knowledge representation. *Information Systems Journal*, 15:147–178, 2005.

[26] Joerg Evermann and Jennifer Fang. Evaluating ontologies: Towards a cognitive measure of quality. *Information Systems*, 2008.

[27] J. Fang and J. Evermann. Evaluating ontologies: Towards a cognitive measure of quality. In *2007 Eleventh International IEEE EDOC Conference Workshop*, pages 109–116, 2007.

[28] Jennifer Fang. Evaluating ontologies through human cognition. Technical report, Victoria University of Wellington, 2006.

[29] Dieter Fensel, James Hendler, Henry Lieberman, and Wolfgang Wahlster, editors. *Spinning the Semantic Web - Bringing the World Wide Web to its Full Potential*. The MIT Press, Cambridge, MA, 2003.

[30] Dieter Fensel, Frank van Harmelen, Ian Horrocks, Debora L. McGuiness, and Peter F. Patel-Schneider. OIL: an ontology Infrastructure for the semantic web. *IEEE Inteligent Systems*, 16(2):38–45, March 2001.

[31] A. Gangemi, C. Catenacci, M. Ciaramita, and J. Lehmann. A theoretical framework for ontology evaluation and validation. In *Proceedings of Semantic Web Applications and Perspectives (SWAP) – 2nd Italian Semantic Web Workshop, Torento, Italy*, 2005.

[32] S. A. Gelman. The development of induction within natural kind and artifact categories. *Cognitive Psychology*, 20:65–95, 1988.

[33] S. A. Gelman and E. M. Markman. Categories and induction in young children. *Cognition*, 23:183–209, 1986.

[34] Andrew Gemino and Yair Wand. A framework for empirical evaluation of conceptual modeling techniques. *Requirements Engineering*, 9:248–260, 2004.

[35] M. C. Genesereth. Knowledge interchange format. In J. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference (KR'91)*, San Francisco, California, 1991. Morgan Kaufmann Publishers.

[36] G. Gillund and R. Shiff'rin. A retrieval model for both recognition and recall. *Psychological Review*, 91:l–67, 1984.

[37] E. Goldstein, editor. *Cognitive psychology: Connecting mind, research, and everyday experience*. Thomson Wadsworth, CA, USA, 2005.

[38] Thomas R. Gruber. A translation approach to portable ontology specification. *Knowledge Acquisition*, 5(2):199–220, 1993.

[39] Michael Gruninger and Jintae Lee. Ontology applications and design. *Communications of the ACM*, 45(2):39–41, February 2002.

[40] Nicola Guarino and Christopher Welty. Evaluating ontological decisions with OntoClean. *Communications of the ACM*, 45(2):61–65, February 2002.

[41] J. E. Hall. Associative strength and word frequency as related to stages of paired-associate learning. *Canadian Journal of Psychology*, 26:252–258, 1972.

[42] E. Hirshman. The expectation-violation effect: Paradoxical effects of semantic relatedness. *Journal of Memory and Language*, 27:40–58, 1988.

[43] E. Hirshman and R. A. Bjork. The generation effect: Support for a two factor theory. *Journal of Experimental Psychology: Learning*, 1988.

[44] J. M. Keenan. Memcog: Semantic memory. page 1, 2001.

[45] Henry Kim. Predicting how ontologies for the semantic web will evolve. *Communications of the ACM*, 45(2):48–54, February 2002.

[46] John Krogstie, Odd Ivar Lindland, and Guttorm Sindre. Towards a deeper understanding of quality in requirements engineering. In *Proceedings of the CAISE'95 Conference*, pages 82–95, 1995.

[47] George Lakoff. *Women, Fire, and Dangerous Things – What Categories reveal about the Mind*. The University of Chicago Press, Chicagor, IL, 1987.

[48] Odd Ivar Lindland, Guttorm Sindre, and Arne Solvberg. Understanding quality in conceptual modeling. *IEEE Software*, 11(2):42–49, March 1994.

[49] E. F. Loftus. Activation of semantic memory. *American Journal of Psychology*, 86:311–337, 1973.

[50] Adolfo Lozano-Tello and Asuncion Gomez-Perez. ONTOMETRIC: A method to choose the appropriate ontology. *Journal of Database Management*, 15(2):1–18, April 2004.

[51] G. Mandler and J. Huttenlocher. The relationship between associative frequency, associative ability and paired associate learning. *American Journal of Psychology*, 59:424–428, 1956.

[52] J. L. McClelland and D. E. Rumelhart. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, 88:375–407, 1981.

[53] P. D. McCormack. Recognition memory: How complex a retrieval system? *Canadian Journal of Psychology*, 26:19–41, 1972.

[54] J. C. McCullers. Effects of associative strength, grade level and interpair interval in verbal paired-associate learning. *Child Development*, 32:773–778, 1961.

[55] J. A. McGeoch. The influence of associative value upon difficulty of nonsense syllables. *Journal of Genetic Psychology*, 37:421–426, 1930.

[56] Caroline Morin, Marie Poirier, Claudette Fortin, and Charles Hulme. Word frequency and the mixed-list paradox in immediate and delayed serial recall. *Psychonomic Bulletin & Review*, 13:724–730, 2006.

[57] G. L. Murphy and D. L. Medin. The role of theories in conceptual coherence. *Psychological Review*, 92:289–316, 1985.

[58] J. H. Neely. Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited capacity attention. *Journal of Experimental Psychology: General*, 106:226–254, 1977.

[59] Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontologies in Information Systems FOIS, Ogunquit, Maine 2001*, pages 2–9, 2001.

[60] K. Noda and A Tokosumi. Development of an evaluation ontology. In *Symposium on Large-Scale Knowledge Resources, Tokyo, Japan*, March 2005.

[61] N. F. Noy and D. L. McGuinness. Ontology development 101: A guide to creating your first ontology. 2001.

[62] D. Palermo and J. Jenkins. University of Minnesota Press, Minneapolis, 1953.

[63] Jeffrey Parsons and Yair Wand. Choosing classes in conceptual modeling. *Communications of the ACM*, 40(6):63–69, June 1997.

[64] R. Porzel and R. Malaka. A task-based approach for ontology evaluation. In *Proceedings of the ECAI Workshop on Ontology Learning and Population, Valencia, Spain*, 2004.

[65] M. I. Posner and C. R. R. Snyder. Attention and cognitive control. In R. L. Solso, editor, *Information processing and cognition: The Loyola Symposium*. Erlbaum, Hillsdale, NJ, 1975.

[66] L. Postman, P. A. Adams, and L. W. Phillips. Studies in intentional learning II: the effects of associative value and method of testing. *Journal of Experimental Psychology*, 49:1–10, 1955.

[67] M.R. Quillian. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–430, 1967.

[68] M.R. Quillian. The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM*, 12:459–476, 1969.

[69] L. Rips. *Similarity, typicality, and categorisation*. Cambridge University Press, New York, 1989.

[70] L. J. Rips, E. J. Shoben, and E. E. Smith. Semantic distance and the verification of semantic relationships. *Journal of Verbal Learning and Verbal Behavior*, 12:1–12, 1973.

[71] H. K. Rodewald. Symmetry of the paired associate bond. *Psychological Reports*, 25:3–6, 1969.

[72] E. Rosch, C. Simpson, and R. S. Miller. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2:491–502, 1976.

[73] E. H. Rosch, C.B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[74] E.H. Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233, 1975.

[75] E.H. Rosch. The nature of mental codes for color categories. *Journal of Experimental Psychology: Human Perception and Performance*, 1:303–322, 1975.

[76] E.H. Rosch and C.B. Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.

[77] Eleanor Rosch. On the internal structure of perceptual and semantic categories. In T.E. Moore, editor, *Cognitive development and the acquisition of language*. Academic Press, New York, NY, 1973.

[78] Eleanor Rosch. Principles of categorization. In E. Rosch and B. Lloyd, editors, *Cognition and Categorization*. Erlbaum, Hillsdale, NJ, 1978.

[79] M. Rosemann and B. Wyssusek. Enhancing the expressiveness of the Bunge-Wand-Weber ontology. In *Proceedings of the Eleventh Americas Conference on Information Systems, Omaha, NE, USA*, 2005.

[80] R. Rosenthal. *Experimenter Effects in Behavioral Research*. Meredith Publishing Company, New York, USA, 1966.

[81] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 1995.

[82] A. I. Schulman. Memory for words recently classified. *Memory & Cognition*, 2:666–672, 1974.

[83] Reinhard Schütte and Thomas Rotthowe. The guidelines of modeling - an approach to enhance the quality in information models. In *Proceedings of 1998 Conference on Conceptual Modelling ER'98*, pages 240–254, 1998.

[84] Maria D. Sera, Jaime Gathje, and Javier del Castillo Pintado. Language and ontological knowledge: The contrast between objects and events made by Spanish and English speakers. *Journal of Memory and Language*, 41(3):303–326, October 1999.

[85] E. E. Smith. Effects of familiarity on stimulus recognition and categorization. *Journal of Experimental Psychology*, 74:324–332, 1967.

[86] Peter Smith. *An Introduction to Knowledge Engineering*. Thomson Computer Press, London, UK, 1996.

[87] E. N. Sokolov. *The modeling properties of the nervous system*. Basic Books, New York, 1968.

[88] P Spyns. Evalexon: Assessing triples mined from texts. Technical report, STAR Lab, Brussel, 2005.

[89] Steffen Staab and Rudi Studer, editors. *Handbook on Ontologies*. Springer Verlag, Berlin et al., 2004.

[90] Xiaomeng Su and Lars Ilebrekke. A comparative study of ontology languages and tools. In *Proceedings of the 10th International Conference on Advances in Information Systems Engineering (CAiSE), Toronto, ON*, 2002.

[91] S. P. Tipper. Selection for action: The role of inhibitory mechanisms. *Current Directions in Psychological Science*, 1:105–109, 1992.

[92] E. Tulving and T. Y. Areiunkle. Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, 1:321–334, 1963.

[93] T. W. Turnage. Pre-experimental associative probability as a determinant of retention. *Journal of Verbal Learning and Verbal Behavior*, 2:352–360, 1963.

[94] B. J. Underwood and R. W. Schulz. Response dominance and the rate of learning paired associates. *Journal of General Psychology*, 62:84–95, 1960.

[95] Y. Wand and R. Weber. On the ontological expressiveness of information systems analysis and design grammars. *Journal of Information Systems*, (3):217–237, 1993.

[96] 0. S. Watkins and M. J. Watkins. Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, 104:442–445, 1975.

[97] M. J. Watkins, D. C. Lecompte, and K. Kim. Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 26:239–245, 2000.

[98] J. W. Whitlow and A. R. Wagner. *Memory and habituation*. Academic Press, New York, 1984.

[99] D. A. Wicklund, D. S. Palermo, and J. J. Jenkins. The effects of associative strength and response hierarchy on paired associate learning. *Journal of Verbal Learning and Verbal Behavior*, 3:413–420, 1964.

[100] R. Zmud, M. Olson, and R. Hauser. Field experimentation in mis research. In I. Benbasat, editor, *The Information Systems Research Challenge: Experimental Research Methods*, pages 97–112. Harvard Business School, Boston, MA, 1989.

Appendix A

SVT technique - Study by Rips, Shoben, and Smith

A.1 Method

A.1.1 Material

142 sentences were constructed in the form of "An S is a P", almost half of the sentences were true statements and the rest were false. The predicate nouns (P) were the category names (e.g. bird, mammal, car, animal), and the subject nouns (S) were instances of these categories (e.g. bluejay, bear, Cadillac). A Level 1 sentence (S2) has animal or vehicle as predicate noun whereas a Level 2 sentence (S1) has bird, mammal or car as its predicate noun. It was predicted that subset effect should be observed for each instance/set - the Level 1 RT should be greater than the Level 2 RT. See Table A.1 for the stimulus list for the construction of true sentences. False sentences are constructed by matching a subject noun with the predicate noun of the incorrect category. The predicate nouns appeared equally often in true and false statements. 12 practice trials had all subject and predicate nouns different from that in the test trials.

A.1.2 Procedure

12 subjects were instructed to decide and press one button if the sentence was generally true and the other button if it was generally false. They were told to do so as accurately and as quickly as possible. In each trial, a warning light

Subject noun	Level2 (Predicate)	Level1 (Predicate)	Difference (Level)
	Bird	Animal	
Buejay	1364	1455	91
Cardinal	1383	1584	201
Chicken	1362	1463	101
Duck	1280	1339	59
Eagle	1309	1360	51
Goose	1350	1417	67
Hawk	1239	1726	487
Parakeet	1210	1398	188
Parrot	1284	1342	58
Pigeon	1214	1481	267
Robin	1346	1424	89
Sparrow	1339	1477	138
Category mean	*1307*	*1456*	*149*
	Mammal	Animal	
Bear	1318	1258	-60
Cat	1355	1278	-77
Cow	1258	1322	64
Deer	1342	1305	-37
Dog	1466	1279	-187
Goat	1442	1315	-127
Horse	1356	1296	-60
Lion	1318	1244	-74
Mouse	1440	1288	-152
Pig	1476	1268	-208
Rabbit	1418	1290	-128
Sheep	1266	1250	-16
Category mean	*1371*	*1283*	*-88*
	Car	Vehicle	
Cadillac	1303	1490	187
Continental	1362	1553	191
Corvette	1292	1446	154
Dodge	1208	1278	70
Edsel	1232	1445	213
Model T	1215	1522	307
Porsche	1348	1395	47
Rolls	1160	1286	126
Studebaker	1318	1405	87
Toyota	1320	1136	-184
Triumph	1256	1227	-29
Volkswagen	1202	1380	178
Category mean	*1268*	*1380*	*112*

Table A.1: Rips et al.'s stimulus list [70]

would signal 2 seconds before the sentence appears in an exposure device. The timer would start as the sentence appeared and would stop as the subject responded. This response would terminate the sentence presentation. The intertrial interval was 7 seconds, and the subject was informed if he had made an error.

A.2 Results

In this study, error rates were on average 4 percent. The RTs for each Subject's True correct responses was the main interest of analysis. The figures on the second and third columns of Table A.1 are the mean RTs of the associated sentence statement. For example, the sentence "A bluejay is a bird" takes on average 1364 mille-seconds (msec) for a subject to verify. The final

column presents the result of the subset effect obtained for each instance. The findings show that subset effect was obtained for all but two instances in the birds and cars categories. However, most of the mammal instances showed the opposite effect. The difference due to levels was significant overall and within each category. The opposite subset effect found amongst mammal instances was an indication that memory structure does not necessarily mirror logical structure.

Appendix B

Recall technique - Study by Hirshman

B.1 Method

B.1.1 Material

A 2 x 2 within-subject factorial design was used with Associative Strength (strongly related pairs vs. weakly related pairs) and Type of Test (free recall vs. cued recall) as within-subject factors.

Two alternative lists each comprised of 19 word pairs served as the to-be-remembered materials. A third list comprised of 6 pairs served as a practice list. The 19 response words in each list were the same words presented in the same order, but if a given response word was a strong associate on one list, it was a weak associate on the other list. Strength of association was determined from published norms [12, 62]. These materials are presented in Table B.1. Long-short is an example of a strongly related pair, while Quick-Short is an example of a weakly related pair. In one list the odd-numbered response words (in terms of list position) are strong associates, and the even-numbered words are weak associates. In the other list, the odd-numbered response words are weak associates and the even-numbered words are strong associates. Each of the two alternative lists was presented to half the subjects. Across subjects, therefore, assignment of strong associates and weak associates to list position was counterbalanced.

The practice list consisted of three strongly related pairs and three weakly related pairs presented in alternation. For each list, slides were constructed

Weakly related stimuli	Strongly related stimuli	Responses
Colour	Grass	Green
Patch	Cabbage	Lettuce
Couch	Bed	Sleep
Swift	Fast	Slow
Glue	Table	Chair
Brave	Strong	Weak
Bright	Dumb	Stupid
Leaf	Stem	Flower
Ruler	King	Queen
Quick	Long	Short
Blade	Scissors	Cut
Rabbit	Lamb	Sheep
Pray	Want	Need
Mate	Man	Woman
Glass	Soft	Hard
Roll	Carpet	Rug
Head	Mind	Think
Worm	Insect	Bug
Room	House	Home

Table B.1: Hirshman's word pair stimuli [42]

with one word pair per slide. A cued-recall test was constructed for each list by randomly reordering the stimulus terms of the word pairs in the list. Subjects wrote the response word in a space provided to the right of the stimulus term.

B.1.2 Procedure

Subjects were shown 19 pairs of words on a slide projector, and they were instructed to write each pair on a sheet in their note pads during the 10 seconds each pair was shown. When a "turn" command was spoken, subjects should turn the page in their note pads and write the new pair on the next page.

Subjects were told that there would be a memory test, but the nature of the test was not specified. A practice list of six items was presented prior to the presentation of the critical list. Subjects then received a one-minute

Type of test	Strongly related pairs	Weakly related pairs
Free recall	.23	.34
Cued recall	.93	.73

Table B.2: Hirshman's results: Proportion of response words recalled as a function of type of test and associative strength

cued-recall test on this list. The critical list was then presented and the study phase ran as described above. Subjects then received a word-search puzzle for 5 min. Following the word-search task, they received a blank sheet and were asked to recall the response word members of the pairs of words they had studied. These sheets were collected after 3 min and the cued-recall test was given. On this test, the stimulus words were presented on a sheet and the subjects were asked to write the response words next to the stimulus words which had accompanied them on the study list. Subjects were given 3 minutes for this test.

B.2 Results

Table B.2 shows the mean proportions of correct responses in free recall and cued recall for responses from strongly related and weakly related pairs. In computing the proportions, the first two and the last one of the studied pairs were omitted from the analysis (considered primacy and recency items).

The finding of primary interest is that while recall of responses from strongly related pairs was significantly superior to recall of responses from weakly related pairs in cued recall (.93 vs. .73), the recall of responses from weakly related pairs was significantly superior to recall of responses from strongly related pairs in free recall (.34 vs. .23). This interaction between Associative Strength and Type of Test was also significant.

In a follow up experiment (Experiment 6 [42]), subjects were asked to recall the word pairs instead of the response terms, and it was found that the expectation-violation effect occurred - weakly related word pairs were better recalled than strongly related word pairs. This finding suggests that expectation-violation effect occurs also in the condition of recalling word pairs in which both subject term and response term are included for recall. Thus, this experimental design can be applied to the recalling of sentence statements where both subject and response terms are included.

Appendix C

Subject design of the methodology

Between-subject design This study chose to use this subject design.

Group 1 subjects do Experiment 1 which consists of the replication study of the recall technique; practice/replication study of the SVT technique; test trials of the SVT technique.
Group 2 subjects do Experiment 2 which is the test trials of the recall technique.

Advantage: each subject would be presented with the test stimuli once only, this avoids the possible priming effect; the experimental design of the test trials of the recall technique is not restricted by the design of Experiment 1 (it can be conducted in groups).
Disadvantage: possible bias due to subject effect (different subjects used in the test trials of the two techniques); more subjects needed as we have two separate experiments to run; need two $50 prizes, one for each subject group.

Within-subject design One group of subjects does both the SVT and the recall experiments (replication and test trials of the SVT technique, and replication and test trials of the recall technique).
Advantage: no subject effect; less subjects needed; need only one $50 prize.
Disadvantage: possible maturation effect, such as fatigue effects caused by the duration of the experiment and the number of tasks each subject needs to do; possible pre-test effect, as subjects might work out the

nature, measure, and purpose of the recall test trials after the recall replication study; possible priming effect due to the repetition of stimuli in the test trials of the two techniques.

Appendix D

FLXLab experimental program

The experimental program for the SVT technique was created by following these steps:

1. Install the FLXLab software, which can be downloaded from the open source SourceForge.

2. Open the demos application from the start menu, as shown in D.1.

3. The demos opens up a folder with various demo experiments provided. The 2 practice trial programs and 4 test trial programs used in this study were modified from the lexical_decision_demo. This is done by modifying the scripts of the demo. The scripts for the SVT experimental program are shown below.

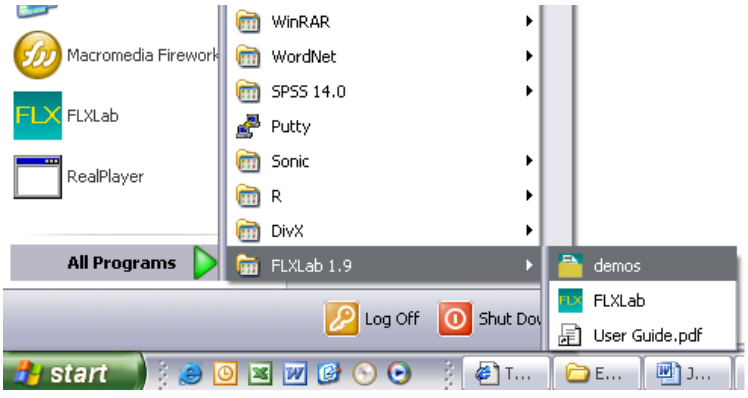

Figure D.1: Demo application in the start menu

```
# FLXLab v1.9 - A program for running psychology experiments.
# Copyright (C) 2006 Todd R. Haskell (thaskell@usc.edu)

# Get the subject id from the user
EditDialog subject_id "Enter subject id:" subject01
# Use the subject id to generate the name of the data file
JoinStrings data_file "data_files $path_separator $subject_id .txt"
# Set the name of the data file
UseDataFile $data_file

# Create labels for the three columns in the stimulus file
StimulusList stimulus_list stimulus_list.txt
LabelListColumn 1 item_number
LabelListColumn 2 target_item
LabelListColumn 3 item_condition
LabelListColumn 4 item_type

# Use 36 point URWNimbus throughout the experiment
Font URWNimbus 30

UseMilliseconds

ClearScreenEvent clear_screen

# This creates an event to display the instructions
LoadTextFromFile instructions_text instructions.txt
TextBoxEvent instructions $instructions_text

WaitEvent wait_for_key "until key any"

TextEvent mask +
WaitForRefresh
ResetEventTime

TextObject target_text $target_item
DisplayEvent target
Colour white
AddObject target_text
# Reaction times will be relative to the onset of the target
ResetDataTime

DataEvent record_response
DataColumn $item_number
DataColumn $item_condition
DataColumn $item_type
```

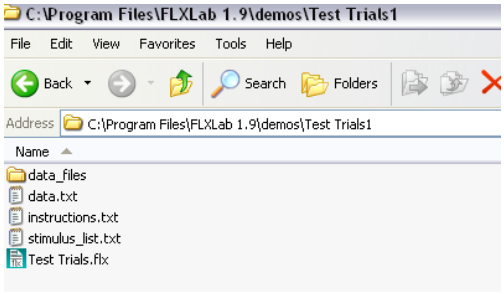

Figure D.2: Sample demo folder

```
DataColumn $key
DataColumn $time

TrialEvent trial "until event record_response"
AddEvent mask
AddEvent target "when time 2000"
AddEvent record_response "when key any"

BlockEvent mainblock "until list end"
AddEvent trial

ExperimentEvent lexical_decision
AddEvent clear_screen
AddEvent instructions
AddEvent wait_for_key
AddEvent mainblock

Start lexical_decision
```

The demo folder of test trial 1 (after modification) is shown in Figure D.2. The script of the demo file can be accessed by first starting up the .flx demo file (Figure D.3 shows that FLXLab interface), and click on the "Edit" button. The same script applies to all practice and test trial programs.

4. The presentation sentences are created by modifying the contents in stimulus_list.txt. Texts below show the stimulus list for test trial Program 1 as an example. Column 1 is the trial numbers. Column 2 is the test sentences in the order of presentation. Column 3 represents

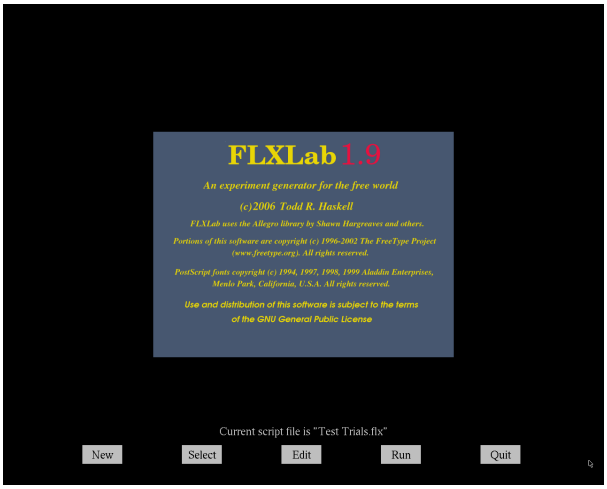

Figure D.3: FLXLab interface

the identification of the sentences. For example, the first presented sentence is 18b (a=S2, and b=S1), which means the sentence is the S1 of the 18th set (S1 of the sixth set of SUMO false sentences - by the order of 6 true SUMO sets and 6 true WordNet sets (Table 4.1), 6 false SUMO sets and 6 false WordNet sets (Table 5.1). Column 4 represents the truth of the sentences presented, and the correct response keys for the sentences.

```
1 "every City is a Text" 18b false-v
2 "every Virus is a Motion" 15b false-v
3 "every Nation is an Agent" 01b true-b
4 "every Socialism is an Organisation" 19a false-v
5 "every Sport is a Game" 02b true-b
6 "every Nation is an Object" 01a true-b
7 "every Election is a Move" 21a false-v
8 "every Book is a Text" 06b true-b
9 "every Noise is a Perception" 10a true-b
10 "every Necklace is a Work" 24a false-v
11 "every Cooking is a Creation" 05b true-b
12 "every Vitamin is a Creation" 17b false-F
13 "every Noise is a Sensation" 10b true-b
14 "every Institution is a Sensation" 22b false-v
```

```
15 "every Fish is a Contest" 14a false-v
16 "every Reasoning is an Object" 13a false-v
17 "every Magazine is a Work" 12a true-b
18 "every Sport is a Contest" 02a true-b
19 "every Virus is a Process" 15a false-v
20 "every Escape is a Diversion" 08b true-b
21 "every Socialism is an Unit" 19b false-v
22 "every Cooking is a Process" 05a true-b
23 "every book is an artifact" 06a true-b
24 "every Background is a Diversion" 20b false-v
25 "every City is an artifact" 18a false-v
26 "every Music is a Radiation" 04b true-b
27 "every Fish is a Game"   14b false-v
28 "every Walk is a Motion" 03b true-b
29 "every Vitamin is a Process" 17a false-v
30 "every Date is a Change" 23b false-v
31 "every Government is a Motion" 16a false-v
32 "every Necklace is a Publication" 24b false-v
33 "every Institution is a Perception" 22a false-v
34 "every Music is a Motion" 04a true-b
35 "every Walk is a Process" 03a true-b
36 "every Escape is an Activity" 08a true-b
37 "every Date is an Act" 23a false-v
38 "every Reversal is a Change" 11b true-b
39 "every Background is an Activity" 20a false-v
40 "every Union is an Organisation" 07a true-b
41 "every Government is a Radiation" 16b false-v
42 "every Step is a Move" 09b true-b
43 "every Reasoning is an Agent" 13b false-v
44 "every Reversal is an Act" 11a true-b
45 "every Union is an Unit" 07b true-b
46 "every Step is a Change" 09a true-b
47 "every Election is a Change" 21b false-v
48 "every Magazine is a Publication" 12b true-b
```

5. After briefing the subject, the experimenter clicks on the Run button on the FLXLab interface, and enters the subject's id in the subject id popup (see Figure D.4) for the participant. The subject id number is

assigned by the participation order, for example, the first participant is subject01; the second participant is subject02 and so on. After entering the subject id, an instruction page, as shown in Appendix M, appears on the screen (Instruction for the experimental procedures is typed in instructions.txt). Appendix M shows the instructions of the practice trial Program 1 and test trial Program 1 and 3 which has the key B for true responses and key V for false. Practice trial Program 2 and test trial Program 2 and 4 have the same instruction page except for the reverse key response assignment. To ensure readability, the font 30 point Urwnimbus is used throughout the experiment (both instruction and sentence presentation, in both practice trials and test trials). Following the instruction are trials of sentence verifications (for an example of a sentence presentation, see Figure D.5).

6. Each subject's data output (see texts below), is saved to the data_files folder as an individual document. The name of the document is the subject id specified when beginning the trial programs, as shown in Figure D.4.

```
1 18b false-v v 3231
2 15b false-v b 5079
3 01b true-b v 2868
4 19a false-v b 4162
5 02b true-b v 2036
6 01a true-b v 4627
7 21a false-v b 1954
8 06b true-b b 1526
9 10a true-b v 5287
10 24a false-v b 3525
11 05b true-b b 1952
12 17b false-F v 2026
13 10b true-b b 2897
14 22b false-v v 2853
15 14a false-v v 2258
16 13a false-v v 2841
17 12a true-b b 1623
18 02a true-b v 2183
19 15a false-v b 2177
```

```
20 08b true-b v 2679
21 19b false-v b 4668
22 05a true-b b 1531
23 06a true-b v 2789
24 20b false-v v 3006
25 18a false-v v 1822
26 04b true-b v 5372
27 14b false-v v 1631
28 03b true-b b 1612
29 17a false-v v 2373
30 23b false-v b 3765
31 16a false-v b 3960
32 24b false-v v 2427
33 22a false-v v 3298
34 04a true-b v 3990
35 03a true-b v 2790
36 08a true-b b 1527
37 23a false-v b 1570
38 11b true-b b 3507
39 20a false-v v 1973
40 07a true-b b 1648
41 16b false-v v 1429
42 09b true-b b 1451
43 13b false-v v 2593
44 11a true-b b 2283
45 07b true-b b 3006
46 09a true-b b 1773
47 21b false-v b 2166
48 12b true-b b 1760
```

The experimenter remained by the subject's side observing them during the practice trials to insure that the subject fully understood the instructions.

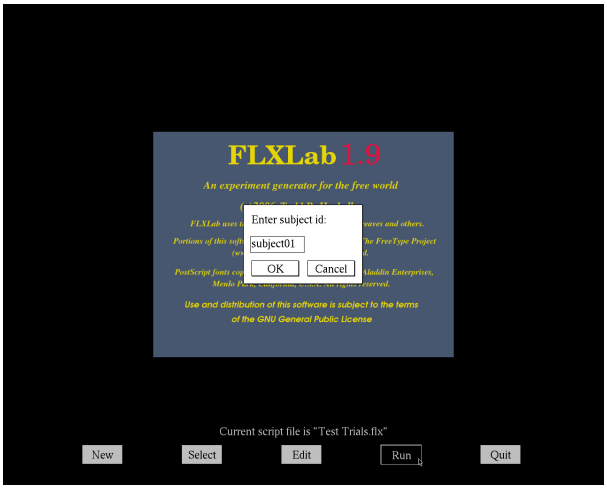

Figure D.4: FLXLab interface of the test-trial program when asked to enter subject id

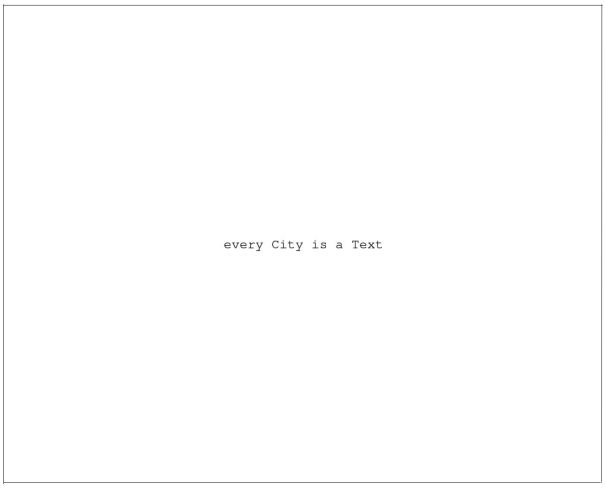

Figure D.5: FLXLab sentence presentation

Appendix E

Recruitment poster

A CHANCE TO WIN $50!!!

I am a MCA- Information Systems student, recruiting participants for my research project on evaluating category structures. This research project is designed to evaluate the quality of two ontologies (abstract level category structures) by assessing people's perception about the representation of the world. The total duration of the experiment will take no longer than 30 minutes (including briefing and de-briefing). A full information sheet will be provided to prospective participants. A $50 prize will be awarded to the participant who has the best performance in the experimental trials.

To participate-

- English must be your first language
- You must be a University student

If you are interested in participating in this study, or if you have any questions regarding this –
Contact me at xxxxxxxxxx@vuw.ac.nz.

Figure E.1: Sample recruitment poster

Appendix F

Information sheet for Experiment 1

EVALUATING ONTOLOGIES THROUGH HUMAN COGNITION

INFORMATION SHEET

This research is a Masters Thesis project. It is designed to evaluate the quality of two ontologies (abstract level category structures) by assessing people's perception about the representation of the world. The following experiment tests you on verifying category memberships. There are three independent parts to this experiment. The first part of the experiment is a simple recollection task. The next two parts will be conducted on the computer and require you to interact with a simple testing program. The experimental session should take no more than 20 minutes. The data you provide will be added to those of other participants.

Before you take part in this study, please feel free to ask any questions that you may have. Your participation is completely voluntary, and you may withdraw from the study at any time prior to the completion of the experiment without having to give reasons. Fifty dollars will be awarded to the participant who performs the best in the experimental trials. Details of how the winner will be picked and how the person will be awarded will be explained once the experiment is finished. The data set will be saved in a password-protected file for one year following the submission of my research project and then deleted. Only myself and Dr. David Mason (Supervisor) will have access to the data. The data will be published in an aggregated form only. The data will be used in a Masters degree thesis, which will be deposited in the Victoria University library. Results of this study may also be published at a scholarly conference or in an academic journal.

You will be de-briefed after data collection has been completed. Your identity will be confidential to the researcher. Your completion of the experiment implies your consent to participate in this research project.

Thank you for your help and cooperation.

Jennifer Fang
School of Information Management
Victoria University of Wellington
Email: Jennifer.Fang@vuw.ac.nz

Dr. David Mason
School of Information Management
Victoria University of Wellington
Email: David.Mason@vuw.ac.nz

Figure F.1: Information sheet for Experiment 1

Appendix G

Debriefing sheet for Experiment 1

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

EVALUATING ONTOLOGIES THROUGH HUMAN COGNITION

DEBRIEFING

Thank you for completing this experiment.

Ontology is a new buzzword in the field of information systems; it is often defined as "a formal explicit specification of a shared conceptualization". The importance of having a standardized ontology has become apparent especially to developers and system integrators in areas such as electronic commerce and artificial intelligence. In this study, we develop an evaluation methodology (multi-method technique) to examine the representativeness of ontologies, adopting two evaluation techniques – sentence verification task and stimulus recall task. The evaluation technique we examine in this experiment is the sentence verification task (tested in Part 1 of the experiment). The correctness of the ontological structures is examined by comparing the time taken to verify the truth of a category membership statement with a more distant member to that of a less distant member. The numbers of correct sets in two different ontologies are then compared. The Part 2 of the experiment is a replication study of the stimulus recall task.

$50 will be awarded to one person who has the best performance in the Part 1 test trials. The correctness and the timeliness of the responses will be the two factors of evaluation. Please neatly print your email address and sign on the paper slip attached, and give the paper to me when you leave. The evaluation process should take roughly one week. You will be informed the result of your participation as soon as the result is out.

If you have any additional questions please feel free to contact me at: Jennifer.fang@vuw.ac.nz

Thank you again for your time.

Jennifer Fang
School of Information Management
Victoria University of Wellington

Figure G.1: Debriefing sheet for Experiment 1

Appendix H

Information sheet for Experiment 2

EVALUATING ONTOLOGIES THROUGH HUMAN COGNITION

INFORMATION SHEET

I am a Masters student in School of Information Management at Victoria University of Wellington. As part of this degree I am undertaking a research project leading to a thesis. The purpose of this project is to develop an evaluation technique that allows examination of the quality of two ontologies (abstract level category structures) by assessing people's perception about the representation of the world. The University requires that ethics approval be obtained for research projects involving human participants.

The experiment tests you on stimulus recall. You will be shown on a computer screen a series of statements to remember. You will also be asked to play 3 minutes of Sudoku, after which you will be asked to recall what you saw earlier. The experimental session should take no more than 15 minutes. The data you provide will be aggregated to those of other participants and no individual results will be published.

Before you take part in this study, please feel free to ask any questions that you may have. Your participation is completely voluntary, and you may withdraw from the study at any time prior to the completion of the experiment without having to give reasons. Fifty dollars will be awarded to the participant who performs the best on the recall task. Any information (including your identity) you provide will be kept confidential to myself and Dr. David Mason (Supervisor), and only we will have access to the data. The data will be published in an aggregated form only. The data set will be saved in a password-protected file for one year following the submission of my research project and then deleted. The data will be used in a Masters degree thesis, which will be deposited in the Victoria University library. Results of this study may also be published at a scholarly conference or in an academic journal.

You will be de-briefed after data collection has been completed.

Thank you for your help and cooperation.

Jennifer Fang
School of Information Management
Victoria University of Wellington
Email: Jennifer.Fang@vuw.ac.nz

Dr. David Mason
School of Information Management
Victoria University of Wellington
Email: David.Mason@vuw.ac.nz

Figure H.1: Information sheet for Experiment 2

Appendix I

Debriefing sheet for Experiment 2

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

VICTORIA
UNIVERSITY OF WELLINGTON

EVALUATING ONTOLOGIES THROUGH HUMAN COGNITION

DEBRIEFING

Thank you for completing this experiment.

Ontology is a new buzzword in the field of information systems; it is often defined as "a formal explicit specification of a shared conceptualization". The importance of having a standardized ontology has become apparent especially to developers and system integrators in areas such as electronic commerce and artificial intelligence. In this study, we develop an evaluation methodology (multi-method technique) to examine the representativeness of ontologies, adopting two evaluation techniques – semantic verification task and stimulus recall task. The evaluation technique we examine in this experiment is the stimulus recall task. The correctness of the ontological structures is examined by comparing the number of correct recalls made with a more distant member to that of a less distant member. The study is based on the assumption of the expectation-violation effect which claims that a failure to understand the relation between the items in two ontological concepts (more distant member) can improve memory performance on that statement regarding the relationship between the two concepts. Additionally, Sudoku was served as a distraction task

$50 will be awarded to one person who has the best performance (ranked the highest) in the recall task. The number of correct recalls of stimuli presented will be the factor of evaluation. If there are multiple people who have the same top score, I will place the names of those people into a hat and draw out the name of the winning participant. Please neatly print your name on the answer sheet, and give the paper to me when you leave. The evaluation process should take roughly one week. You will be informed the result of your participation as soon as the result is out.

If you have any additional questions please feel free to contact me at: Jennifer.fang@vuw.ac.nz

Thank you again for your time.

Jennifer Fang
School of Information Management
Victoria University of Wellington

Figure I.1: Debriefing sheet for Experiment 2

Appendix J

Consent form for Experiment 2

Consent form

VICTORIA UNIVERSITY OF WELLINGTON

CONSENT TO PARTICIPATION IN RESEARCH

I have been given and have understood an explanation of this research project.

I have had an opportunity to ask questions and have them answered to my satisfaction.

I understand that I may withdraw myself (or any information I have provided) from this project (before data collection and analysis is complete) without having to give reasons or without penalty of any sort.

I understand that any information I provide will be kept confidential to the researcher and the supervisor

I understand that the published results will not use my name, and that no opinions will be attributed to me in any way that will identify me. Data will be reported in form only.

The data set will be saved in a password-protected file for one year following the submission of my research project and then deleted.

I understand that the data I provide will not be used for any other purpose or released to others without my written consent.

I understand that the $50 prize will be given to the best performer.

I agree to take part in this research

Signed:

Name of participant
(Please print clearly) Date:

Figure J.1: Consent form for Experiment 2

Appendix K

Recall experiment - Distraction task

Su Doku

Instructions for Sudoku

Fill in the grid so that every row, every column, and every 3x3 box contains the digits 1 through 9 exactly once.

Game 1:

	8		9	3			5	7
	3	7					8	9
5		9	2	8		4	3	
	7	8	5	6	4	9		3
3	9			1			4	8
	2	6				7		5
7		1		2	9		6	
	4	3		5		8	7	2
8	6		4	7	3	5		1

Game 2:

	5	3	1			4		
		7						
		8						2
8			6			2	7	3
3		6						
	1		2	3	9	8		
	3				4			
	7		9	5			8	
9	8		3			6		

Figure K.1: Recall technique distraction task

Appendix L

Recall experiment - Stimulus lists

List A	List B
every Advertising is a Dissemination	every Advertising is a Communication
every Muscle is a Part	every Muscle is an Organ
every Book is a Text	every Book is an Artifact
every Step is a Change	every Step is a Move
every Nation is an Agent	every Nation is an Object
every Cooking is a Process	every Cooking is a Creation
every Reversal is a Change	every Reversal is an Act
every Union is an Organization	every Union is a Unit
every Noise is a Sensation	every Noise is a Perception
every Music is a Motion	every Music is a Radiation
every Magazine is a Publication	every Magazine is a Work
every Escape is an Activity	every Escape is a Diversion
every Walk is a Motion	every Walk is a Process
every Sport is a Contest	every Sport is a Game
every Blood is a Mixture	every Blood is a Substance
every Tourism is a Business	every Tourism is a Commerce

Table L.1: Two stimulus lists (List A & B) of the recall test trials, differed by the alternation order of strength associations

Appendix M

Experimental instructions for subjects

This appendix contains the on-screen instructions that were presented to subjects before the practice/replication and test trials of the SVT experiments, and the on-screen instructions shown to subjects before the recall experiments.

SVT instructions

Instructions for Practice/replication Runs

A series of sentences will be presented to you one at a time. Your task is to decide whether each sentence presented is true or false. Indicate your decision by pressing either a True or False button. It is important that you give only one key response to each sentence.

Press the key labelled 'T' if the sentence is TRUE. Press the key labelled 'F' if the sentence is FALSE.

A plus sign ("+") will appear at the center of the screen two seconds before each sentence appears. Please focus your attention when you see the plus sign and be prepared to verify the coming sentence.

Please put your right index finger on the key labelled 'T', and left index finger on the key labelled 'F' now.

Press any key to move on to the practice trials. The practice trials are a practice run for your understanding of the experiment. Please respond as

quickly and accurately as possible.

Instructions for test trials

The test trials are constructed the same as the practice trials. Your task is to decide whether each sentence presented is true or false. Please respond as quickly and accurately as possible. It is important that you give only one response to each sentence.

Press the key labelled 'T' if the sentence is TRUE. Press the key labelled 'F' if the sentence is FALSE.

Please focus your attention when you see the plus sign ("+") and be prepared to verify the coming sentence.

Please put your right index finger on the key labelled 'T', and left index finger on the key labelled 'F' now.

Press any key to start the test trials.

Recall experiment instructions

You will be shown 16 sentence statements one after one on the screen. Please write each sentence on a sheet in the note pad provided during the 10 seconds each sentence is shown. You will hear a sound at the end of each 10 seconds, it is an indication for you to turn the page in your note pad and start writing the new sentence on the next page.

You will then be asked to play 3 minutes of Sudoku.

There will be a memory test at the end.

Appendix N

R commands

N.1 SVT analysis

N.1.1 SVT analysis - Practice/replication trials

```
# Extract the data for SVT replication study
> data <- read.csv ("F:/Masters/Analysis/SVTReplication/SVTreplication.
  csv", header=TRUE)

# Normalize to unit SD
> z <- (data$S2_S1_TD / sd(data$S2_S1_TD, na.rm=T))

# Test to see if the data is normally distributed: not normally
distributed!!
# plot a histogram, looks somewhat peaked, so probably too large
a kurtosis to be normal
> hist(z, 25)
> qqnorm(z)
> qqline(z)
> ks.test(z, pnorm)
> shapiro.test(z)

# Load the kurtosis function
> library(e1071)
> kurtosis(z, na.rm=T)

# Transform data: the atan transform can lower the kurtosis, so
let's apply
> kurtosis(atan(1.15*z), na.rm=T)
```

```
# Test normality after transformation: no significant departerues
> hist(atan(1.15*z), 25)
> qqnorm(atan(1.15*z))
> qqline(atan(1.15*z))
> shapiro.test(atan(1.15*z))

#But, the Kolmogorov-Smirnoff test does show non-normality
> ks.test(atan(1.15*z), pnorm)

# Let's work with this data anyway
> zz <- atan(1.15*z)

# Do the ANOVA with the transformed data
> pair.fac <- factor(data$Pair)
> subj.fac <- factor(data$Subject)
> cate.fac <- factor(data$Category)
> SVTR.aov <- aov(zz ~ cate.fac + cate.fac/pair.fac + Error(subj.
  fac))
> summary(SVTR.aov)

# Boxplot S2-S1-TD (standardized and transformed data) by Category
> plot(zz ~ cate.fac[drop=TRUE], ylab="Verification Time", xlab=
  "Category")

# Boxplot S2-S1-TD (standardized and transformed data) by Sets
> plot(zz ~ pair.fac[drop=TRUE], ylab="Verification Time", xlab="Sets")

# Let's see whether each pair has S2--S1--TD that is greater than 0.
> wilcox.test(data[data$Pair=="1",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="2",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="3",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="4",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="5",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="6",]$S2_S1_TD, na.rm=T, alternative="greater")
```

N.1.2 SVT analysis - Test trials

```
# Extract the data for SVT test trial study
> data <- read.csv ("F:/Masters/Analysis/SVTTest/SVT.csv", header=TRUE)

# Normalize to unit SD
> z <- (data$S2_S1_TD / sd(data$S2_S1_TD, na.rm=T))
```

```
# Histogram (looks normally distributed but somewhat peaked)
> hist(z, 25)

# Test kurtosis: too large a kurtosis to be normal
> library(e1071)
> kurtosis(z, na.rm=T)

# Draw QQ-plot
# The curve is far from the straight line so I strongly suspect
the normality.
# Test the normality by performing Kolmogorov-Smirnov and Shapiro-
Wilk tests
> qqnorm(z)
> qqline(z)
> ks.test(z, pnorm)
> shapiro.test(z)

# Transform data: the atan transform can lower the kurtosis, so
let's apply
> kurtosis(atan(1.17*z), na.rm=T)

# Test normality after transformation: no significant departerues
> hist(atan(1.17*z), 25)
> qqnorm(atan(1.17*z))
> qqline(atan(1.17*z))
> shapiro.test(atan(1.17*z))

# But, the Kolmogorov-Smirnoff test does show non-normality
> ks.test(atan(1.17*z), pnorm)

# Let's work with this data anyway
> zz <- atan(1.17*z)

# Do the ANOVA with the transformed data
> pair.fac <- factor(data$Pair)
> ont.fac <- factor(data$Ontology)
> subj.fac <- factor(data$Subject)
> SVTatan.aov <- aov(zz ~ ont.fac + ont.fac/pair.fac + data$AVE_DIST
  + Error(subj.fac))
> summary(SVTatan.aov)

# Boxplot TD vs. ontology
> plot(zz ~ ont.fac[drop=TRUE], ylab="Verification Time", xlab="Ontology")
> dev.off()
```

```
# Boxplot TD vs. Sets
> postscript(onefile=F, file="sets.eps", width=8, height=4, title=
  "Verification time by set(1-6 = SUMO, 7-12 = WordNet)",
  horizontal=F)
> plot(data$S2_S1_TD ~ pair.fac[drop=TRUE], sort.names=FALSE, ylab=
  "Verification Time", xlab="Set (1-6 = SUMO, 7-12 = WordNet)")
> dev.off()

# Because the KS test still shows non-normality, let's do Non-parametric
test
# The Kruskal-Wallis test is a rank-based test, but can only compare
single factors

# Ontology has no effect
> kruskal.test(data$S2_S1_TD ~ ont.fac)

# Pairs have no effect
> kruskal.test(data$S2_S1_TD ~ pair.fac)

# Let's see whether each pair has a verification time that is greater than
0, as we would expect
> wilcox.test(data[data$Pair=="1",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="2",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="3",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="4",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="5",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="6",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="7",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="8",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="9",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="10",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="11",]$S2_S1_TD, na.rm=T, alternative="greater")
> wilcox.test(data[data$Pair=="12",]$S2_S1_TD, na.rm=T, alternative="greater")
```

N.2 Recall analysis

N.2.1 Recall analysis - Test trials

```
# Extract the data for Recall test trial study
> data <- read.csv ("F:/Masters/Analysis/Recall test/Recall.csv",
  header=TRUE)

# Normalize to unit SD
```

```
> z <- (data$Remembered / sd(data$Remembered, na.rm=T))

# Normality test: Not normally distributed - too large a skewness
> hist(z, 25)
> library(e1071)
> kurtosis(z, na.rm=T)
> skewness(z, na.rm=T)
> shapiro.test(z)
> ks.test(z, pnorm)

# Transform data: the log transform can lower the skewness value, so
let's apply
> skewness(log(z+0.4))

# Test normality after transformation: no significant departerues
> shapiro.test(log(z+0.4))
> ks.test(log(z+0.4), pnorm)
> qqnorm(log(z+0.4))
> qqline(log(z+0.4))

# Let's work with this data then
> zz <- log(z+0.4)

# Do the ANOVA with the transformed data
> pair.fac <- factor(data$Pair)
> ont.fac <- factor(data$Ontology)
> level.fac <- factor(data$Level)
> recall.aov <- aov(zz ~ ont.fac*level.fac + data$FREQ_ave)
> summary(recall.aov)
```