Human Reasoning and Description Logics: Applying Psychological Theory to Understand and Improve the Usability of Description Logics

Thesis

HUMAN REASONING AND DESCRIPTION LOGICS

APPLYING PSYCHOLOGICAL THEORY TO UNDERSTAND AND IMPROVE THE USABILITY OF DESCRIPTION LOGICS

A THESIS SUBMITTED TO THE OPEN UNIVERSITY (UNITED KINGDOM)
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE, TECHNOLOGY, ENGINEERING & MATHEMATICS

2017

by
Paul W Warren
Knowledge Media Institute

Contents

Abstract 1
Acknowledgements 3
1. Introduction – Description Logics and human reasoning 5
1.1 Research questions 5
1.2 Structure of the dissertation 7
2. Ontologies and Description Logics 9
2.1. Ontologies 9
2.2. Description Logics 10
2.3. Key features of OWL 15
2.4. Difficulties with DLs 19
2.5. Ontology development 20
2.6. Discussion 27
3. Theories of reasoning 29
3.1. Rule-based 31
3.2. Mental models 32
3.3. Relational complexity 39
3.4. Language – implicatures and ambiguities 43
3.5 Discussion 44
4. Methodology 49
4.1. Ethical considerations 49
4.2. The survey 50
4.3. The study questions 51
4.4. Data collection and analysis 52
4.5. Avoiding the use of prior knowledge 54
4.6. Avoiding bias due to question ordering 55
4.7. Statistical significance 56
4.8. Understanding the reasoning processes 57
4.9. Threats to validity 58
4.10. Discussion 59
5. The user experience 61
5.1. The surveys and survey respondents 61
5.2. Ontology user survey: ontologies 63
5.3. Ontology user survey: ontology editors 64
5.4. Ontology user survey: ontology languages 65
5.5. Ontology user survey: DL language features 66

5.6. Ontology user survey: ontology languages – respondents' comments 68
5.7. Ontology user survey: ontology visualization 69
5.8. Ontology patterns 71
5.9. Threats to validity 74
5.10. Conclusions 75
6. The cognitive difficulties of some common DL constructs 77
6.1. Identifying commonly used DL features 77
6.2. The study 81
6.3. Response time data 82
6.4. The componency pattern 84
6.5. The modified coparticipation pattern 87
6.6. The types of entities pattern 88
6.7. Participant feedback 91
6.8. Participant performance 91
6.9. Effect of question position 93
6.10. Effect of number of reasoning steps 94
6.11. Response time – effect of validity, response and accuracy 95
6.12. Discussion 98
7. Further investigations – controlled comparisons 103
7.1. Organisation of the study 103
7.2. Response time data 105
7.3. Functional object properties 106
7.4. Negation, disjunction and conjunction 114
7.5. Negation and restriction 123
7.6. Nested restrictions 131
7.7. Participant feedback 140
7.8. Effect of participant prior knowledge and experience 143
7.9. Effect of question position 145
7.10. Varying effects of valid and non-valid putative conclusions 148
7.11. Discussion 150
8. Study 3: Modifications to Manchester OWL Syntax 155
8.1. Organisation of the study 155
8.2. Response time data 156
8.3. Functional and inverse functional object properties 158
8.4. Boolean concept constructors 163
8.5. Negation and restrictions 169
8.6. Nested restrictions 172

8.7. Participant feedback 176
8.8. Effect of participant prior knowledge and experience 177
8.9. Effect of question position 179
8.10. Varying effect of valid and non-valid putative conclusions 182
8.11. Discussion 183
9. Discussion – theory and methodology 187
9.1. The uses of theory 187
9.2. Ambiguity in natural language – the examples of and and or 190
9.3. Reasoning in the wild and in the laboratory 193
9.4. Decidability and tractability 195
9.5. Sample size, effect size and significance 196
9.6. Experimental methodology – lessons learned and recommendations 199
10. Conclusions and future research 203
10.1. The research questions – summarising progress 203
10.2. The research questions – further work 205
10.3. Expanding the research horizon 207
10.4. Guidance for practitioners 209
10.5. Final remarks 211
Glossary of acronyms 213
References 215
Appendix A: study 1 - reasoning steps for valid questions 225
Appendix B: study 2 – reasoning steps for Boolean operator questions 231
Appendix C: effect of sample size on response time comparison 235

Abstract

Description Logics (DLs) are now the most commonly used ontology languages, in part because of the development of the Web Ontology Language (OWL) standards. Yet it is accepted that DLs are difficult to comprehend and work with, particularly for ontology users who are not computer scientists. The Manchester OWL Syntax (MOS) was developed to make DLs more accessible, by using English keywords in place of logic symbols or formal language. Nevertheless, DLs continue to present difficulties, even when represented in MOS. There has been some investigation of what features cause difficulties, specifically in the context of understanding how an entailment (i.e. an inference) follows from a justification (i.e. a minimal subset of the ontology that is sufficient for the entailment to hold), as is required when debugging an ontology. However, there has been little attempt to relate these difficulties to how people naturally reason and use language.

This dissertation draws on theories of reasoning from cognitive psychology, and also insights from the philosophy of language, to understand the difficulties experienced with DLs and to make suggestions to mitigate those difficulties. The language features investigated were those known to be commonly used, both on the basis of analyses reported in the literature and after a survey of ontology users. Two experimental studies investigated participants' ability to reason with DL statements. These studies demonstrate that insights from psychology and the philosophy of language can be used both to understand the difficulties experienced and to make proposals to mitigate those difficulties. The studies suggest that people reason using both the manipulation of syntax and the representation of semantics with mental models; both approaches can lead to errors. Particular difficulties were associated with: functional object properties; negated conjunction; the interaction of negation and the existential or universal restrictions; and nested restrictions. Proposals to mitigate these difficulties include the adoption of new language keywords; tool enhancement, e.g. to provide syntactically alternative expressions; and the introduction during training both of De Morgan's Laws for conjunction and disjunction, and their analogues for existential and universal restrictions. A third study then investigated the effectiveness of the proposed new keywords; finding that these keywords could mitigate some of the difficulties experienced.

Apart from the immediate applicability of these results to DLs, the approach taken in this dissertation could be extended widely to computer languages, including languages for interacting with databases and with Linked Data. Additionally, based on the experience of the three studies, the dissertation makes some methodological recommendations which are relevant to a range of human-computer interaction studies.

Acknowledgements

The author would like to thank his supervisors, Dr. Paul Mulholland, Dr. Trevor Collins and Prof. Enrico Motta for their advice, support and friendship. Without their help this work would not have been possible.

The author is also grateful to Prof. Marian Petre for her work in sustaining the community of computing research students, and in particular ensuring the inclusion of external students such as the author.

More generally, the author has benefited enormously from conversations with people in the CRC and KMi research community at the Open University. The author is grateful for having been part of that community.

Thanks are also due to the survey respondents and the participants in the studies. Research of this kind is entirely reliant on such volunteers. Particular thanks are extended to Dr. Gem Stapleton, of Brighton University, and Dr. John Davies, of BT Research, for facilitating study sessions at their respective institutions.

Finally, thanks are due to my wife, for her patience throughout this work.

1. Introduction – Description Logics and human reasoning

The fundamental problem of ontology is managing the complexity - and doing so in a way that people can understand.

John Sowa[1]

1.1 Research questions

The goal of this work is to understand the difficulties people experience using Description Logics (DLs[2]) to create and edit ontologies, and from that understanding, to seek to mitigate those difficulties. In particular, the work draws on insights from theories of reasoning developed by cognitive psychologists, and also from the philosophy of language.

DLs are discussed in detail in Chapter 2. In brief, they are modelling languages based on First Order Logic (FOL). Over the past few decades their decidability and computational properties have been extensively studied. Baader et al. (2003) provide a comprehensive overview of the theory of DLs, as well as describing applications in areas such as software engineering, medical informatics, digital libraries, natural language processing, and databases. A Web Ontology Language, OWL, which is based on DLs has been standardised by the W3C (2001) and is used in a number of application domains for creating ontologies. In fact, OWL exists as a number of profiles, the properties of which are well understood (Motik et al., 2012).

However, the usability of these languages, in particular their comprehensibility, has been much less studied. There have been some notable exceptions. For example, Rector et al. (2004) discuss the problems that people have working with DLs, based on their experiences of teaching OWL.

Other researchers have investigated how to create comprehensible explanations of why an ontology gives rise to a particular inference (Nguyen et al. 2012, 2013). Typically the ontology developer will be using a tool such as Protégé[3] and will, from time to time, run a reasoner. The reasoner may create an inference, or *entailment*, which the ontology developer does not expect and which may indicate an error in the ontology design. At the same time the reasoner may provide a *justification* for this entailment, defined by Horridge et al. (2011, p. 241) as "a minimal subset of an ontology that is sufficient for an entailment to hold". Horridge et al. (2011) present an intuitive model for determining the cognitive complexity of a particular justification, with a view to enabling a reasoner to choose an optimum justification[4]. Horridge et al. (2011) compared the predictions of the model with the relative difficulty experienced by people in understanding a set of entailments, concluding that the model "fared reasonably well". Nguyen et al. (2012) were interested in automating the

[1] In a post on ontolog-forum, 30th November 2013, see: http://ontolog.cim3.net/forum/ontolog-forum/2013-11/msg00117.html#nid09

[2] A glossary of acronyms is provided at the end of the dissertation, after the References and immediately before the Appendices.

[3] http://protege.stanford.edu/

[4] The model comprised twelve factors, with varying weights. For example, the number of different axiom types in the justification and the entailment together is weighted 100, i.e. the number of axiom types is multiplied by 100. As another example, each axiom type in the entailment which is not in the justification causes 50 to be added to the overall score. These two scores can be calculated from a syntactic analysis. Other factors depended on the semantics and required entailment checking.

production of a proof tree, in English, to show why a particular entailment is a consequence of the ontology. Such a proof tree consists of a sequence of deduction rules and the problem is to determine, from a number of possible proof trees, the one which is the most comprehensible. To do this requires understanding the comprehensibility of each deduction rule. Nguyen et al. (2012) investigated the comprehensibility of fifty-one typical deduction rules.

There has been other work not directly looking at comprehensibility but investigating common ontology faults, some of which may originate from problems of comprehension. For example, Kalyanpur et al. (2006) have analysed problems in existing ontologies and identified a number of underlying user errors. Poveda-Villalon et al. (2012) have compiled a catalogue of potential ontology design faults, a number of which are classified as arising from problems of 'human understanding'.

This previous work has catalogued the kinds of errors people make and highlights the need to understand why those errors occur and to mitigate them. However, none of this work has made use of findings from cognitive psychology. Other areas of computing, in particular relating to the human-computer interface, have drawn on and contributed to cognitive psychology, going back to pioneering work in the last century (Card, Moran, & Newell, 1983). Theoretical models have been used to create an understanding of human performance. An example is the use of the 'Goals, Operators, Methods, Selection' (GOMS) model to understand text editing performance (Card et al., 1983, chapter 5). Card et al. (1983) were concerned with physical issues such as screen and keyboard layout. The work of Yamauchi (2007) is more related to the inferencing studied in this dissertation. He investigated the extent to which people make predictions about individuals based on class tags and property tags; finding that the former are more important for prediction than the latter. However work to understand the difficulties of DLs, and of OWL in particular, has been limited to the anecdotal (Rector et al., 2004) and to controlled experiments to identify which constructs give difficulty (Horridge et al., 2011; Nguyen et al., 2012). This work does not explain why those constructs give difficulty. Without that understanding it is not possible to generalize beyond the particular constructs studied. Yet there has been considerable research on human reasoning which offers the possibility of providing generalizable explanations of human performance in dealing with these DL constructs. The goal of this work is to find those generalizable explanations and use them to motivate solutions to overcome the difficulties.

Thus, the research questions investigated in the current work are:

(1) In what way can the difficulties experienced in using Description Logics be understood in terms of underlying theories, e.g., theories of reasoning, already developed within the cognitive psychology community?

(2) In what way could such theories contribute to improving the usability of Description Logics?

More specifically, the work investigates:

(1A) What theories are available, and what are their relevant strengths in the context of DLs?

(1B) How can these theories explain how people understand DL statements?

How do these theories motivate:

(2A) notational extensions;

(2B) tool enhancements;

(2C) enhancements to training?

1.2 Structure of the dissertation

Chapter 2 discusses DLs, their role in describing ontologies, and some of the approaches and tools used. Chapter 3 then gives an overview of the various theories of reasoning developed by psychologists, in particular the three used in the work reported here, i.e. theories based on the use of rules, mental models, and relational complexity. This chapter also briefly discusses the concept of the *implicature*, taken from the philosophy of language. Chapter 4 then describes the methodology adopted for the current work, and the reasons for such a methodology.

Chapter 5 describes a survey into the use of ontologies. This survey served to confirm the importance of DLs, identified the most commonly used DL constructs and also identified some of the problems people experience in using DLs.

The next three chapters describe three laboratory studies. Chapter 6 describes a study specifically designed to investigate difficulties with frequently used DL constructs in the context of common ontology patterns. This chapter starts to answer research question (1) above. Chapter 7 continues to answer research question (1), by describing a study which further investigated the difficulties identified in the first study, as well as looking at some additional DL constructs. The work reported in Chapters 6 and 7 led to some proposed modifications to the syntax, in response to research question (2A). The effect of these modifications was investigated in the final study, described in Chapter 8.

Chapter 9 reviews the underlying themes, both theoretical and methodological which have run through the dissertation. Chapter 10 draws some conclusions, reviewing progress against the research questions, including a discussion of (2B) and (2C). Finally, Chapter 10 also outlines some areas for future work.

2. Ontologies and Description Logics

> *... since Aristotle, it [logic] has been unable to advance a step and, thus, to all appearance has reached its completion.*
>
> *Immanuel Kant, 'Critique of Pure Reason', Preface to second edition, 1787*

This chapter provides an overview of ontologies and Description Logics (DLs). The purpose of the chapter is to provide the background required for later chapters, and presentation is skewed towards aspects relevant to those chapters. Section 2.1 provides a general discussion of ontologies. Section 2.2 then gives an overview of DLs and section 2.3 discusses the OWL family of DLs. Section 2.4 discusses some of the known difficulties with using OWL. Section 2.5 talks about ontology development, describing the mainstream tools as well as some more experimental approaches. This section is left until after the discussion of DLs, since most ontology development today makes use of DLs. Finally, section 2.6 reviews the main themes of the chapter and makes some conclusions.

2.1. Ontologies

The classic definition of an ontology, given by Gruber (1993) is "an explicit specification of a conceptualization". This has the virtue of precision but the defect of abstraction. Its precision defines exactly what is and is not an ontology. However, its abstraction gives no insight into what an ontology looks like or how it is created. For the purposes of this dissertation a more concrete description of an ontology will be given.

Ontologies are used to describe knowledge[5]. More specifically, ontologies contain the three basic elements of any knowledge representation system: sets, members of sets, and relations between those set members. In the context of an ontology, the sets are normally referred to as **classes**, the members of the sets as **instances** or **individuals**, and the relations as **properties**. Note that properties are defined as potentially existing between the instances of two classes, such that one class provides the subject and the other the object, e.g. *John plays tennis*, where *John* is the subject and *tennis* the object of the property *plays*. Here *John* might be an instance of a class *Person* and *tennis* an instance of a class *Game*. The two classes providing the subject and object can, of course, be identical. In fact, in the context of OWL, these kind of properties are more fully referred to as **object properties**; another type of property, the **datatype property**, will be introduced in subsection 2.3.3. In this dissertation, wherever the term *property* is used unqualified, it is assumed to mean object property. In principle, relations could be defined between classes, e.g. *has_more_members_than*[6]. However, apart from the predefined subsumption relation, this is not generally done and this possibility will not be considered in this dissertation. The reason is that this takes us outside of First Order Logic (FOL) and introduces problems of decidability, see the discussion below.

Some writers reserve the term *ontology* for the classes, the properties, and the axioms relating classes and properties, e.g. the subsumption relations and domain and range statements. The term **knowledgebase** is then used for the instances and the instantiation of properties between instances. However, this distinction is not always observed.

---

[5] Antoniou and van Harmelen (2004) state this precisely and succinctly: "an ontology describes formally a domain of discourse".

[6] When property names comprise more than one word, they will be written with underscores.

An overview of knowledge representation, including a discussion of ontologies, is given in Sowa (2000).

2.2. Description Logics

2.2.1. An overview of Description Logics

This subsection provides a brief overview of DLs. For a more detailed introduction, see Krötzsch et al. (2012a). DLs are formal modelling languages used to create ontologies. Because they are formal they enable automated reasoning; in this respect they differ from, say, the Unified Modeling Language (UML). They comprise the three kinds of entities described above: individuals, classes and properties. DLs possess three defining characteristics:

1. They are based on First Order Logic (FOL)[7]. In practice, most DLs are based on decidable fragments of FOL.
2. They employ the Open World Assumption (OWA).
3. They make use of operators which are analogous to the operators of FOL, but which have a different interpretation because they are concerned with classes not propositions. In the place of conjunction, disjunction and negation, they have intersection, union and complement. In the place of the existential and universal quantifiers, they have existential and universal restrictions, which are used to define classes.

Each of these will be considered in turn.

Full FOL is undecidable. For example, in full FOL, no algorithm to determine whether one class is a subclass of another can be guaranteed to terminate. However, various judicious choices of subsets of FOL enable such guaranteed-to-terminate algorithms to be constructed.

The OWA means that the absence of a fact from the knowledgebase does not mean that the fact is false. This differs from the typical database application, where the absence of John Smith from a company employee database would imply that John Smith is not an employee. With a DL knowledgebase, if John Smith is not stated to be a company employee, all that one could infer is that the knowledgebase does not know whether he is an employee. To exclude the possibility of his being an employee, the knowledgebase would need to include the explicit fact that he is not an employee, or else contain sufficient other facts to enable the deduction that he is not an employee. Another consequence of the OWA is that one cannot assume, without other evidence, that two different names refer to different entities. For example, *John* and *Smith* may refer to the same individual. To infer that the names represent different individuals there either needs to be an explicit statement to that effect, or it needs to be possible to deduce that they are different from the statements in the knowledgebase. The OWA fits well with Web applications which need to work with incomplete data, in contrast to a corporate database which might be assumed to be complete.

The third point above is more subtle, but differentiates the use of logic and logic symbols in DLs from their normal use in mathematical logic. As an example, consider the use of the existential restriction. In the so-called **German DL** syntax, standard logical symbols are used, including ∃ for existential restriction and ⊤ for Everything, i.e. the domain of discourse. Assuming that P is a property, it is possible to write:

[7] "Description Logics (DLs) are a family of knowledge representation languages … equipped with a *logic*-based semantics which, up to some differences in notation, is actually the same semantics as that of classical first-order logic" (Baader et al. 2017).

$$\exists P.T$$

This defines an anonymous class containing all the individuals which are the subject of a property P, i.e. all the individuals *a* for which an individual *b* exists such that *a P b*. Thus, the symbol normally used for existential quantifier (∃) is being used to represent existential restriction. The symbol is here not explicitly stating the existence of an individual, it is being used to define a class. It might be construed as implicitly stating that, for each *a* in the anonymous class, then there exists *b* such that *a P b*. However, the existence of any such individual *b* is not guaranteed. It might be that the anonymous class defined above is empty, i.e. that there is no instance of the property *P* in the knowledgebase, and hence no individuals *a* and *b* acting as subject and object of the object property, *P*.

DL syntaxes created by logicians, like the German DL syntax, are generally not regarded as ideal for domain experts who are non-logicians, e.g. see Horridge et al. (2006). However, it is also likely that the use of a restriction to define classes, rather than a quantifier to make statements about individuals in classes, may present difficulties even to those familiar with logical notation.

The same difficulty applies to the use of the universal restriction. However, here there is an extra difficulty which creates problems in writing and interpreting DL statements. Consider the following statement:

$$\forall P.X$$

Here P is again a property, and X is assumed to have been declared to be a class. This statement defines the anonymous class containing all those individuals *a* such that either:

- whenever *a* is the subject of an instance of P, the object is in the class X; or
- *a* is never the subject of an instance of P.

The first option is equivalent to stating that, *for all b* such that *a P b* and *a* is in the anonymous class defined above, it is the case that *b* is in X. The use of *for all* in the last sentence corresponds to the common usage of the phrase. The second option occurs because, in mathematical logic, any statement will be valid for the members of the empty set. The existence of the possibility exemplified by the second bullet point above is sometimes referred to as the 'trivial satisfaction of the universal restriction', e.g. see Rector et al. (2004). As a consequence, in DLs, if no *b* exists such that *a P b*, then *a* will also be in the anonymous class.

To the non-logician this can be confusing, and the second option above is often overlooked. In everyday language, the phrase *for all x* has the implicature that *x* exists. That is to say, it does not logically imply that *x* exists, but conveys an assumption of existence. The word *implicature*, coined by Grice (1975) is used rather than *implication*, since the latter has a variety of meanings. An overview of implicatures in general, along with a detailed discussion of one particular sort, is given by Carston (1998).

In fact, in Aristotle's study of syllogisms *all* implied existence, see Khemlani and Johnson-Laird (2012) who cite Strawson (1952). The introduction of the more general sense of the universal quantifier, which leads to the second option above, was introduced into the mathematical treatment of logic developed in the late nineteenth century.

In this context it is worth noting the role of the domain and range statements which are implemented in some DLs. These statements are not provided for error checking in the manner of, for example, typed variables in some programming languages. They are simply additional axioms which can be used to make inferences. Thus, if the domain of a property, P, is given to be class X then the statement $a\ P\ b$ enables the inference that the individual a is a member of X. Similarly if the range of P is given as Y, then $a\ P\ b$ enables the inference that b is in Y.

In fact, domain and range statements are not necessary as they can be replaced with existential and universal quantification respectively. The statement that P has domain X is equivalent to the statement that all individuals which are the subject of an instance of P are in X, i.e.:

$$\exists P.\top \subseteq X$$

The statement that P has range Y is equivalent to the statement that all individuals which serve as objects of instances of P are in Y. This is equivalent to stating that every individual is either the subject of an instance of P such that the object is in Y, or else is not the subject of an instance of P, i.e.:

$$\forall P.\ Y \equiv \top$$

DLs also typically include the Boolean concept constructors *intersection*, *union* and *complement*. As noted by Krötzsch et al. (2012a), these are also knowns as *conjunction*, *disjunction* and *negation*, because of their relationship to conjunction, disjunction and negation in logic. Throughout this dissertation, both terminologies will be used interchangeably. However, this question of terminology will be returned to later. In the German DL notation intersection, union and complement are represented by ⊓, ⊔ and ¬ respectively.

2.2.2. The Web Ontology Language (OWL)

DLs were developed during the last few decades of the 20th century, see Nardi and Brachman (2003) for background and history. Much early research went into investigating which subsets of FOL are decidable. Considerable research was also dedicated to investigating the computational tractability of these subsets, i.e. the space and time requirements of algorithms to determine subsumption and class membership.

By the beginning of the 21st century, the success of the World Wide Web (WWW) led to calls to make the Web 'semantic'. The intention was to go beyond a web of information purely for direct human consumption to a web of data which enabled automatic reasoning across various data sources (Berners-Lee et al. 2001). The decision was taken to base an enhanced, semantic Web on DLs. An alternative knowledge representation formalism considered was frame logic (F-Logic). F-Logic is an object-oriented database logic (Kifer & Lausen, 1989). It uses the Closed World Assumption (CWA); if something cannot be shown to be true, then it is assumed to be false. As already observed, one of the reasons for adopting DLs for the Semantic Web was that the latter's OWA was believed to be more suited to the open environment of the Web where not all facts on a given topic are available.

HyperText Markup Language (HTML) was the original markup language for the WWW, and this language was purely aimed at the presentational level. The requirement for a semantic web necessitated a language with which ontologies could be constructed. The Resource Description Framework (RDF) was developed to provide a very basic

functionality, enabling the creation of subject, predicate, object triples. Since any given entity can be involved in any number of such triples, the resultant structure is a graph, which is usually serialized in XML. The RDF schema language (RDFS) extended RDF, e.g. with the introduction of the concepts of subclass and subproperty. Antoniou and van Harmelen (2004), and Allemang and Hendler (2011) provide information on these developments.

There was a perceived need to go beyond RDFS and create a DL for use on the WWW. DAML+OIL was an early attempt at such a language, see Horrocks (2002)[8]. This developed into the Web Ontology Language (OWL), created and standardised by the World Wide Web Consortium (W3C). OWL is a set of standardised DL profiles. Originally there were three species of OWL: OWL Full, OWL DL and OWL Lite, see Antoniou and van Harmelen (2004). OWL Full is not decidable. OWL DL is a subset of OWL Full chosen to be decidable, whilst OWL Lite is a subset of OWL DL chosen to improve computational tractability.

A second generation of OWL contained three profiles:

- OWL 2 QL; designed for database-like applications with a simple schema and large amounts of data.
- OWL 2 EL; designed for applications with a large number of classes, as are found in the life sciences.
- OWL 2 RL; designed for applications which employ rules, e.g. business rules.

These profiles differ in the particular constructs used and in the precise manner in which they can be used. Allemang and Hendler (2011) and Horridge et al. (2012) provide more detail on these. Motik et al. (2012) provide an overview, with links to various W3C documents. OWL is one of the technologies supporting the semantic web[9]. Specifically, OWL builds on RDF and XML. Thus an OWL ontology can be expressed as an RDF graph and serialized as XML. For a detailed description of the relation between OWL and DLs, see Baader et al. (2017, Chapter 8).

OWL exists within an ecosystem of tools, frequently employing XML as an exchange format. One of the most commonly used tools is Protégé[10], a free open-source ontology editor developed at Stanford University. Protégé makes use of a graphical user interface with panes to provide a class, individual and property-oriented view. The software has a long history; its early development is recorded by Gennari et al. (2003). It was originally developed for biomedical applications and used a frame-based, rather than DL paradigm. The frame-based approach originated with Minsky (1974) and is loosely related to the object-oriented approach in programming. One particular feature not shared with DLs is that frame-based systems allow for defaults and the creation of exceptions. However, the widespread adoption of DLs has meant that current versions of Protégé are entirely DL-based. The software incorporates a reasoner to compute class subsumption and membership relations. A wide variety of plugins are available, e.g. for ontology visualization[11]. A guide to using Protégé 4 is provided by Horridge (2011). The current version is Protégé 5. There

[8] DAML+OIL was the result of a merger between DAML-ONT, developed by the US DARPA programme, and OIL, developed chiefly in Europe. DAML is an acronym for DARPA Agent Markup Language; OIL is an acronym for Ontology Inference Layer.
[9] https://www.w3.org/standards/semanticweb/
[10] http://protege.stanford.edu/
[11] For the Protégé plugin library, see http://protegewiki.stanford.edu/wiki/Protege_Plugin_Library

is also a web-based version of Protégé created particularly for collaborative ontology development.

2.2.3. OWL syntaxes

OWL is not associated with one prescribed syntax or notation. Rather, various syntaxes have been adopted. As already noted, DLs were conceived by logicians and they used the standard mathematical symbols for the existential quantifier (∃) and universal quantifier (∀) along with the standard set-theoretic symbols for union (⊔), intersection (⊓), the formation of a complement (¬) and the subset relationship (⊒). These symbols were originally used by Protégé.

At an early stage an abstract syntax was constructed, to enable OWL ontologies to be more easily shared between people, see Patel-Schneider et al. (2004). A functional syntax was used which contained keywords such as *unionOf*, *intersectionOf*, *complementOf*. The existential and universal restrictions were achieved by the use of the keywords *restriction* and *SomeValuesFrom* or *AllValuesFrom*. As an example, the expression ∃ P.⊤ would be written:

restriction(P someValuesFrom(owl:Thing))

P would normally be replaced by a mnemonic property name, including a namespace prefix.

Somewhat later a related functional-style syntax was developed for the second generation of OWL (OWL 2), see Motik et al. (2012). Like the abstract syntax this used *some* and *all*, in compound words, to represent the existential and universal restrictions. In this the authors of these syntaxes were following the use of these words in a common English rendering of the Aristotelian syllogisms[12]. The functional-style syntax also used *union*, *intersection* and *complement* in compound words, e.g. *ObjectUnionOf*.

The abstract and functional syntaxes may have been more readable to the non-logician than the mathematical notation, although this does not appear to have been tested. However, this was at the expense of verbosity. The Manchester OWL Syntax (MOS) was created to combine readability and succinctness, see Horridge et al. (2006) for the original description of MOS, and Horridge and Patel-Schneider (2008) for some additional features. The former authors used capital letters for the keywords. However, the latter adopted lowercase and this convention is the one currently used. The current status of the syntax is described by Horridge and Patel-Schneider (2012). MOS is now used in the Protégé tool.

MOS uses the keywords *or*, *and* and *not* for union, intersection and complement. This is a departure from the previous syntaxes. Note that these are not set-theoretic terms; apart from their usage in everyday language they are used in logic, e.g. FOL and Boolean algebra. They are, however, being used in MOS as operators on classes, i.e. as replacements for the set-theoretic terms. They were presumably adopted because they were regarded as more familiar than the set-theoretic terms. However, there may be some ambiguity about the words *or* and *and*. This point will be returned to in chapters 3 and 9.

MOS uses the keywords *some* and *only* for the existential and universal restrictions. For *some* the authors were following the convention in the previous syntaxes. As already noted this has a history going back to Aristotle. The use of *only* is, however, another departure

[12] In Aristotle all syllogisms are expressed using premises and conclusions of the form: *All A are B*; *No A are B*; *Some A are B*; *Some A are not B.*

from previous practice. Note that the Aristotelian premise *all X are Y* can be rewritten *only Y are X*.

MOS also uses an infix notation. As regards the set-theoretic operators *or* and *and* this was a return to the convention used in the German DL notation. For the restrictions, the German DL syntax used a prefix notation, e.g. as in ∃ P.⊤. The abstract syntax uses a combination of prefix and infix. Consider the example given previously:

restriction(P someValuesFrom(owl:Thing))

This is prefix as regards the keyword *restriction* but infix as regards *someValuesFrom*. However, MOS drops the use of *restriction* and this results in an entirely infix notation:

P some Thing

The replacement of *all* with *only* may have been to avoid an ambiguity arising from the use of the shortened keyword. In the abstract syntax *P allValuesFrom X* makes it relatively clear that all the entities selected as an object of the property P (for subjects in this anonymous class) will be chosen from the class X. However, *P all X* might give the false impression that all members of X are to be used as objects of P. *P only X* is arguably less ambiguous.

The relative advantages of the various possible alternative keywords does not seem to have been evaluated in any systematic way. Nor has the advantages and disadvantages of prefix and infix notation been studied. Chapter 8 considers the effect of some alternative keywords and also compares human performance with prefix and infix notation in some particular situations.

2.2.4 DLs, OWL and MOS

To finish this section, it is useful to recap the distinction between DLs, OWL and MOS:

- DLs are knowledge representation languages based on FOL. In practice, they are frequently based on decidable subsets of FOL.
- OWL is a set of standardized DLs, i.e. a subset of all possible DLs. With the exception of OWL Full, they are chosen to be decidable. OWL Lite and the three species of OWL 2 are also chosen for their computational tractability.
- MOS is a particular syntax frequently used for the OWL species.

Chapters 6 to 8 investigate difficulties with DLs. These investigations concentrate on language constructs occurring in several of the OWL species, and make use of MOS. Thus, this work is concerned with difficulties arising from the specific syntax and also with difficulties arising from the underlying DL constructs. In the latter case, the difficulties may be mitigated or exacerbated by the choice of syntax.

2.3. Key features of OWL

This section describes the key features of DLs, as implemented in the various species of OWL. Not all the features here are included in all the species of OWL. The treatment here is skewed towards those features which are relevant to the work described in this dissertation. Throughout the remainder of the dissertation the syntax used for OWL statements is a slightly modified version of MOS. Specifically, the syntax follows that in Horridge and Patel–Schneider (2012) with the exceptions that:

- The colon which follows certain keywords is omitted.

- The keyword *Types* was replaced by its singular form, *Type*[13]. In studies 2 and 3, the keyword *Characteristics* was replaced by *Characteristic*[14].
- As already noted above, the study reported in Chapter 8 experiments with certain variations in keywords and syntax.

2.3.1. Classes

The set-theoretic operators union, intersection and complement, plus the relationship subclass have already been described. OWL also permits classes to be defined as equivalent or as disjoint. Additionally, a class can be described as a disjoint union of a number of other classes.

The use of the existential and universal restrictions to define classes has already been described. As already noted, these are generically referred to as *property restrictions*. They are frequently used with the subclass and equivalence relations, e.g.:

X SubClassOf has_child some MALE[15]

X EquivalentTo has_child some MALE

The first of these statements specifies that all members of X have one or more sons, and says nothing about whether or not they have daughters. The second statement defines X to be the class of individuals who have one or more sons, again irrespective of whether they have daughters.

Note also the analogous statements using *only*:

X SubClassOf has_child only MALE

X *EquivalentTo has_child only MALE*

The first of these specifies that all members of *X* have only sons, or no children at all. The second that *X* is exactly the set of individuals who have either only sons or no children at all.

Another category of property restriction is based on cardinality, specifically maximum, minimum and exact cardinality. As an example, using MOS, we could write:

has_child exactly 4

This defines a class of individuals with exactly 4 children. Using min or max we could define class of individuals with a minimum or maximum of 4 children respectively.

Similarly, we could write:

has_child exactly 4 MALE

to define the class of individuals with exactly 4 sons. Note that this places no restriction on the number of daughters. Individuals in this class might have no daughters or very many.

[13] This was done for simplicity in the questions used for the studies, since in these questions the *Type* and keyword was only followed by one entity.
[14] In study 1 some object properties were declared to have more than one characteristic, hence the keyword *Characteristics* was used. In studies 2 and 3 this was not the case, and the keyword *Characteristic* was used.
[15] Throughout, class names are written with capital letters, or combinations of capitals, numbers and underscores. Property and individual names are written with lowercase letters.

2.3.2. Individuals / instances

Individuals can be defined as being members of a class, e.g. to define *a* as a member of class *X*:

a Type X

Two names can be defined to refer to different individuals:

a DifferentFrom b

This is important because the OWA does not permit us to assume that different names represent different individuals.

Alternatively, we can specify that two different names do refer to the same individual:

a SameAs b

2.3.3. Properties

Properties can be defined to have certain characteristics, based on the characteristics of relations studied by logicians. In the following *R* represents an arbitrary relation; *a* and *b* represent arbitrary individuals. Many of the examples are taken from Hitzler et al. (2012):

symmetric, i.e. *a R b* implies that *b R a*. Example: *has_spouse*.

asymmetric, i.e. *a R b* implies that it is not the case that *b R a*. Example: *has_child*.

Note that asymmetric is stronger than not symmetric. All asymmetric properties are not symmetric; however there are properties which are both not symmetric and not asymmetric, e.g. *has_brother*. Similarly, symmetric is stronger than not asymmetric.

reflexive, i.e. *a R a* for all a. Example: *has_same_age*

irreflexive, i.e. it is never the case that *a R a*. Example: *parent_of*

functional, i.e. *a R b* and *a R c* implies that *b* and *c* represent the same instance, or alternatively for any given property and any given subject for that property, there can only be one corresponding object. An example is *has_national_insurance_number*. As suggested by this example, functional properties are often used to indicate unique identifiers, analogously to a unique key in a database.

inverse functional, i.e. *a R b* and *c R b* implies that *a* and *c* represent the same instance, or alternatively for any given property and any given object for that property, there can only be one corresponding subject. Example: *is_father_of*.

transitive, i.e. *a R b* and *b R c* implies *a R c*. Example: *has_ancestor*

In this dissertation, property characteristics are defined using the *Characteristic* keyword, e.g.:

has_father Characteristic asymmetric

It is possible to state that one property is the inverse of another, e.g.:

is_father_of InverseOf has_father

It is also possible to state that one property is a subproperty of another, e.g.:

has_child SubPropertyOf has_descendant

Note that, *in general* subproperties do not inherit characteristics from their superproperties. For example, the symmetric, reflexive and transitive characteristics are not necessarily inherited. Indeed, the concept of inheritance is inappropriate in this context. That said, it is the case that there are certain characteristics which, if possessed by a property, will also be possessed by any subpropropety. An example is functionality, since if a subproperty has two different object entities for the same subject, then so will its superproperty. Hence if a superproperty is functional, its subproperty must also be functional. Thus, each characteristic must be considered separately and this needs to be stressed when teaching DLs. For reference, Table 2-1 lists the OWL property characteristics, indicating whether the characteristic is transmitted to subproperties, and sketching a proof or providing a counterexample as appropriate[16].

The property characteristics in OWL are only a subset of the possible relational characteristics which can be studied logically. For example, OWL does not possess an intransitive characteristic. Such a characteristic would mean that, for an intransitive property *R*, *a R b* and *b R c* implies that it is not the case that *a R c*. Halpin and Curland (2011) discuss a wider range of property characteristics available in Object Role Modelling[17], commenting on which of these characteristics are available in OWL. They note that some characteristics, or combinations of characteristics imply other characteristics. As an example, they show that a symmetric and transitive property will also be reflexive. In addition, some combinations of characteristics are not possible, e.g. a property cannot be both reflexive and irreflexive[18].

---

[16] Characteristics can be grouped into non-inheritable and inheritable characteristics:

(i) Non-inheritable characteristics
These are characteristics which imply the truth of additional property statements. E.g., if *S* and *T* are symmetric and transitive properties respectively:
$a\ S\ b \Rightarrow b\ S\ a$
$a\ T\ b;\ b\ T\ c \Rightarrow a\ T\ c$
These characteristics are not inherited because there is no guarantee that the inference will be valid when the superproperty is replaced with a subproperty.

(ii) Inheritable characteristics
These are characteristics which imply the falsehood of additional property statements. E.g., if *A* and *F* are asymmetric and functional properties respectively:
$a\ A\ b \Rightarrow \neg\ b\ A\ a$
$a\ F\ b;\ \neg\ (b = c) \Rightarrow \neg\ a\ F\ c$
These characteristics are inherited because the inference remains valid when a superproperty is replaced with a subproperty; were the inference with the subproperty not valid, then the original inference with the superproperty could not be valid.

Note that the reflexive and irreflexive characteristics are particular cases of (i) and (ii) respectively, where the set of statements on the left-hand side of the inference is the empty set. If *R* and *I* are reflexive and irreflexive properties respectively, then *a R a* is always true and *a I a* is never true. Using $\emptyset$ to represent the empty set, it is possible to write:
$\emptyset \Rightarrow a\ R\ a$
$\emptyset \Rightarrow \neg\ a\ I\ a$

[17] http://www.orm.net/

[18] Strictly speaking, an exception to this is the bottom property. Baader et al. (2017) define the bottom property as "a property whose extension is empty in every interpretation". Thus it is the property for which there are no property assertions, i.e. the set of property assertions is the empty set. Any statement is, therefore vacuously true of the elements of this set, and hence the bottom property can be said to possess any characteristic.

Table 2-1 OWL property characteristics, indicating whether a characteristic is transmitted to subproperties.

characteristic of P	**transmitted to subproperty S**	**proof or counterexample**
transitive	No	property *descendant_of*, subproperty *child_of*
functional	Yes	$a\ S\ b \Rightarrow a\ P\ b$ and $a\ S\ c \Rightarrow a\ P\ c$. So $a\ S\ b$; $a\ S\ c \Rightarrow a\ P\ b$; $a\ P\ c \Rightarrow b \equiv c$ (by functionality of P). Hence *S* is also functional.
inverse functional	Yes	$a\ S\ b \Rightarrow a\ P\ b$ and $c\ S\ b \Rightarrow c\ P\ b$. So $a\ S\ b$; $c\ S\ b \Rightarrow a\ P\ b$; $c\ P\ b \Rightarrow a \equiv c$ (by inverse functionality of P). Hence *S* is also inverse functional.
symmetric	No	property *sibling*, subproperty *sister*
asymmetric	Yes	Given $a\ S\ b$ assume $b\ S\ a$. Then: $a\ S\ b \Rightarrow a\ P\ b$ and $b\ S\ a \Rightarrow b\ P\ a$ However, we cannot have both $a\ P\ b$ and $b\ P\ a$, by asymmetry of P. Hence we cannot have $b\ S\ a$, i.e. $a\ S\ b \Rightarrow \neg\ b\ S\ a$
reflexive	No	property *greater_than_or_equal_to* subproperty *greater_than*
irreflexive	Yes	$a\ S\ a \Rightarrow a\ P\ a$, which contradicts the irreflexivity of P. Hence $\neg\ a\ S\ a$, i.e. *S* is irreflexive.

N.B. in the proofs, P is the superproperty, assumed to possess the particular property characteristic; S is a subproperty of P

Strictly speaking, much of the foregoing discussion is related to object properties, i.e. relations between instances. Besides classes, instances and properties, OWL also contains **literals**. As a result, besides object properties there are also **datatype properties**, i.e. describing relations between an instance and a literal, e.g. *John has_age 30*. Datatype properties can also have subproperties, which are also datatype properties. However, it is not possible to define the inverse of a datatype property, and with the exception of functional, the characteristics above relate only to object properties.

2.4. Difficulties with DLs

Chapter 1 noted that there has been a small number of papers reporting on user difficulties with DLs. Rector et al. (2004), based on their experience of teaching OWL DL, found four logical issues which users find particularly difficult. Expressed in MOS, these are:

1. A tendency to assume that *only* implies *some*.
2. Confusion between *and* and *or*.
3. A combination of (1) and (2), where the use of *and* between disjoint classes creates Nothing (⊥), which leads to the trivial satisfaction of a universal restriction.
4. Confusion between *P some (not X)* and *not (P some X)*.

To illustrate (3), referring to the pizza ontology (see Horridge et al. (2011)) which was used by Rector et al. for training, students may write *has_topping only (Meat and Fish)* when they mean *has_topping only (Meat or Fish)*. The former is a valid class definition, but can only be satisfied by pizzas with no topping, since *Meat and Fish* is the empty class, i.e. the class *Meat* and the class *Fish* are disjoint. The authors also proposed a set of paraphrases for the main OWL constructs, suggesting that these be used in the comments on the ontology to help confirm that the constructs are achieving what is desired.

Kalyanpur et al. (2006) have added four additional issues to this list, based on their analysis of unsatisfiable classes found in ontologies. Expressed in MOS, these are:

1. Writing *EquivalentTo* when *SubClassOf* is meant. They regard this as "the most common reason for unsatisfiability [of a concept]".
2. Writing *A SubClassOf B* and *A SubClassOf C* when what is intended is *A SubClassOf (B or C)*.
3. Writing *P Domain A* and *P Range B* instead of *A SubClassOf P only B*.
4. Writing *P Domain A* and *P Domain B* instead of *P Domain (A or B)*.

Note that (2) and (4) are variants on the confusion between *and* and *or* identified by Rector et al. (2004).

As noted in section 1.1, Horridge et al. (2011) were interested in the relative comprehensibility of different justifications for a particular entailment. They proposed an intuitive cognitive complexity model to compute the complexity of any particular entailment, based on the DL constructs used in the justification. They tested the model by providing study participants with six justifications and in each case asking the participants to confirm whether or not a putative entailment could be inferred. The model categorised the justifications into three easy and three hard and found that this categorisation was consistent with participants' experience for four of the six justifications.

Nguyen et al. (2012) were interested in the relative comprehensibility of deduction rules. They presented participants with 51 deduction rules with valid inferences, plus some with non-valid inferences. Participants were asked to confirm the validity, with success rates for the 51 valid rules ranging from 1 to 0.04. These success rates, which they termed facility indexes, could then potentially be used by an algorithm to optimise the comprehensibility of a deduction chain.

2.5. Ontology development

2.5.1. General purpose ontology development tools

Mention has already been made of the GUI-based tool, Protégé, which was one of the earliest ontology editors. Protégé continues to be developed and is one of the most widely used. TopQuadrant[19] (2014) provide a commercial ontology editor, TopBraid Composer, similar in philosophy to Protégé. There are now a very wide range of plug-ins available for Protégé[20], including reasoners. The current version of Protégé is downloaded with the Hermit reasoner (Glimm et al. 2014), one of a number of highly optimised reasoners. Some reasoners are optimised for particular species of OWL. For example ELK, also available for Protégé, is optimised for OWL 2 EL (Kazakov et al. 2012).

A considerable number of ontology visualization tools have also been developed, some as plug-ins for Protégé. Visualizers are frequently designed primarily to navigate all or part of an ontology. An early one was Jambalaya (Storey et al., 2001) which adopted techniques previously used in software visualization. Another application of visualization is to compare ontologies, e.g. OWLDiff, see Křemen et al. (2011). Both these visualizers are available as Protégé plug-ins. An early review of ontology visualization tools was provided by Katifori et al. (2007). Ramakrishnan and Vijayan (2014) provide a more recent review.

[19] http://www.topquadrant.com
[20] For Protégé plug-ins, see http://protegewiki.stanford.edu/wiki/Protege_Plugin_Library

Some visualization tools show only the subsumption hierarchy. The classic problem here is enabling users to focus in on a part of the ontology whilst seeing that part in the context of the whole ontology. Katifori et al. recommend different visualizers for different size ranges. Some do show object properties as well as the subsumption hierarchy. OWLPropViz[21] was designed specifically to show the object properties; the user can select which properties are to be shown, and also whether to show subsumption, disjointedness and class equivalence. However, where object properties are shown there is very rarely any support for human reasoning about those properties. Systems which do support such reasoning are described in subsection 2.5.5.

Apart from problem of actually displaying large ontologies, the layout of such ontologies creates computational problems. A recent visualizer, NavigOWL (Hussain et al., 2014), is based on the observation that when an ontology is represented as a graph, a small proportion of the nodes are *power nodes*, i.e. highly connected to other nodes. Taking account of this fact enables the design of a more efficient layout algorithm than would be possible if all nodes were treated equally.

2.5.2. Textual approaches

GUI tools, such as Protégé, are widely popular; however Lord (2015) has noted that they are not ideal for large ontologies, or ontologies with a great deal of repetition. Such ontologies were part of the motivation for developing textual approaches to creating OWL ontologies, based on particular programming languages. Ogbuji (2008) developed InfixOWL building on Python and using an infix syntax based on MOS. For example the InfixOWL *some* and *only* operators are associated with methods that construct the existential and universal restrictions. Nickles (2014) has taken a functional-logic programming approach based on Javascript and JSON.

Matassoni et al. (2014) developed a language based on the LaTeX[22] syntax. They surveyed ten knowledge engineers to obtain their views on the intuitiveness and conciseness of their syntax, compared with five other OWL syntaxes, including MOS. To do this they presented the engineers with ten samples of OWL usage, each one represented in each of the six syntaxes. The LaTeX-like syntax and MOS were very similarly regarded for intuitiveness, and clearly ahead of the next best, which was the functional syntax. The LaTeX-like syntax was the clear favourite for conciseness, well ahead of functional syntax and then MOS. The other three syntaxes used, Turtle, OWL/XML and RDF/XML, were very poorly regarded for both intuitiveness and conciseness.

Lord (2015; 2013) has taken a related approach in developing Tawny OWL. Tawny OWL makes use of Clojure, a dialect of Lisp built on a java virtual machine, thereby providing easy access to the OWL API. It uses the MOS keywords but a prefix notation taken from Lisp. Functions are variadic, e.g. *or X Y Z* indicates the disjunction of *X*, *Y* and *Z*. Abstraction, e.g. the definition of patterns, is enabled via Lisp functions and macros. There are also some in-built patterns, e.g. value partition[23]. Apart from supporting repetition, the programming approach enables the use of environments and tools for software development rather than developing tools specifically for ontology development. The use of macros and functions means that the programmer's intention is more evident from the source. Relatively

[21] http://protegewiki.stanford.edu/wiki/OWLPropViz
[22] https://www.latex-project.org/
[23] For a discussion of value partitions, see Rector (2005).

compact source code, with well chosen macro or function names, is more readable than longer source code with repetitive sections.

2.5.3. Controlled Natural Languages

The textual systems described in the previous subsection were based on formally defined computer languages. An alternative approach is to use a controlled natural language (CNL), i.e. one based on a natural language. CNLs have a long history; they were used for communication, e.g. between non-native speakers of English, prior to their use in computing. They have also been used as simplified languages to facilitate translation. Kuhn (2014) provides a survey and traces their development back to the 1930s. The essence of a CNL is that it is an approximate subset of a natural language designed to simplify the language. The adjective approximate is necessary because, as Kuhn (2014) points out, there may be "small deviations from natural grammar or semantics". Kuhn provides a classification of language based on four dimensions, represented by the acronym PENS, for precision, expressiveness, naturalness and simplicity. The last of these refers to the effort required to describe the language. Each dimension is divided into five categories. English lacks precision, is very expressive and natural, and very complex to describe. This is categorised as $P^1E^5N^5S^1$. Propositional Logic (PL), on the other hand, is very precise, neither expressive nor natural, but simple to describe. It is categorised as $P^5E^1N^1S^5$. Thus English and PL are at diametrically opposite corners of the four dimensional space used to describe language. MOS is categorised as $P^5E^2N^2S^4$, i.e. high in precision and simplicity but low in expressiveness and naturalness. Kuhn (2014) adopts the convention that to be classified as a CNL, a language must have a naturalness of three or greater. By this convention, MOS is not a CNL. Kuhn also notes that CNLs can be defined in a proscriptive or prescriptive fashion, or a combination of the two. In the proscriptive case the language definition specifies what is not allowed; in the prescriptive case what is allowed. More generally, we might regard CNLs as being derived from a natural language by the removal of those constructs which create ambiguity. On the other hand, MOS was created from logical notation by removing those symbols which were felt to be off-putting to some users, e.g. the existential and universal quantifier symbols, and replacing them with words from the English language.

Within the domain of computing, CNLs have been used for software specification, and more recently for ontology generation. Examples of such languages are ACE, Rabbit and Sydney OWL Syntax (SOS). ACE has been designed to translate natural language into first order logic, originally for software specification (Kaljurand, 2007; Kuhn, 2014),. Fuchs et al. (2006) describe its use for knowledge representation. Rabbit was designed specifically to translate into OWL, see Hart et al. (2007) for an introduction to the language. SOS, the design of which is discussed in Cregan et al. (2007) had a similar goal to Rabbit.

A comparison of the three languages is given by Schwitter et al. (2008). They highlight a design decision which differentiates ACE and SOS from Rabbit. With ACE and SOS, when stating the characteristics of an object property, dummy names are used. Thus, using the example from Schwitter et al. (2008), to define *larger than* as asymmetric, one writes in ACE: *if something X is larger than something Y then Y is not larger than X*. Similarly, in SOS one writes: *if X is larger than Y then Y is not larger than X*. However, in Rabbit the ontology author is expected to have an understanding of the terminology associated with object properties and to write: *the relationship "is larger than" is asymmetric*. This is an aspect of a general design decision about how much one wishes to protect the author from the technical language of DLs.

A number of researchers have investigated the use of CNLs for ontology creation. Engelbrecht et al. (2010) investigated the ease with which domain experts could use the Rabbit language. Apart from suggesting some changes to Rabbit syntax, this highlighted the need for an authoring tool.

Indeed, a Protégé plug-in for Rabbit (ROO) has been produced, see Denaux et al. (2010), although not used in the Engelbrecht et al. study. Another Protégé plug-in, ACEView (Kaljurand, 2008), has been developed for writing ACE statements. As well as generating OWL, ACEView can generate SWRL (Horrocks et al., 2004) and DL-Query[24] statements, thereby offering a unified interface to the three languages.

Dimitrova et al. (2008) compared ROO with ACEView. They found a number of errors which occurred with both tools and which may be a feature of ontologies created from CNLs:

- Some classes had a number of immediate superclasses; they regarded this as an error although in some cases it might be the correct approach.
- What should have been classes were recorded as instances.
- Redundant axioms were entered, i.e. axioms which repeated knowledge already present in the knowledgebase.

Kuhn (2013) has compared the understandability of ACE with a simplified form of MOS which he named 'Manchester-like language' (MLL). This was done by showing participants a diagrammatic representation of a small ontology and ten statements in either ACE or MLL. The participants were required to indicate whether each statement was true or false, based on the diagram. Each participant did the test twice, once with ACE and once with MLL, with a different diagrammatic ontology in each case. Overall, performance was significantly better with ACE than with MLL. The design of the experiment enabled comparison of equivalent statements in ACE and MLL. There was a particularly large difference between ACE and MLL in a question where an existential restriction was used with a complemented class ("John buys something that is not a present" versus "John HasType buys some (not present)". All of the 16 participants who were shown this statement in ACE responded correctly, whilst only 8 out of 16 responded correctly when shown the statement in MLL. This may relate to the confusion noted by Rector et al. (2004) between *P some (not X)* and *not (P some X)*, see section 2.4 above.

Another marked difference was between the expression of object property equivalence in ACE ("If X loves Y then X helps Y. If X helps Y then X loves Y") and in MLL ("loves EquivalentTo helps"). The former elicited 14 out of 16 correct responses, the latter 8 out of 14. This difference might disappear or at least be reduced once people had more experience with MLL and were used to the *EquivalentTo* statement. Indeed, Kuhn comments on the need to understand the learning curve for both types of languages.

Kuhn's work suggests that a well-designed CNL can be more understandable than MOS, which is generally regarded as the most readable of the standard DL syntaxes. However, Kuhn was not concerned with the ability to reason using DL statements.

Smart (2008), in a detailed survey of CNLs for the 'Semantic Web' presents an interesting argument and counter-argument. He is one of the few in this area who refers to the cognitive psychology and linguistic literature. He starts by arguing that reasoning is not something which people naturally do well, citing some of the same kinds of studies as will be discussed

[24] http://protegewiki.stanford.edu/wiki/DLQueryTab

in chapter 3 of this dissertation. He then presents the view that CNLs should support people in reasoning better than more formal syntaxes since they are close to natural languages which presumably have been optimised through evolutionary pressure. Finally, he makes the antithetical point that the apparent ease of interaction which is observed with natural language may not equate to ease of understanding. He observes that the psychological studies which have highlighted our difficulty with reasoning have used "sentential structures", i.e. natural language.

Perhaps natural language has been optimised by evolution. However, it is extremely unlikely that it has been optimised for understanding and reasoning about complex logical scenarios.

The debate about the use of natural language in semantic technologies echoes an earlier one about natural language in database systems. Shneiderman (1978), in a paper arguing for a greater consideration of human factors in database research, reports an experiment comparing the use of natural language with a formal database query language. In a within-participants study he provided participants with a scenario and asked them to formulate queries in both English and the formal query language. There was no significant difference between the number of valid queries constructed in the two languages. However, participants created significantly more invalid queries in English than in the formal language. Shneiderman (1978) specifically avoids an outright condemnation of natural language in database applications but concludes that, amongst other things, querying in a natural language requires a knowledge of the structure of the data, since there is no formal language to guide users. This is supported by the fact that, in the experiment, there was a significant order effect; participants who were asked to use first the formal language, and then English, created fewer invalid queries than participants who were asked to use English first.

The largely proscriptive approach of CNLs and the prescriptive approach of MOS may be relevant for different sets of users. Hendler (2015) talks about 'big O' ontologies and 'small O' ontologies. The former find their application in specialized fields such as biological research. The latter are found at Web scale constituting the LOD cloud. He suggests that for the Web scale applications users can tolerate that not every answer is correct, just as they can tolerate incorrect responses to Google searches. It may be that CNLs are of more use for non-specialist users working with these Web scale applications, e.g. semantic wikis, where a degree of error can be tolerated. MOS is likely to be more relevant for specialist users, e.g. in bioinformatics, who are prepared to learn some of the jargon of DLs.

2.5.4. Ontology patterns and pattern tools

A fundamental feature of ontology development is the ability to import and reuse existing ontologies. However, another means of supporting reuse is with ontology patterns. Ontology patterns exist at a number of levels of design; a categorisation framework is provided in Gangemi and Presutti (2009). Falbo et al. (2013) review the categorisation and relate it to the phases of ontology development. The OntologyDesignPatterns.org[25] portal lists a number of patterns, using this categorisation. Since the work reported in this dissertation is concerned with a relatively detailed level of design, only two pattern types will be considered: *logical ontology patterns* and *content ontology patterns*[26]. Logical patterns are independent of content. An example is a pattern to create an n-ary relation using

[25] http://ontologdesignpatterns.org

[26] Other pattern types, not considered here include architectural ontology patterns, which are concerned with the overall shape of an ontology, and reengineering ontology patterns, which are rules to transform one ontology to another.

as building blocks ternary relations available in OWL[27]. The category of logical patterns also includes logical macros; a simple example is *exclusive or*, see Horridge et al. (2006). Content patterns relate to a specific topic. OntologyDesignPatterns.org lists content patterns ranging from the relatively generic, e.g. a pattern to represent time intervals, to the industry specific, e.g. a pattern to represent catch records from the fishing industry. Logical patterns do not contain predefined individuals, classes or properties, except for generic ones such as Thing (⊤). They may contain rules to generate entities from the entities in the ontology, as is the case for the n-ary relation pattern. Content patterns, on the other hand, contain predefined entities, e.g. the time interval pattern contains the class *Time Interval*. Content patterns are, in fact, small ontologies.

There is a significant literature on ontology patterns, of which only a small part can be mentioned. The papers described here include some of the most important usability studies and tool developments as well as describing different approaches to pattern generation and use. Blomqvist et al. (2009) studied the use of content patterns by giving experiment participants a modelling task without access to content patterns and an analogous task with access to patterns. They concluded that patters were perceived as useful by developers and that they led to increased ontology quality. Three particular modelling 'mistakes' were identified when patterns were not available: failure to distinguish "between persons and their roles"; a similar failure to distinguish "between information and its physical realization"; and incorrect modelling of n-ary relations. With the use of patterns neither of the first two mistakes were observed and there were fewer incidences of the third. Blomqvist et al. (2010) describe user experiences with tool support designed specifically for patterns. Horridge et al. (2012) analyse patterns from the OntologyDesignPatterns.org library for their 'language expressivity', e.g. to determine within which species of OWL 2 they fit. Another research area is that of finding patterns, e.g. from a patterns library, and then integrating them into the ontology; Hammar (2014) discusses these issues.

Language and tools have been created to support pattern generation and use. Iannone et al. (2009) describe the Ontology Pre-Processor Language (OPPL). OPPL patterns specify axioms to be added or removed from an ontology. They also include variables, defined to be particular ontology entity types, which the developer associates with entity names whenever the pattern is to be instantiated. Aranguren et al. (2011) provide a user manual for the language. An OPPL plugin has been created for Protégé[28].

Tawny OWL, discussed above, incorporates closure and covering patterns, see Lord, (2013) and (2015). For example, *some-only* enables closure. Using Tawny OWL notation:

$$P \text{ some-only } X\ Y\ Z$$

indicates the class of individuals which are related as subject of property P to individuals from each of the classes X, Y and Z, and individuals from no other classes. Tawny OWL also enables the definition of additional patterns by the developer through use of the Lisp facility for defining macros. As already noted, Tawny OWL is intended

[27] See http://ontologydesignpatterns.org/wiki/Submissions:N-Ary_Relation_Pattern_%28OWL_2%29 and also Hoekstra (2009).

[28] See http://sourceforge.net/projects/oppl2/files/ for plugins for Protégé 4.0 and 4.1.

to support repetitive structures in ontologies, and is motivated by applications in biology where repetition is common.

Jupp et al. (2012) argue that there is a distinction between developing an ontology and populating it with instances. The former requires people with strong ontology skills, the latter may be done by domain experts with less training in ontologies. The two groups also require different tools. The ontology developers will typically use tools such as Protégé which are not appropriate for the domain experts. Moreover, in domains such as the life sciences ontology population is often highly repetitive, based on patterns. Jupp et al.'s solution to this was to design the Populous tool[29]. Populous provides the domain expert with a form, with a particular template appropriate to the repetitive structure. OPPL patterns (see above) are created by the ontologist to generate OWL from the tabular data.

2.5.5. Diagrammatic representation of ontologies

The visualization tools discussed in section 2.5.1 chiefly concentrate on subsumption relations between classes and the membership relation between individuals and classes. Some show properties, but not in a way which supports human reasoning about these properties.

It is, however, possible to reason diagrammatically, i.e. using rules for manipulating diagrams analogously to the way in which we reason with rules to manipulate text in an algebra. Dau and Eklund (2008) demonstrate the soundness and completeness of a diagrammatic reasoning system based on existential graphs (Peirce & Sowa, 2000) and applicable to the small DL called $\mathcal{ALC}$[30]. The operations in their calculus include adding and deleting branches to and from a tree structure. They also provide a survey of the use of diagrams in logic.

Some researchers are now developing diagrammatic approaches to formally representing DL ontologies, thereby aiding users in reasoning about those ontologies. Chapman et al. (2011) describe a formalism they call concept diagrams and provide a formal semantics. They derive sound inference rules, based in part on pattern matching, with which to reason about these diagrams. They argue also that these inference rules are intuitive, a consequence of "the syntactic properties of the diagrams reflecting the semantics". Shams et al. (2017) have developed a set of inference rules "for reasoning about inconsistencies and incoherence in ontologies".[31] Stapleton et al. (2013) illustrate how to represent common DL axioms using concept diagrams. A case study, using the Semantic Sensor Network[32] ontology, is described by Howse et al. (2011), concluding with requirements for tool support. They identify the need to be able to translate from symbolically specified ontologies to diagrams and for sketch recognition to translate from diagrams to symbolic form. Stapleton et al. (2014) provide another case study based on a cemetery ontology. They highlight the need for the development of effective layout algorithms. One aspect of this is ensuring that where

[29] Available from http://e-lico.eu/populous.html

[30] The computational complexity of DLs is conventionally described using letters written in a cursive script, as shown. $\mathcal{AL}$ represents a basic DL with negation of concept names on the right-hand side of axioms, intersection, universal restriction and limited existential quantification. $\mathcal{C}$ indicates the ability to negate concepts made up of other concepts (complex concept negation). A brief overview of this notation is given in Horridge et al. (2012).

[31] An iconsistent ontology "cannot have any model and, so entails anything"; an incoherent ontology "entails an unsatisfiable (i.e. empty) class or property" (Shams, et al., 2017).

[32] See Compton et al. (2012) for a description of the Semantic Sensor Network ontology.

multiple diagrams have common parts, these look identical in each diagram. Another aspect is the development of algorithms which enable the incremental updating of diagrams when an ontology is edited. The longer-term goal is to enable the diagrammatic editing of these diagrams, so that diagrammatic editing and a more conventional approach could be used complementarily.

A visual language for DL has also been developed by do Amaral (2013); he describes a study to compare understanding of this language with MOS[33]. Participants were asked a number of questions to test their understanding of ontologies described using the two approaches. The participants did better with the visual language than with the MOS. There appear to be a number of reasons for this. In one question, the difficulty may have arisen in part from a non-standard use of MOS. In another question involving *only*, the problem seems to be the familiar one of assuming that the universal restriction implies existence, i.e. ignoring the trivial satisfaction of the universal restriction. The visual language avoided this difficulty. Although not entirely without ambiguity, generally the visual language appeared to avoid some of the ambiguities and misinterpretations present in MOS.

2.6. Discussion

This chapter has explained the use of ontologies in formal modelling, and in particular DLs which are the most frequently used languages for creating ontologies. The prevalence of DLs is in part because the OWL family of DLs has been standardised by the W3C. It was explained that the formal notation used initially for DLs was thought to be a problem for non-logicians, and this has led to a variety of other syntaxes. MOS is now commonly used and has been incorporated into the Protégé tool. None of the commonly used syntaxes appears to have been rigorously tested, nor does their design seem to have been influenced by any considerations from cognitive psychology or linguistics. Yet studies have been described which identify difficulties experienced with these notations. The studies reported later in this dissertation will further illustrate some of these difficulties.

MOS uses keywords from English. Its constrained structure and limited vocabulary mean that it would not normally be regarded as a CNL. A number of CNLs have been designed for translation into DLs. Some of these have been subjected to user testing. However, CNLs can be verbose and ambiguous. Kuhn (2013) admits that they can be difficult to write, although tools such as ROO (Denaux et al., 2010) and ACEView (Kaljurand, 2008), can help with this. They can also be verbose and ambiguous. At least one study has compared a CNL with MOS for understandability (Kuhn, 2013); this work suggest that MOS could be improved upon. More fundamentally, none of the studies have compared how easily people can reason when confronted with statements in the various DL languages. Yet this is a key requirement, for example during debugging when attempting to understand how a particular entailment follows from a justification. Indeed, some of the problems people have in reasoning about DL statements may transcend issues of syntax; for example the trivial satisfaction of the universal restriction discussed in subsection 2.2.1. This is not to say that particular syntaxes may not help mitigate them.

The use of patterns offers the opportunity to work at a higher level of abstraction. Whilst there has been some interest in pattern research, there is no evidence of widespread systematic use of ontology patterns. This point is discussed in chapter 5. The textual approach of Lord (2015; 2013) which builds on the availability of software engineering tools

[33] In his study the MOS keywords were translated into Brazilian Portuguese. He maintains that the same ambiguities and "undesired interpretations" exist in the translated version as with the English keywords.

would appear to encourage the use of patterns. However, the DL still needs to be represented in some syntax, e.g. MOS, and the problems of comprehension and reasoning remain.

Another approach is the graphical representation of ontologies. Here there are broadly two categories. Most visualization tools are designed to enable navigation through an ontology. A few support comprehension and reasoning. One such, developed by do Amaral (2013), has been compared with MOS. It is interesting that one of the advantages of the visualization was to remove the confusion over the precise meaning of *only*. It may be that it is not visualization of itself which is valuable, but rather that it avoids some areas of confusion. There may be other, non-visual, ways of achieving this.

CNLs and visualizations are likely to have a continuing role in the development of ontologies. It is also likely that, where conciseness and precision are required, a more formal DL syntax will be required. The widespread acceptance of MOS suggests that this more formal syntax will employ some English words, rather than a purely mathematical notation. Indeed, the widespread acceptance of MOS suggests that any improvements are likely to be evolutionary. This highlights the need to use cognitive and linguistic theories to better understand the difficulties experienced with MOS, and to mitigate those difficulties. This will be the theme of later chapters.

3. Theories of reasoning

For, if some of the moderns have thought to enlarge its [logic's] domain by introducing psychological discussions on the mental faculties, such as imagination and wit, ...: this attempt, on the part of these authors, only shows their ignorance of the peculiar nature of logical science.

Immanuel Kant, 'Critique of Pure Reason', Preface to second edition, 1787

Logic is the mirror of thought, and not vice versa.

Jean Piaget, 'The Psychology of Intelligence', 1947

Psychologists have been investigating human reasoning for a considerable time. Bucciarelli and Johnson-Laird (1999) cite Störring (1908) as reporting early psychological research on syllogisms. However, modern investigations into human reasoning are generally regarded as beginning with the British cognitive psychologist Wason. A classic experiment of his is described in Wason (1968). In the experiment the participant is presented with four cards laid face down on a table and told that each card has a letter on one side and a number on the other. The participant is also told that there is a putative rule expressed as a conditional statement: "if there is a vowel on one side of the card, then there is an even number on the other side" (Wason, 1968, p. 273). The visible sides of the cards show a vowel, a consonant, an even number and an odd number. The participant is required to turn over just those cards necessary to test out the rule. This task has come to be known as the 'Wason selection task'. The correct response is, of course, to turn over two cards displaying the vowel and the odd number. In fact, the most common choices are the cards displaying the vowel and the even number. Wason was surprised by this result, in part because it seemed to contradict the contemporary psychological theory concerning intellectual development.

Much has been written about this, and other similar experiments. It has set the tone for subsequent research into reasoning, both in terms of content and in terms of methodology. In terms of content, research has looked at the traditional logical forms such as the conditional, which underlies the Wason selection task, the disjunction, conjunction and the syllogism. In terms of methodology, there has been a concern with identifying the mistakes which people make, and using consistent mistakes to support or refute particular theories.

Johnson-Laird (2010, page 194) points out that early theories of reasoning assumed "an unconscious logical calculus" which later was assumed to contain "formal rules of inference". The expectation was that people reason using formal rules of logic, in some way similar to the processes of a trained logician. This is variously referred to as the *sentential*, *rule-based* (Stenning & Yule, 1997) or *mental logic* (Oaksford & Chater, 2001) school. The phrase *rule-based* will be used in this dissertation, to emphasize the difference from the *model-based* approach described later. Rips (1983) is representative of the rule-based approach. He suggests that people use rules such as modus ponens, and that there is an availability associated with each such rule. By 'availability', Rips (1983) means the ease with which the rule can be retrieved from memory. He associates an availability parameter, between 0 and 1, with each rule to represent the ease of retrieval. The greater the availability of the rules being used, the greater the probability of arriving at the correct conclusion. Rips used a computer program to simulate his proposed scheme, and this is a feature which recurs in reasoning research.

An alternative school, in particular associated with Johnson-Laird, suggests that in general people do not reason with formal rules, but instead build *mental models*[34] of any given situation. They then test out putative inferences for consistency with the proposed models. If no inconsistency can be found, then the inference is accepted. Ehrlich and Johnson-Laird (1982) describes an early attempt to use mental models to explain experimental results. The models they describe relate to the layout of objects in two-dimensional space. However, mental models can be used to represent more abstract situations. Bucciarelli and Johnson-Laird (1999), for example, interpret relative difficulties with syllogisms in terms of mental model theory.

The distinction between the rules-based and model-based theories of reasoning can be regarded as analogous to the distinction in logic between syntactical and semantic reasoning. A mental model can be regarded as an interpretation of the corresponding syntactical system.

Neither the rules-based nor the mental models approaches are able to represent probabilities. However, Oaksford and Chater (2001) point out that there are statements which are commonly made, but commonly understood to be not 100% true. For example, the statement 'birds fly' is not true for penguins, and neither for very young birds. Thus 'Probability (birds fly) = 0.9', say, better captures our understanding of the statement. Oaksford and Chater (2001) propose a probabilistic interpretation of human reasoning which provides a mathematical framework to understand people's reasoning performance. Unlike the rules-based and mental model theories, the probabilistic approach does not in general attempt to provide an explanation of how people reason. However, Oaksford and Chater (2001) describe a probability heuristic model which interprets the syllogism quantifiers probabilistically and proposes a set of heuristics to enable probabilistic reasoning.

Another approach is to view every reasoning step as being about a relation between elements. The number of elements involved can then be regarded as a measure of the complexity of the reasoning step. This is known as the *relational complexity* of the reasoning step (Halford et al., 1998). This viewpoint can be regarded as complementary to the other schools. Zielinski et al. (2010) have attempted an integration of the relational complexity and mental model approaches in the case of the syllogism.

The next three sections discuss the three theoretical frameworks which are used to interpret results and form hypotheses in the studies described in this dissertation: rule-based; mental models; and relational complexity. At the end of his paper, Wason (1968) discusses whether the difficulty reported is caused by the structure of the rule used or by the verbal representation of the rule. This theme is taken up in section 3.5 which discusses some of the issues which arise with language. These issues can influence how some of the reported experiments on reasoning should be interpreted and have a bearing on the studies described in this dissertation. Finally, section 3.6 makes some general comments.

Before proceeding, a comment needs to be made on terminology. In the computer science literature, the word *reasoner* refers to software which undertakes formal reasoning. This is the sense in which it was used in chapter 2. In the psychological literature, however, the word refers to an individual who is undertaking a reasoning task. Sometimes the phrase *naïve reasoner* is used to emphasize that an individual, e.g. a study participant, has had no training in logic. In this dissertation, the single word reasoner will always refer to computer

[34] Throughout this dissertation, the phrase *mental model* will be used with the specific meaning described here, rather than any more general meaning with which it might be associated in everyday speech.

software. An individual undertaking a reasoning task will be referred to as a *human reasoner*.

3.1. Rule-based

To anyone with any training in logic, be it the verbal formalisms of Aristotle or the more mathematical formulation developed since the 19th century, it may seem natural to suppose that human reasoning is based on some kind of formal system. Certainly, as already noted, the first approaches to modelling human reasoning were rule-based.

Rips (1983) has developed a model of propositional reasoning, the ANDS model, which he incorporated into a computer program. The model employs 11 logical rules. For example, one rule is modus ponens: *if p then q; p; therefore q*. Another is the use of one of De Morgan's laws to transform *not (p and q)* into *not p or not q*. The rules are used in forward and backward modes, i.e. working forward from an assumption and backward from a goal. Some of the rules exist in both modes. Proofs are constructed in a 'working memory' with a defined memory limitation. Heuristics are used to determine how to deploy rules.

The availability of each rule has a probability associated with it. Thus it is possible to compute the probability of constructing a particular chain of reasoning, from assumption to goal. To calculate these probabilities, Rips undertook an experiment in which participants were given a number of reasoning tasks, each of which could be regarded as requiring a sequence of rules. Specifically, participants were asked to confirm whether a putative conclusion followed from a set of statements. Based on the results of this experiment, and using a least-squares criterion[35], Rips was able to estimate the probabilities for each of his rules. He reports a correlation of 0.933 between the observed and predicted accuracies.

Rips does not claim his system is complete, in the sense of containing all the rules which human reasoners use. It does, though, demonstrate the feasibility of the approach. Although it does not prove that this is the way in which humans actually reason, Rips claims that the system was motivated by insight gained from previous experiments in which participants thought aloud about particular reasoning tasks.

Braine (1978) provides another example of this rule-based approach applied to propositional reasoning. He argues that our reasoning with natural language does not correspond to standard logic. For example, *if p then q* has a directionality, from *p* to *q*, in natural language which is absent from standard logic. In everyday discourse we are not concerned with what happens when *p* is not true. As a result, in natural language one would not normally associate a truth value with the statement when *p* is false, as one does in standard logic. He develops a natural propositional logic based on the rules he claims we ordinarily use, but which is equivalent to standard propositional logic.

Braine and O'Brien (1991) have revised Braine's (1978) theory of propositional reasoning and undertaken a detailed analysis of how this revised theory can explain reasoning with conditional statements. The theory proposes "lexical entries" for the logical operators *and*, *or* and *if*. These lexical entries contain inference schemas which differ in some respects from what would be expected from standard logic. Moreover, there is nothing analogous to the truth table approach of standard logic. The lexical entry for *if* contains a schema for modus ponens, i.e. *p; if p then q; therefore q*. This explains why people generally apply modus ponens correctly. However, there is no inference schema for modus tollens, i.e. *not q; if p*

[35] Presumed to mean that the availability parameters were chosen so as to minimize the sum of the squares of the differences between the observed and calculated proportion of correct responses for each question.

then q; therefore not p. This has to be achieved by reductio ad absurdum, i.e. by assuming *p*, deducing *q* from modus ponens, and finding a contradiction between this deduction and the original assumption of *not q*. This explains the greater difficulty experienced with modus tollens.

Additionally, Braine and O'Brien's (1991) logic rejects the so-called "principle of explosion" of standard logic, whereby anything can be proven from a contradiction. Rather, the only conclusion from a contradiction is that one of the assumptions is wrong. They claim that their revised model better explains a variety of reported data.

3.2. Mental models

The theory of mental models has been developed over a number of years by a number of researchers, in particular Philip Johnson-Laird and Ruth Byrne. There have been a great number of publications on the subject. The publications discussed in this section are chosen to provide an overview. In particular, they cover the various logical structures investigated: disjunction and conjunction; the conditional and biconditional; negation; and the syllogism. Johnson-Laird (2005) provides a review of the subject, whilst Johnson-Laird (2004) puts the theory in its historical context, tracing it back to the work of the American logician C.S. Peirce, through Wittgenstein's (1922) picture theory of meaning, and Craik's (1967) view that cognition is based on forming models of the world. Of particular significance is Johnson-Laird (2004) reference to Peirce's categorisation of the properties of signs into: iconic, i.e. possessing structural similarity; indexical, i.e. entailing some sort of physical connection; and symbolic, i.e. making use of an implicit or explicit representational convention. Johnson-Laird (2004) argues that mental models cannot be purely iconic, since some sort of symbolic representation is required, e.g. to represent negation and disjunction.

The essential principle of the mental model theory of reasoning is that, in order to reason, we create a model or models of a given situation and then check any putative conclusion against those models. Any putative conclusion needs to be checked for consistency with all the models. Mental model theorists argue that with a greater number of models there is a danger of one or more being forgotten, or not even created. This could lead to the acceptance of an invalid conclusion, i.e. one which is consistent with some but not all the necessary models.

3.2.1. Conjunction and disjunction

An example of the simplest possible situation, i.e. with only one model, is given by conjunction: *there is a circle and there is a triangle*. Using C and T to represent circle and triangle, this can be represented by:

C T

Exclusive disjunction: *there is a circle or there is a triangle, but not both*; requires two mental models:

C

T

Inclusive disjunction is: *there is a circle or there is a triangle, or both*; requires three models:

C T

C

T

The first of the three lines above describes the situation in which there is both a circle and triangle, the second line the situation in which there is only a circle, and the third line in which there is only a triangle.

On this basis we would expect people particularly to experience a difference in difficulty between the two extremes of these three situations, i.e. between conjunction and inclusive disjunction. Khemlani et al. (2012a) did find that human reasoners made more errors in interpreting inclusive disjunction than conjunction. In an experiment, the former yielded 68% correct responses compared with 86% correct for the latter.

3.2.2. The conditional and biconditional

The conditional, *if there is a circle, then there is a triangle* illustrates another feature of the theory. Initially, there may appear to be only one model, identical to that for conjunction:

C T

This is sufficient for a certain form of reasoning, i.e. modus ponens. Johnson-Laird et al. (1992) have shown experimentally that people do reason more accurately with a conditional, when the problem would require the application of modus ponens, than with an exclusive conjunction. Given that the former requires one mental model and the latter two, they regard this as evidence that people are using mental models with which to reason.

However in certain situations, specifically where a logician would apply modus tollens, three models need to be constructed to represent the conditional. These models correspond to the three valid rows of the truth table for implication, i.e. excluding only the row corresponding to a true antecedent and a false consequent. Using the symbol '¬' to represent negation, the models are:

C T

¬C T

¬C ¬T

These models are more detailed than those for the conjunction and disjunction, in that they explicitly show the negated statements, i.e. in this case the absence of a circle or a triangle. These are referred to as *explicit models*.

The biconditional, *if and only if there is a circle, then there is a triangle*, requires only two models, again corresponding to the two valid rows in the truth table:

C T

¬C ¬T

Johnson-Laird et al. (1992) showed that modus tollens with a biconditional was significantly easier than with a conditional. They attribute this to the fewer number of models required for the biconditional than for the conditional.

In fact, a key aspect of mental model theory is the claim that frequently human reasoners do not form fully explicit models but neglect the negated, i.e. false models. Johnson-Laird (1999, page 116) formulates this as *the principle of truth*:

> *"Individuals minimize the load on working memory by tending to construct mental models that represent explicitly only what is true, and not what is false."*

Not only does this reduce the number of models required in working memory but avoids negative statements which may require more cognitive work than affirmative ones (Pinker, 2014, chapter 5). In certain situations the principle of truth reduces the number of models but still leads to the correct conclusion. In other situations it can lead to false conclusions.

Johnson-Laird et al. (1992) observe that for propositional logic, the fully explicit model corresponds to the disjunctive normal form (DNF) of the expression, where each line corresponds to a disjunct.

3.2.3. Negation

It was noted in the previous subsection that conjunction required one mental model whilst inclusive disjunction required three. When these operators are negated the situation is reversed[36]. *It is not the case that there is a circle and a triangle* corresponds to three mental models:

¬C ¬T

¬C T

C ¬T

Whilst *it is not the case that there is a circle or a triangle or both* corresponds to one mental model:

¬C ¬T

Thus we would expect reasoning with negated conjunction to be harder than reasoning with negated inclusive disjunction. Khemlani et al. (2012a) did indeed find this to be the case. In the same experiment as already mentioned in subsection 3.3.1, they found that negated conjunction was interpreted correctly by 18% of the participants whilst negated inclusive disjunction was interpreted correctly by 89%. In fact, 45% of the participants interpreted negated conjunction as equivalent to "not-A and not-B", i.e. they created only the first of the three mental models shown above.

Khemlani et al. (2012a) also report on negation of the conditional, where they found that 59% of responses equated a negation of *if A then B* as *if A then not B*. In a later study they elaborate their ideas (Khemlani et al. 2012b), pointing out that negation involves creating a set of models complementary to the set describing the negated assertion. This is illustrated by comparing the models for negated conjunction and negated inclusive disjunction in this subsection with the models for conjunction and inclusive disjunction in subsection 3.2.1. They suggest that people tend to "assign a small scope to negation in order to minimize the number of models of possibilities that they have to consider" (Khemlani et al., 2012b, page 1). This makes sense in the case of negated conjunction where, in the example above, a

[36] This reversal is evident from the application of De Morgan's Laws:
not (circle and triangle) ≡ not circle or not triangle
not (circle or triangle) ≡ not circle and not triangle

small scope negation would presumably imply negating *there is a circle* and *there is a triangle*, rather than the whole statement. However, in the case of the conditional, both the assertion (*if A then B*) and the small scope interpretation of its negation (*if A then not B*) require three mental models for their full interpretation, whereas the true negation (*A and not B*) requires only one.

An alternative view might be that the difficulty of creating the correct mental model for the conditional leads to a syntactic strategy, which in turn leads to the error of changing the scope of the negation from the whole statement to the consequent. This might be seen as analogous to the confusion between the OWL statements *P some (not X)* and *not (P some X)* reported by Rector et al. (2004) and discussed in section 2.4.

3.2.4. The syllogisms

The syllogisms were first formulated by Aristotle. Because of their importance in psychological research into reasoning, an overview is provided here. Briefly, they consist of two premises and a conclusion. Each of these three statements is in one of four forms:

All X are Y

Some X are Y

No X are Y

Some X are not Y

Each syllogism has three terms, e.g.: *A*, *B*, *C*. One of the premises contains the terms *A* and *B*. The other premises contains the terms *B* and *C*. The conclusion contains *A* and *C*. In a valid syllogism the elimination of the *B* term allows the conclusion to state a relationship between *A* and *C* in one of the four forms. Thus a simple valid syllogism is:

All A are B

All B are C

Therefore, all A are C

There are four ways of ordering the terms in the two premises, referred to as the four *figures*. These are shown in table 3-1.

Table 3-1 Four figures of syllogism

1st premise	**2nd premise**	**position of common term (B)**
A-B	B-C	end of 1st premise, beginning of 2nd premise
B-C	C-B	beginning of 1st premise, end of 2nd premise
A-B	C-B	end of both premises
B-A	B-C	beginning of both premises

Thus, with four possibilities for each of the two premises and four figures, there are 64 (4^3) possible syllogisms. Of these, 27 syllogisms have valid conclusions, i.e. one can write a relationship between A and C using one of the four statements above. The other 37

syllogisms have no valid conclusion[37]. More detail is provided in Khemlani and Johnson-Laird (2012).

The syllogisms have been studied by a number of psychologists, including the mental model theorists. They encompass an enormous range of difficulty; Johnson-Laird and Byrne (1991, chapter 6, page 106) comment that "some of these 27 [i.e. the syllogisms with valid conclusions] are so easy that a nine year old child can spontaneously draw a correct conclusion, whereas others are so hard that barely any adults perform better than chance with them".

3.2.5. Mental models for the syllogism

The mental model representation of syllogism premises is more complex than that for propositional logic, and the notation different. Newstead (1995) uses as an example *Some A are B*. One mental model can be represented as:

a b

a b

a

b

There are two immediate differences from the notation for propositional logic. Firstly, the lowercase letters *a* and *b* are representative of individuals in the classes *A* and *B*, whereas for propositional logic the elements of the mental model represent specific individuals. Secondly, all four lines are part of the same mental model. The model states that there are individuals which are both *a* and *b*, and that there can be individuals which are *a* and not *b* or *b* and not *a*. It is not clear how the necessity of the first option is implied, possibly by the repetition of the line 'a b'.

However, Newstead (1995) also gives a second model for the same premise:

[a] b

[a] b

The square brackets indicate that *a* is exhaustively represented, i.e. there is no possibility of there being an *a* which is not a *b*. This is, in fact, the model for the premise *All A are B* which is included as a possibility in the original premise *Some A are B*. A fuller version of this model would require a line with only *b*, to indicate the possibility that some *B* may not be *A*.

Johnson-Laird and Bara (1984) suggest that a likely strategy when dealing with syllogisms is to construct a mental model for the first premise and then integrate the information from the second premise. They explain procedures by which this could be achieved.

Integration of the mental models from two premises can lead to more than one mental model for the syllogism. Johnson-Laird and Bara (1984) give as an example the syllogism:

[37] To be precise, there are no valid conclusions using the four statement forms provided by Aristotle. Stenning and Yule (1997) point out that, if we permit conclusions of the form *some not X are not Y*, then some of these other 37 syllogisms have valid conclusions.

All of the A are B

some of the B are C

This leads to two mental models. In the terminology of Johnson-Laird and Bara (1984), which is different from that of Newstead (1995), the two models can be represented:

(1)	(2)
a = b = c	a = b
a = b 0c	a = b 0c
0b	0b

Here the = indicates that an individual is in two classes and *0* indicates an individual which may or may not exist. The first model says that there is an individual which is in all three classes, an individual in classes *A* and *B* but possibly not *C*, and possibly an individual in class *B* which is in neither of the other two classes. The second model does not include the possibility of a model in all three classes, but does include the possibility that an individual in *B* may also be in *C*.

Any conclusion to the syllogism must be valid in both mental models. The first model permits the conclusions *some of the A are C* and *some of the C are A*. However, the second model rules out both these, and in fact the syllogism has no valid conclusion.

Johnson-Laird and Bara (1984) present a review of the then current psychological theories of syllogistic inference. They report experiments which lead them to propose that the variation in performance between the syllogisms can chiefly be explained in terms of the figure of the syllogism and the number of mental models required to represent it.

Bucciarelli and Johnson-Laird (1999) developed a computer program to simulate reasoning about syllogisms using mental models. They then conducted experiments in which participants were presented with a number of syllogisms and asked to reason about those syllogisms. In the first experiment they were given pencil and paper, their use of which was videoed. The camera was focused on where the participants were writing or drawing, so that they themselves were anonymous. In the second experiment they were additionally asked to think aloud. In two other experiments they were provided with cut-out shapes which represented the terms of the syllogisms, e.g. chefs, musicians and painters, and asked to use these shapes in their reasoning tasks. They claim that when their observations were compared with the behaviour of the computer program, similar operations were being used, thus confirming their mental model theory of syllogistic reasoning. The human reasoners did use more varied search strategies. However, they did not always create all the necessary models.

3.2.6. Mental models and Euler diagrams

Euler diagrams, or Euler circles, are a logical formalism closely related to mental models. They are a means of representing sets by closed areas, usually circles. Inclusion of one set within another is shown by one circle being totally contained within another. Overlap of sets is shown by overlap of the circles. Euler diagrams are related to Venn diagrams. However, Venn diagrams show all the potential areas of overlap between sets, indicating by shading where there is no permitted overlap. As a result, Venn diagrams can be more complex than Euler diagrams. Euler diagrams have been used to reason about syllogisms. Figure 3-1 illustrates the Euler diagram representation of each of the four forms of statement in a syllogism.

Graphical representations such as Euler diagrams are logically equivalent to mental models. However, Euler diagrams represent classes, whereas mental models represent exemplars of classes. Johnson-Laird and Bara (1984) argue that the processing of Euler circles is too complex to constitute a model of human reasoning[38]. In particular, a large number of possibilities arises when the Euler circle representations of two premises are combined. They claim that the mental model representation, based on tokens to represent elements of a class, does not suffer from this drawback.

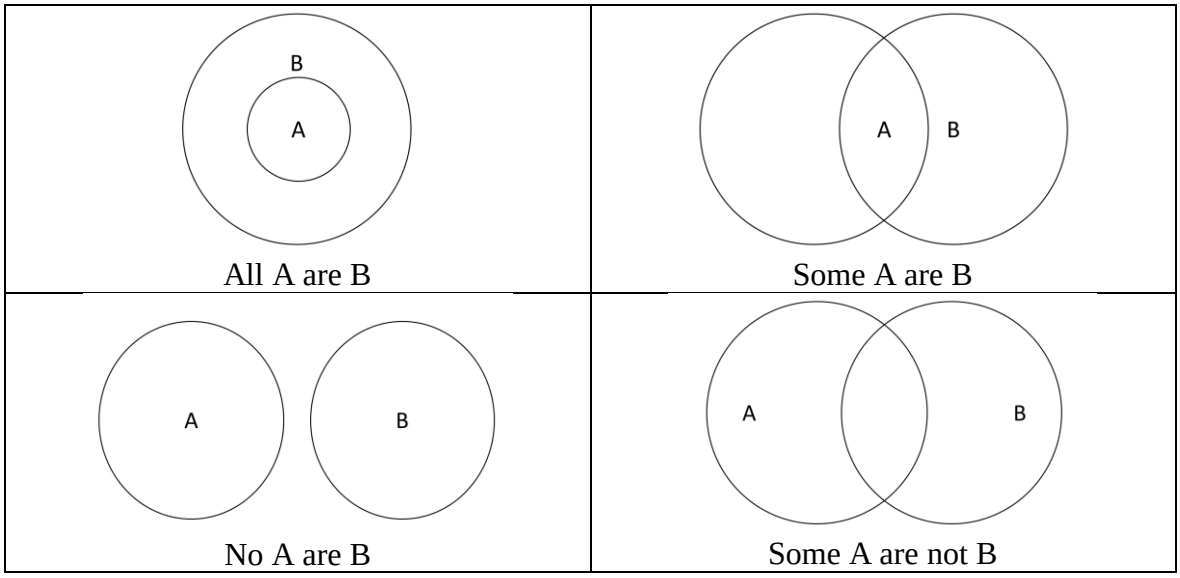

Figure 3-1 The four syllogistic forms represented as Euler diagrams
From: (Euler, 1843, pp. 260–261)

N.B. *some A are B* and *some A are not B* are geometrically identical. They are differentiated by the location of the class names.

Evidence for the efficacy of reasoning with diagrams comes from Sato and Mineshima (2016). They looked at the effect that providing Euler diagram representations had on syllogistic reasoning. They compared participants who had only a linguistic representation of the reasoning problem with those who had both a linguistic and diagrammatic representation. Some questions were based on standard first-order quantifiers, whilst other questions included the second-order quantifier, *most*. In both cases, speed and accuracy were significantly better for the group with diagrams than for the group without diagrams. An interesting aspect of this work was that it was conducted in Japanese; to this author's knowledge, most reasoning research has been conducted in European languages, in particular English.

3.2.7. Evidence from cognitive linguistics

Lakoff (1987) takes a similar view to the mental model theorists, but starting from a linguistic standpoint. His thesis is that categorization is at the heart of human reasoning, but that human categorization does not fit the classical view of categorization. The classical view is that we categorize entities by reference to a set of characteristics. However, when we examine the use of language, we find that the categorization is more complex. For

[38] The problem arises in part because, with the exception of *no A are B*, each of the four forms of premise requires to be represented in several different ways. For example, the statement *some X are Y* requires four representations: one for the case that the classes overlap but that each class has members not in the other; one for the case that the classes are exactly equal; and two for the possibilities that one class is a proper subset of the other (Johnson-Laird & Bara, 1984, p. 9). In reasoning each case needs to be considered individually.

example, we commonly used radial structures where there is a central case and variations on that case which most native speakers would recognise but which cannot be predicted. An example he gives is *mother*, which gives rise to *birth mother*, *genetic mother*, *adoptive mother*, *foster mother* etc. Consequently, human categorization does not align with the simple and well-defined set theoretic model which underlies logic. This undermines theories of reasoning based on an axiomatic approach akin to that used in mathematics. This axiomatic approach is based on the construction of axioms which can be manipulated syntactically, removed from any consideration of the original interpretation of these axioms; so that only at the end of the process is the conclusion interpreted in the real-life context. Whilst in no way critical of this approach in mathematics, where he acknowledges its considerable success, he claims that it is inconsistent with natural language and that it is natural language which we normally use for reasoning. His conclusion is that "we reason *about things in terms that are meaningful*" (Lakoff, 1987, Chapter 14, page 226). Lakoff's view on reasoning is different from that of the mental model theorists; for example, he ascribes a significant role to metaphor. However, neither does it support an approach based entirely on the rules of logic.

3.3. Relational complexity

The theory of relational complexity, as its name implies, is concerned with quantifying the complexity associated with reasoning about relations. It begins by defining the complexity of any particular relation, as explained in subsection 3.3.1 and also considers the complexity of an inference, as explained in subsection 3.3.2. Relational complexity theory and mental model theory have been compared in the particular case of the syllogisms; this is explained in section 3.3.3. In fact, both these two theories are based on consideration of working memory limitations, and subsection 3.3.4 discusses the connection between relational complexity and working memory.

3.3.1. The complexity of a relation

Relational complexity theory defines the complexity of a particular relation as the number of elements associated with that relation. The theory has its roots in the study of children's intellectual development and the observation that relations of increasing complexity can be understood with increasing age. Table 3-2, taken from Halford et al. (2007), illustrates relations of complexity one to four, and the median age of attainment.

Table 3-2 Example of relations of arity one to four; from: Halford et al. (2007)

Relational complexity	Example	Median age of attainment (yrs)
unary (1)	class membership, e.g. cat (Marcus)	1
binary (2)	larger (elephant, mouse)	1.5
ternary (3)	addition (2, 3, 5)	5
quaternary (4)	proportionality (2, 3, 6, 9)	11

Like the rules-based and mental model theories, relational complexity theory takes account of the limited capacity of working memory. This sets an upper limit on the complexity of any relation that can be completely comprehended at any one time. Halford et al. (2005), working with adults, found that there was no significant difference in accuracy or time between problems of relational complexity two and three. However, problems of relational complexity four were answered significantly less accurately and in significantly longer time than problems of relational complexity three. Problems of relational complexity five were

answered no better than chance, suggesting that for most people a relational complexity of four imposes an upper bound on what can be handled. After that, strategies such as 'conceptual chunking' and 'segmentation' are used. The former consists of combining tokens together, e.g. letters to form a word. The latter consists of breaking a task into serial steps.

3.3.2. The complexity of an inference

A key point is that we are frequently concerned with the complexity of a particular reasoning step involving relations, rather than with the complexity of a single relation. Halford et al. (1998) give the following example. The statements *John is taller than Mary* and *Mary is taller than Sue* both represent binary relations. However, when we make a transitive inference to deduce that *John is taller than Sue*, then this involves three elements and the relational complexity of the inference is said to be three. Halford and Andrews (2004) illustrate the ternary nature of transitive inference by comparing with concatenation. They argue that the transitive inference $a > b; b > c \Rightarrow a > c$ is ternary because it requires all three elements to be considered simultaneously. However, using $a > b$ and $b > c$ to form the concatenation *a, b, c* is binary because at each of the two stages used in the concatenation, only two elements need to be considered. After the first stage the result is externalised, and the human reasoner proceeds to the second only needing to consider the $b > c$ relation.

3.3.3. Relational complexity and mental models

Zielinski et al. (2010) have compared how well the relational complexity and mental model theories explain human reasoners' performance on all 64 syllogisms, based on data from their own experiments. To compute the relational complexity metric of each syllogism they counted the number of classes that needed to be considered to arrive at the correct conclusion. In doing this they take account of any potential for chunking to reduce the complexity[39]. As an initial example they give the easiest syllogisms[40]:

All X are Y
All Y are Z
⇒*All X are Z*

Zielinski et al. (2010) illustrate their argument graphically, as reproduced in figure 3-2. The first premise, shown in figure 3-2(a), tells us that there must[41] be an *X* which is a *Y*. Hence the class *XY* is shown patterned to illustrate that it must be populated. It may be the case that there is a *Y* which is not an *X*. Hence the class $\neg X\ Y$ is also shown. However, this class is not patterned because it may not necessarily be populated. The second premise, shown in figure 3-2(b) is similar. The class *YZ* is shown patterned because it must be populated, whilst the class $\neg YZ$ may be populated.

Figure 3-2(c) shows the effect of combining the premises. There are now three classes to consider: *XYZ* which must be populated and $\neg XYZ$ and $\neg X \neg YZ$ which may be populated. Thus we expect a relational complexity of three. However, Zielinski et al. (2010, page 394) argue that "the relation between $\neg X\ Y\ Z$ and $\neg X\ \neg Y\ Z$ does not need to be processed". Hence these two classes can be chunked to form, $\neg X\ (Y, \neg Y)\ Z$. This gives an effective relational complexity of two, as illustrated in figure 3-2(d).

[39] In a limited number of syllogisms they also take account of certain heuristics which may be used to reduce the complexity.
[40] In fact, this was the only syllogism which all their participants answered correctly.
[41] Note the comment from subsection 2.2.1 that in Aristotelian syllogisms, unlike in modern logic, the universal quantifier implies existence.

Similar arguments can be used to arrive at the relational complexity of each of the other 63 syllogisms.

For comparison Zielinski et al. (2010) use the number of mental models for each syllogism proposed by mental model theorists, in particular citing Bucciarelli and Johnson-Laird (1999) and Johnson-Laird and Byrne (1991).

Overall both theories performed similarly, in each case explaining approximately 80% of the variance in accuracy. This is not surprising. As Zielinski et al. (2010, page 418) point out "relational complexity theory is essentially a theory of mental models". The difference is that "it conceptualises models in a different way than have previous mental model theories of reasoning".

3.3.4. Relational complexity and working memory

Halford et al. (2010) review relational knowledge and include a discussion of the relationship to working memory. They regard working memory as "the workspace where relational representations are constructed" (Halford et al., 2010, page 1). Indeed, relational complexity theory can be seen as a formulaic way of thinking about the limitations posed by a finite working memory. As such it can only be an approximate description of these limitations.

As one example of this, relational complexity theory makes no distinction between an n-ary relation in which the ordering of elements is not significant and one in which the ordering is significant. Yet we would expect the former to pose less difficulty than the latter. Consider the following two sequences of relations, where we assume that in the relation *descendant (X, Y)*, *X* is the descendant of *Y*:

sibling(B, A); sibling(C, D); sibling(E, C); sibling(A, E)

descendant(B, A); descendant(C, D); descendant(E, C); descendant(A, E)

In the first case, consider whether *sibling(B, D)*, and in the second case whether *descendant(B, D);* the answer is yes in both cases. It may be that the second will be harder to answer, since at every stage of the reasoning we need to take account of directionality. Similarly, in the context of DL object properties, we might expect symmetric object properties (e.g. *is_sibling_of*) to pose less difficulty than object properties with directionality (e.g. *is_father_of*).

Another potential limitation of relational complexity theory is that it takes the particular relation as a given. In the example of subsection 3.3.2 an inference of relational complexity three is constructed from two premises, each using the relation *is taller than*. Where two different relations are used in an inferential step, this may cause additional load on working memory. Similarly, where a human reasoner is obliged to switch rapidly between thinking about different relations, this may cause additional difficulties. These are situations which could occur when reasoning about DL object properties.

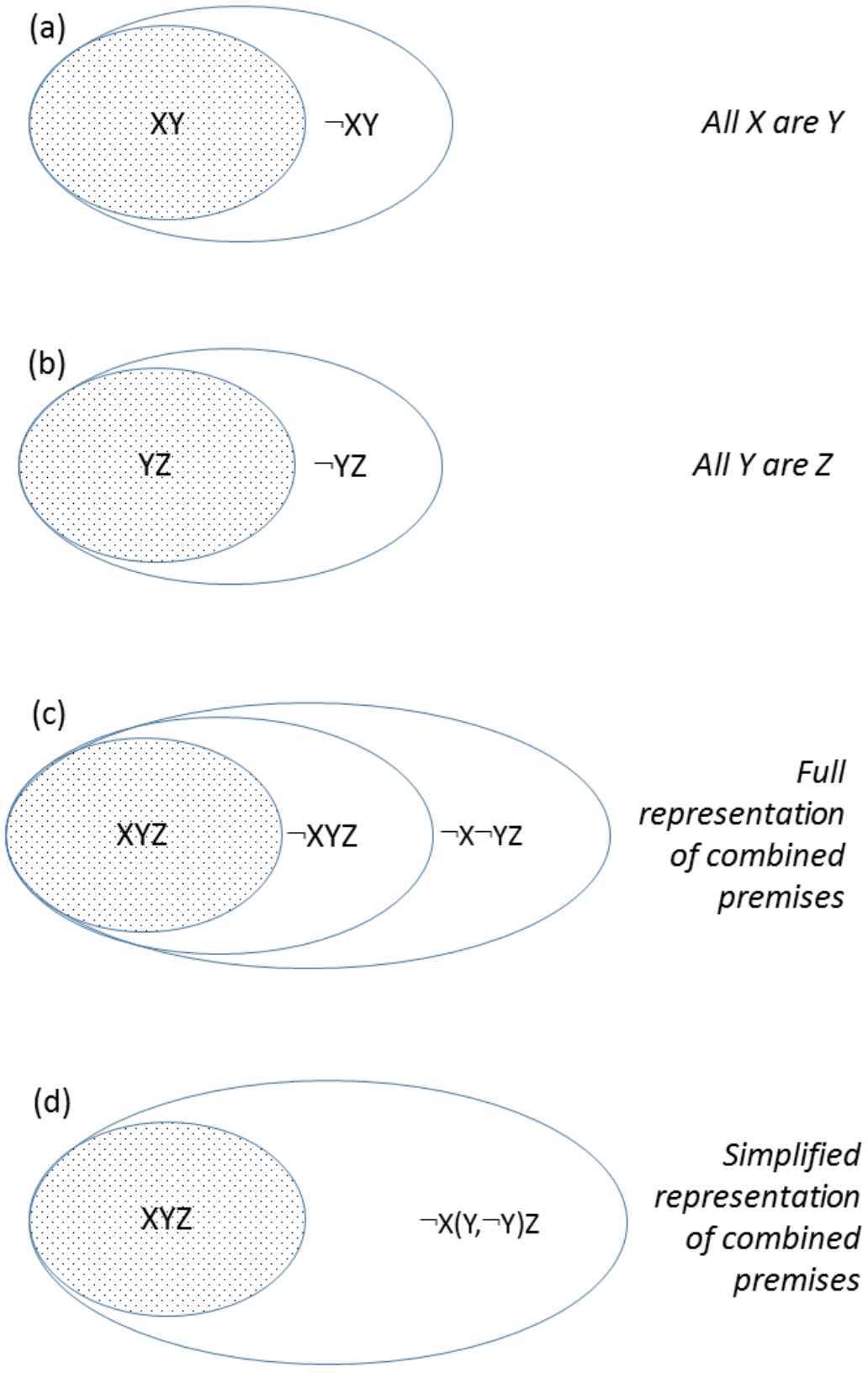

Figure 3-2 Illustrating the derivation of RC=2 for the syllogism 'all X are Y; all Y are Z ⇒ all X are Z'. The patterned areas indicates the classes which must be populated. *From Zielinski et al. (2010)*

3.4. Language – implicatures and ambiguities

As already noted, Wason (1968) posed the question whether the verbal representation of a rule influences the difficulty people experience. Grice (1975) has made the point that language may convey ideas which are not logically implied by the words used, but which nonetheless are understood by the hearer. The phrase *Gricean implicature*[42] has come to be used for ideas which are conveyed but not logically implied. Of particular relevance here are *scalar implicatures* (Carston, 1998). The underlying idea is that when we make a weaker statement than we might make, there is an implicature that the stronger statement is not true. 'Some of the students work hard' does not contradict the possibility that they all work hard. However, most people hearing this sentence would assume the existence of some less industrious students. Otherwise the speaker would have said 'all of the students work hard'. As a result, apparent mistakes in reasoning may not always be at the reasoning stage, but at the preliminary stage of interpretation. Braine (1978) makes the point that, in reasoning according to standard logic, it is necessary first to extract the "minimum commitments" from the premises, whereas in everyday discourse we make use of implicatures.

Newstead (1995) investigated the extent to which the implicature that *some* implies *some not*, and its converse, affect performance on syllogisms. This was based on a survey of previously published results and on experimental work of his own. He found that advising participants about the logical significance of *some*, i.e. that it does not imply *some not*, makes a significant difference to the effect of the implicature, compared with when no such advice is given. He also found that, when the logical meaning of *some* was made clear, these kind of Gricean errors were still common in "simple, interpretational tasks" but less common in tasks which required more reasoning. He interprets this in terms of mental model theory. He is careful to add that, whilst this particular Gricean implicature may not be important in affecting syllogism performance, other aspects of Gricean theory may well be relevant.

Another source of confusion is between *and* and *or*. 'The cars are available in red and white' and 'the cars are available in red or white' would be accepted by most listeners as reasonable representations of the same fact: that you can have a car in whichever of the two colours you like. However, the two sentences 'the Polish flag is red and white' and 'the Polish flag is red or white' have quite different meanings. The first correctly states that the flag contains both red and white areas. The second, if it means anything at all, seems to imply that the speaker is not sure of the colour of the flag but believes it to be either red or white. There is the added confusion that, to the logician, *or* is interpreted as *inclusive or*; *exclusive or* has to be explicitly stated. In everyday language, depending on context, *or* may be interpreted exclusively[43]. These varying interpretations of *and* and *or* are not just facts of everyday discourse, but pervade areas where precision might be expected. Mendonça et al. (1998) analysed the occurrence of *and* and *or* in medical terminology[44]. They found that, for *and*, 51% of the occurrences represented conjunctions, 46% represented inclusive disjunctions, and 3% represented exclusive disjunctions. The occurrences of *or* were split almost equally between inclusive and exclusive disjunctions, with 0.3% representing conjunction. Braine (1978) noted the differences between the interpretations in natural language and standard

[42] The word *implicature* was coined by Grice. It is intended to convey the idea of an assumption which a reasonable person would make but which is not justified in logic.

[43] We do sometimes, particularly for emphasis, precede the *or* with *either*. However, this is not essential.

[44] It should be noted that this was based on a version of the medical terminology system (SNOMED) prior to its conversion to a logic-based representation.

logic of *and* and *or*. He took account of those differences in developing his system of natural logic discussed in section 3.2.

Even where two statements are semantically identical, human reasoners may perform differently depending on the verbal representation. Johnson-Laird and Byrne (1989) exemplify this in a comparison of *all* and *only*. These two quantifiers are interchangeable. *All X are Y* is equivalent to *only Y are X*. Johnson-Laird and Byrne (1989) found that performance with syllogisms using *all* was better in general than using *only*, although there were differences depending on the figure of the syllogism.

Johnson-Laird and Byrne (1989) also compared *all* and *only* in the context of simple inferencing. For example, they compared the two versions of modus ponens:

All the artists are beekeepers
Lisa is an artist
Therefore Lisa is a beekeeper

and

Only the beekeepers are artists
Lisa is an artist
Therefore Lisa is a beekeeper

Participants performed less well with *only* than with *all*, although not significantly so. Johnson-Laird and Byrne (1989) also experimented with modus tollens, i.e. they replaced the second premise with *Lisa is not a beekeeper*, giving rise to the valid conclusion *Lisa is not an artist*. Here there was significantly better performance with *only* than with *all*. They discuss their results in terms of the mental model theory. However, whatever theoretical interpretation one chooses to place on these results, logically equivalent but verbally different statements can emphasize different aspects of the relationship and lead to different levels of performance.

The existence of implicatures, of ambiguities in language, and of different apparent emphases in equivalent statements, are important for the cognitive psychologists interpreting the results of reasoning experiments. They also need to be considered by the computer scientist whenever designing a notation which contains elements from normal human language.

3.5 Discussion

The study of human reasoning remains an active area of psychological research. One topic of the debate is how to interpret the experimental data; this is discussed in subsection 3.5.1. Another topic, which is less often considered but nevertheless raised by some researchers, is the relationship between conscious and unconscious reasoning; this is the theme of subsection 3.5.2. It may not be possible to develop one over-arching theory of human reasoning. It may not even be possible to create experimental situations which absolutely differentiate between rival theories. Additionally, different reasoning processes may be used by different people, or by the same person at different times. Lack of such a unifying theory is the theme of subsection 3.5.3.

3.5.1 Interpreting the data

Research into human reasoning has by now accumulated a considerable quantity of experimental data. However, this data is open to a variety of interpretations.

Much of the argument hinges on the mistakes which people apparently make in reasoning. It is possible to make a case that many of these so-called mistakes are not truly mistakes in

reasoning but arise because the participant's reasoning processes and assumptions are different from those of the logician. At the same time, those processes and assumptions may be completely valid for the individual.

Stenning and van Lambalgen (2008, chapter 5) cite studies with people from illiterate cultures who had difficulty answering simple reasoning problems, posed in everyday language. As they note, the issue seemed to be a failure to understand the 'discourse genre'. In particular, they were unfamiliar with the idea that the question should be answered purely on the basis of the information supplied. They either used their own experience to answer the question in a way which made sense in their context but was logically incorrect; or they felt unable to answer the question because of a lack of personal experience. This may not seem relevant to the studies cited here[45], or described in later chapters. However, even with highly literate subjects where the participants are told to take into account only the information provided, one cannot be sure that prior knowledge is not being made use of. This may be happening consciously or unconsciously. In either case, this could lead to incorrect answers, or to correct answers but arrived at by short-circuiting the reasoning process.

Related to this are the findings of Wason and Shapiro (1971) that people perform worse on abstract tasks than realistic ones. They compared an abstract and thematic version of the Wason selection task. The abstract version used letters and numbers on opposite sides of the cards, as in the original selection task experiment. The thematic version used the rule: *every time I go to Manchester I travel by car*. Each card contained a town (Manchester or Leeds) on one side and a mode of transport (Car or Train) on the other. Performance was significantly worse on the abstract than on the thematic task.

Citing Wilkins (1929), Wason and Shapiro (1971, p. 69) note "it is well known that concrete material is better remembered than abstract material, and that in syllogistic reasoning familiar terms inhibit fallacious inferences". There are several issues here, which will be considered again in the context of DLs. Firstly, as Wilkins (1929) notes, the use of realistic names may aid memorisation, e.g. of the rule. Secondly, as he also notes, realistic names may prevent fallacious inferences. They may also, of course, encourage correct ones. Rips (2001a)[46] supports this by showing that when faced with a logically incorrect but plausible conclusion, many people will state that it is logically correct. Rips suggests that this effect may be particularly strong when the inference is difficult. However, the effect is not likely to have been a factor in the Wason and Shapiro study. There seems no reason to suppose a greater preference for using a car rather than a train when travelling to Manchester, than when travelling to Leeds[47]. Finally, it may simply be the case that people can reason more easily with concrete problems than with abstract ones. Perhaps the concrete problem is easier to visualize than the abstract and this helps reasoning.

[45] In point of fact, Johnson-Laird and Bara (1984) do report rejecting one participant during an experiment "because she could not make any more inferences since she did not know the particular individuals referred to in the premises". The author is not aware of any other similar event being reported in such laboratory experiments. Indeed, this may have been an excuse to withdraw from a situation which the participant found stressful.

[46] Note that the "two kinds or reasoning" of the title of this paper is quite separate from other distinctions referred to in this dissertation, e.g. between rules-based and mental model based. The "two kinds of reasoning" refers instead to the distinction between reasoning deductively and reasoning inductively. By the latter, Rips means supporting a conclusion based on its plausibility.

[47] Presumably the implicit origin of the hypothetical journey was London, since the experiments were conducted with students at University College London.

Even where the genre of discourse is totally understood and no contextual knowledge can influence reasoning, there may still be problems of language. Bucciarelli and Johnson-Laird (1999), in the work studying syllogisms discussed in subsection 3.3.4, observed the use of an implicature. They report that a participant, given the statement *some of the A are not C* wrongly concluded that *some of the A are C*. This is an example of the scalar implicature, discussed in section 3.5. In the context of everyday language it is reasonable to suppose that if all of the A had not been C, then the statement would have said so. In the context of logic, this assumption cannot be made. Ford (1995) and Johnson-Laird and Bara (1984) have observed exactly the same implicature.

3.5.2 Conscious and unconscious reasoning

Another factor which may influence reasoning processes is the extent to which they are unconscious, sometimes referred to as System 1 processes, or conscious, referred to as System 2 processes. Evans (2003, page 454) notes that "System 1 processes are rapid, parallel and automatic in nature; only their final product is posted in consciousness". He uses the difference between System 1 and System 2 to explain the different results achieved with the content based and abstract versions of the Wason selection task. Indeed, Evans cites neurophysiological research which indicates that different parts of the brain are used when solving these two kinds of problems. System 1 is also seen as enabling the reasoning by plausibility which leads to the errors reported by Rips (2001a) and discussed above. This is not the same as the distinction between the mental model and rules-based theories. Evans (2003, page 457) explicitly notes that "the dual-process theory [i.e. of System 1 and System 2] does not take sides on the issue of whether this competence is achieved by the manipulation of mental models or mental rules". Johnson-Laird and Bara (1984, page 29) state of mental models that they "may take the form of vivid images or they may be largely outside conscious awareness". Johnson-Laird et al. (1989) agree that mental models may be within or outside consciousness, and maintain that "most people claim to be unaware of how they reason". They explicitly state, of the theory of mental models, that it is the structure of the model which is important, "not the subjective experience". Braine (1978) states of his system of natural language reasoning that the "schemata"[48] are not accessible to introspection, but only their products. This implies that his rules-based system should be categorised as System 1.

Sloman (1996) also considers the difference between unconscious and conscious reasoning. He describes reasoning in which the human reasoner is "conscious only of the result of the computation, not the process" as "associative" reasoning (Sloman, 1996, p. 6). He sees this associative reasoning as being based on considerations of similarity.

Oaksford and Chater (2001) suggest that their probabilistic approach, as discussed in section 3.1 is applicable to System 1 reasoning. They comment that probabilistic System 1 reasoning might be used to generate conclusions to be tested with System 2 reasoning.

Braine and O'Brien (1991) make an interesting observation about consciousness. Citing Ericsson and Simon (1984) they claim that working memory is accessible to consciousness, i.e. in the terminology used here working memory processes are System 2 process. An implication is that any working memory limitations, e.g. as proposed by the relational

[48] An inference rule schema is a formula which contains variables, e.g. *p* and *q*, which becomes an inference rule on substituting values for the variables.

complexity theory, would not apply to unconscious processes. Presumably such process would be subject to other, but perhaps more generous, constraints.

From a practical viewpoint, the extent to which reasoning is conscious or unconscious need not be of great concern. The question for this dissertation is how people reason, and in particular how they reason with DLs. The conscious or unconscious nature of the reasoning is, however, relevant from a methodological viewpoint. In particular, when people are asked about their reasoning processes, they may not be reporting the whole truth.

3.5.3 No one theory of reasoning

It can be the case that experimental results support more than one theory. For example, Ford (1995), in a study of the syllogism, argues that there is a tendency to create the conclusion to a syllogism with a form[49] used in one of the premises, even when this is incorrect. Thus performance tends to be better when the correct conclusion is of the same form as one of the premises. It happens that all the syllogisms requiring three mental models are syllogisms in which the conclusion is of a different form from either of the premises. All the two model syllogisms use a form in the conclusion which is present in a premise. This is also true for the one model syllogisms[50]. Thus, at least for the syllogism, it is difficult on the basis of the data to discriminate between the mental model theory and a theory which proposes a bias towards reuse of the form of one of the premises.

The search for one theory of reasoning may well not be satisfiable. This is illustrated by Ford (1995) who gave experiment participants reasoning tasks based on syllogisms. After the experiment, they were asked to go through the questions again, explaining with the help of pencil and paper how they arrived at their conclusions. As a result, Ford (1995) concluded that human reasoners could be split into two groups: verbal reasoners and spatial, i.e. graphical, reasoners[51]. Significantly, the relative difficulty of some of the syllogisms varied between the two groups. This difference in difficulty could be explained by considering the different strategies used by the verbal and spatial reasoners, i.e. syllogisms which are difficult using a verbal strategy are not necessarily difficult using a graphical strategy, and vice versa.

When interviewing the verbal reasoners, she notes that a few rules can render many of the apparently difficult syllogisms much easier. For example, the syllogism:

Some X are Y
All X are Z

becomes more obvious when the first premise is replaced by its equivalent *some Y are X*. The second premise then permits the substitution of *Z* for *X* in the first premise, leading to the conclusion *some Y are Z.*

When interviewing the spatial reasoners, Ford (1995) notes the construction of figures akin to Euler diagrams. As already pointed out in subsection 3.3.5, this is different from the process proposed by mental model theorists, which makes use of exemplars. It may be, of

[49] By *form* is meant one of: *all X are Y, no X are Y, some X are Y, some X are not Y*; as discussed in subsection 3.3.4.

[50] There is one possible exception to this. One syllogism reported by Johnson-Laird and Bara (1984) as a one-model syllogism has a conclusion of different form from either of the premises. This is a syllogism on which human reasoners perform badly and which was, in fact, later reclassified as a three model syllogism by Johnson-Laird and Byrne (1991).

[51] In fact, the view that human reasoners use either visual or verbal techniques goes back much further. Johnson-Laird (2004) cites Störring (1908) as having formed the same view from a study of syllogistic reasoning.

course, that when provided with pencil and paper people reason differently than in the absence of such an aid.

In a detailed theoretical and experimental study of the syllogism, Stenning and Yule (1997) argue that, at least for the case of the syllogism, the experimental evidence does not permit discrimination between the rules-based, mental model and graphical approaches. They do admit that these approaches represent "different implementations", but "of the same abstract processes" (Stenning & Yule, 1997, p. 143). They suggest that part of the learning process is searching for a representation system to use in reasoning. They argue against looking for one right system of mental representation but suggest that the representation system used depends upon the person, their level of expertise and on the particular task.

In summary, the evidence available to us does not permit the construction of one overarching theory of reasoning, and indeed the variation in how people reason makes such a theory impossible. Rather it seems that different people use different approaches. It may even be that an individual will use different approaches for different types of problems, or even experiment with different approaches to the same type of problem.

From the practical viewpoint of this dissertation, the lack of a unique theory of reasoning need not be of concern. The thesis put forward here is that the various theories of reasoning can be used in different contexts to provide insights; insights which help us understand the difficulties experienced with DLs, and also to mitigate those difficulties.

4. Methodology

We have to remember that what we observe is not nature herself, but nature exposed to our method of questioning.

Werner Heisenberg,
'Physics and Philosophy: The Revolution in Modern Science', 1958

This chapter provides an overview of the methodology used in the subsequent four chapters; a methodology which was designed to answer the research questions posed in Chapter 1.

In their most general form these were stated as:

In what way can the difficulties experienced in using Description Logics be understood in terms of underlying theories, e.g., theories of reasoning, already developed within the cognitive psychology community?

In what way could such theories contribute to improving the usability of Description Logics?

Before answering the first of these it was necessary to understand the ways in which DLs are being used. The next chapter describes a survey intended to gain a better understanding of how ontologies are being created and edited. Of particular relevance to the first research objective, the survey investigated what features of DLs were commonly used. This helped to ensure the ecological validity of the studies described in subsequent chapters.

Chapters 6 and 7 then describe two studies designed to answer the first of these research questions. In each study participants were given reasoning tasks and on the basis of their responses, and response times, information was gained about the relative difficulty of the various DL features. The theories of reasoning described in Chapter 3 were used to help explain these difficulties. Chapter 8 then addressed the second research question. Motivated by the observed difficulties, and by their explanation in terms of the theories of reasoning, changes to MOS were proposed, and the effects of those changes investigated.

In this chapter, section 4.1 discusses eithical considerations, for both the survey and the studies. Section 4.2 then provides some general background to the survey. The rest of this chapter is concerned with the three studies reported in Chapters 6 to 8. Section 4.3 describes the general format of the study questions. Section 4.4 describes the tools and techniques used for collecting and analysing the data. The next two sections discuss the techniques used to avoid two particular sources of bias; section 4.5 is concerned with avoiding bias due to participants' prior knowledge of the question subject matter and section 4.6 is concerned with avoiding bias due to question order. Section 4.7 discusses statistical significance and section 4.8 discusses qualitative techniques for gaining insights into the reasoning process. Section 4.9 discusses threats to the validity of conclusions drawn from the work reported in the subsequent chapters. Finally, section 4.10 reviews the themes of the chapter and makes a general comment about the nature of the work reported.

4.1. Ethical considerations

The nature of this work is such that no sensitive personal information was required either from the survey respondents or the study participants. Nevertheless, there were some ethical issues to be addressed. Subsection 4.1.1 discusses the treatment of confidentiality issues arising from the survey and subsection 4.1.2 discusses some broader ethical issues arising from the studies.

4.1.1. The survey – confidentiality issues

Most of the personal information sought from the survey respondents could be regarded as public domain, e.g. the sector in which the respondent worked. In any case, there were no mandatory questions; participants were free to omit any question. All reporting of the data was anonymous. In particular, published quotes from the respondents were not attributed. Respondents were asked to provide their email addresses; again, this question was non-mandatory. Respondents were also asked if they were willing to be contacted after the survey.

4.1.2. The studies – confidentiality and other ethical issues

As with the survey, all reporting was anonymised. In general, reported data comprised statistical rather than individual data. An exception to this was the reporting of participant comments. Here care was taken to avoid providing any information, e.g. gender, which might enable identification of the participant.

Participants were not subject to any intentional stress during the study. They were not required to complete the study in any particular time period. They were encouraged at the beginning not to spend too long over any question. In a few cases participants were reminded of this during the study when they seemed to be spending an excessive time on a question. This was from the practical viewpoint of avoiding excessively long study times, rather than any intention to induce stress.

Participants were free to withdraw from the study at any time. In total, one participant did do this. The stated reason for this was that the participant did not feel in possession of sufficient background knowledge to be able to answer the questions. However, the participant did not show any signs of stress during the study or in withdrawing from the study.

Each study was approved by the Open University Human Research Ethics Committee. The relevant references are: HREC/2013/1551/Warren/1; HREC/2014/1767/Warren/1; HREC/2015/2075/Warren/1.

4.2. The survey

Research into the usage of ontologies takes a number of forms. A common approach is to analyse ontologies available on the Web. This approach was adopted early in the history of the Web by Tempich and Volz (2003), who were interested in creating benchmark ontologies and for this they needed an understanding of the kinds of ontologies being used. More recently, Power and Third (2010) reported on the usage of OWL language features, Khan and Blomqvist (2010) were interested in the frequency of occurrence of particular content patterns, and Glimm et al. (2012) were interested in which OWL features were being used by the Linked Data community.

Another approach is to interview users. Vigo et al. (2014a) were interested in ontology tool design and interviewed 15 ontology authors in order to arrive at recommendations for tool improvement. A related approach is to observe users as they work. Vigo et al. (2015) followed up their previous study with a study of user behaviour based on a combination of eye-tracking data and user logs obtained by instrumenting Protégé.

Analysis of existing ontologies provides detailed insight into the kinds of ontologies being used in a way which can be automated to consider large numbers of ontologies. On the other hand, it provides limited insight into authoring processes and authors' thinking. User interviews and user observation do provide this kind of richer insight, which complements

what can be learned by analysing ontologies. However, user interviews and observation are necessarily labour-intensive, although Vigo et al. (2015) have automated the process of user log analysis.

The survey approach, used in the work described in Chapter 5, enables a far greater number of ontology users to be reached than is possible with interviews. This provides information about user processes and user thinking, although the lack of interaction with the users means that the data will not be so rich.

A problem with both interviews and surveys is to achieve an unbiased sampling. The likelihood is that the samples used will be influenced by the researchers' access to ontology users. In the survey described later, respondents were chiefly approached via mailing lists. These were mostly general mailing lists for those interested in ontology use across a wide range of application areas. Life sciences are a major application area and a mailing list for life scientists interested in semantic technologies was also used. The result was a wide range of respondents, although there may have been some bias towards life scientists.

4.3. The study questions

The research questions stated above refer to the difficulties of using DLs. Such difficulties can take various forms. Here the focus is on the difficulties people have in reasoning with DLs, e.g. when attempting to understand how a justification leads to an entailment. An alternative focus might have been on modelling with DLs. The focus on reasoning has been adopted partly because of the importance of understanding how a justification leads to an entailment in the ontology debugging process. Moreover, during the modelling process it is necessary to be able to reason with DLs, to understand the consequences of the axioms being created. However, the modelling process offers additional challenges, and is a valuable complementary area of study; for an example of such a study see Scheuermann et al. (2013).

In the studies, accuracy and response time are used as a proxy for difficulty. These are objective measures of the difficulty experienced, and are generally used in psychological research. An additional possibility would have been to have asked participants to rate the difficulty of each question once completed, e.g. on the scale: 'easy', 'moderate', 'difficult'. This would have provided a subjective measure as a complement to the two objective measures.

The questions in each study followed the same pattern. A set of DL statements were provided plus a putative inference. The participant was required to confirm or refute the validity of the inference by clicking on a button labelled 'valid' or 'not valid'. In general, in each study the questions were divided approximately equally between valid and non-valid inferences.

The structure of the questions follows a pattern common in reasoning research, as described in chapter 3. They are similar to the questions used by Rips (1983) who asked participants to indicate whether a conclusion was "necessarily true" or "not necessarily true". Other studies (e.g., Johnson-Laird & Bara, 1984) require the participants to create their own conclusions from a set of premises. This type of question is particularly used for studying syllogisms, where there are a limited number of valid conclusions[52] and each conclusion is in one of four forms. However, in each of the three studies described later there were a

[52] Of the 27 syllogisms with valid conclusions the great majority have only one valid conclusion in one of the four Aristotelian forms. However, there are a few with more than one.

variety of logical structures being investigated and this would not have lent itself to the use of open-ended questions.

4.4. Data collection and analysis

4.4.1. The data

A feature of the Rips (1983) study, and of a number of other early studies, was that it was conducted with pencil and paper, i.e. participants were required to indicate the correct answer on question sheets. This restricts the quantitative data to the accuracy of the answers. In fact, much of the discussion in the reasoning literature is based on accuracy, in part because the debate between the various schools turns on the errors which people make. From the statistical viewpoint the data is relatively limited since the responses are all binary (correct / incorrect) and sampling is assumed to be from a binomial distribution. Some studies, e.g. Johnson-Laird and Bara's (1984) study of syllogisms, use computer technology to record the response time for each question, thereby supplementing the accuracy data.

In situations where the questions are relatively easy, the accuracy data may be of little value and the timing data is the chief source of information. An example is given by Blake et al. (2012), who were interested in whether the orientation of an Euler diagram affected comprehension. The overall error rate was 3% and the authors noted that "there is little useful information that can be derived from this error data". The timing data, however, provided a significant amount of diversity.

Johnson-Laird and Bara (1984) used a measure which combined accuracy and timing. They were working with questions based on the 27 valid syllogisms. For each participant they put the participant's correct responses into the top category, the responses for which the participant had indicated a conclusion which was incorrect into the middle category, and the responses for which the participant had indicated 'no valid conclusion' into the bottom category[53]. Within each of these categories they ranked each of the responses by time. This produced an overall ranking of difficulty[54].

Where there are an appreciable number of incorrect responses one approach is to analyse the accuracy and timing data separately but to limit all statistical analysis of response time purely to the correct responses. The rationale for this is that the time taken to incorrectly respond is not a meaningful measure of the difficulty of the question. However, this may not necessarily be the case; participants are likely to take some considerable time over the more difficult questions even when they get these questions wrong. One disadvantage of not using timing data from the incorrect responses is the reduction in sample size. Moreover, this strategy may introduce a bias. If we remove the incorrect responses, as the questions become more difficult not only do the samples become smaller but in general they concentrate more on the better-performing participants, thus biasing timing comparisons between questions of different levels of difficulty. For these reasons, in the studies reported in Chapters 6 to 8, analysis of response time data has included the timings from the incorrect responses.

4.4.2. Data analysis

In analysing time data it is common to use techniques such as the t-test and analysis of variance (ANOVA) which assume an approximately normal distribution. The data should

[53] I.e. they regarded 'no valid conclusion' as a more serious error than an incorrect conclusion.

[54] They used this ranking to demonstrate that the difficulty of the 27 valid syllogisms is dependent on number of models, and for the one and three model syllogisms, on 'figure', i.e. the order in which the terms are presented in the syllogism, see Table 3.1 of Section 3.3.4 of Chapter 3.

be checked for normality prior to using such techniques, e.g. by using a normal probability plot[55]. If the data are not normal, it may be possible to transform the data such that they are normal and techniques such as the t-test and ANOVA can then be applied to the transformed data. An alternative is to use a non-parametric test. However, Hopkins et al. (2009) note that a transformation to reduce skewness followed by a parametric test provides greater statistical power at small sample sizes than does a non-parametric test. Osborne (2005) provides an overview of how data can be transformed so that its distribution is closer to the normal distribution. An example of this approach is given by Blake et al. (2012) who found it necessary to apply a log transformation to their timing data. In fact, the log transformation is one of a number of transformations on Tukey's ladder of powers[56] (Scott, 2012). The technique is to move up or down this 'ladder' until the data best approximate normality. Data can be tested for normality either by graphical inspection or the Shapiro-Wilk test (Shapiro & Wilk, 1965). More detail is provided in Chapters 6 to 8, but in general the log transformation was found most satisfactory in transforming the time data to an approximately normal distribution.

For the accuracy data a binomial distribution was assumed and Fisher's Exact Test and logistic regression (Peng et al., 2002) were used.

Note also that all tests are two-sided, unless it is explicitly stated otherwise.

4.4.3. Practical details

For the first two studies, described in chapters 6 and 7, a web-based tool was used to provide the questions and capture the responses. More details are provided in the appropriate chapters. The screen-recording tool *Camtasia*[57] was used to record the sessions and hence capture the timing information. This approach was used in laboratory experiments. In the second study, it was also used in remote sessions, using *Skype*[58] for screen-sharing. Specifically, the participant shared his or her screen with the experimenter and accessed the web-based tool which collected the responses. The experimenter ran Camtasia on his machine and recorded the session.

The approach has a number of disadvantages. Collection of the timing data after the experiment is a lengthy process. Moreover, the various web survey tools available are limited in their ability to randomize question order. The need for randomization is discussed in section 4.5 below. For these reasons the third study, described in chapter 3, used the *MediaLab* tool from *Empirisoft*[59]. This software, which was run on the experimenter's laptop computer, collects the responses and the timing data in a form which can be exported to a spreadsheet. It also enables the order of the sections within the study and of the questions within each section to be randomized.

Statistical analysis was performed using the *R* software environment[60] (R Core Team, 2014) run within *RStudio*[61].

---

[55] In a normal probability plot, normal data appears as a straight line.

[56] These include: … $-x^{-2}$, $-x^{-1}$, $-x^{-0.5}$, log x, $x^{0.5}$, x, x^2 …The negative sign before the transformations with negative index is to preserve the order of the variable, i.e. to ensure that the relations '<' and '>' are preserved by the transformation.

[57] https://www.techsmith.com/camtasia.html

[58] http://www.skype.com/en/

[59] http://www.empirisoft.com/

[60] https://www.r-project.org/

[61] https://www.rstudio.com/

4.5. Avoiding the use of prior knowledge

It is clearly essential in all reasoning experiments to ensure that the participants are actually reasoning about the questions, or at least attempting to reason, and not using prior knowledge about a subject area. There are various approaches to this, which can be regarded as being on a spectrum depending on their degree of abstraction.

In psychological research, an example of the completely abstract approach is provided by the original selection task described in Wason (1968), where letters and numbers were used. In DL research, the completely abstract approach is exemplified by Horridge et al. (2011), who gave participants a set of axioms and a putative conclusion to confirm or refute. The axioms and the conclusions made use of classes and properties; these were referred to as C1, C2, … and prop1, prop2, …

A variant on the abstract approach is illustrated by Nguyen et al. (2012, 2013), who used 'nonsense' statements. An example of a statement used by Nguyen et al. (2012) is "Everything that has a worship leader is a fomorian". As can be seen from this sentence, the approach combines made-up words, e.g. "fomorian" and real words. Some of the real words are employed with their actual meanings, e.g. in the phrase "that has". Others are used meaninglessly, e.g. the noun phrase "worship leader"[62]. The result is a set of grammatical sentences employing a combination of phrases which have meaning ("Everything that has a … is a") combined with meaningless words and phrases ("worship leader", "fomorian") such that the sentence itself is meaningless. This is useful where one wants questions to be in natural language, rather than a logical symbolism. The resultant statements do have the general flavour of nonsense language, akin to some of the writing of Edward Lear. Some participants may find this disorienting.

Zielinski et al. (2010) took an abstract approach in their study of syllogisms, using the letters *X*, *Y* and *Z*. However, they began by giving participants two practice questions. The first of these used occupations; whilst the second used the letters *A*, *B*, *C* to acquaint participants with the abstract approach. Johnson-Laird and Bara (1984), in their study of syllogisms, take an approach which does not permit the use of prior knowledge but which has a realistic flavour. They constructed questions in terms of occupations (e.g. judges), interests (e.g. ornithologists) and preoccupations (e.g. vegetarians). Bucciarelli and Johnson-Laird (1999), also studying syllogisms, take a similar approach. As already described in Chapter 3, in one experiment they provided participants with pictures to help them reason, choosing occupations and interests which were pictorially distinguishable, e.g. cooks, musicians and swimmers.

In some experiments it is, of course, required to pose realistic questions. Such was the case with the thematic questions of Wason and Shapiro (1971) discussed in chapter 3. When realistic questions are used, great care needs to be taken to ensure that there are no natural associations which might bias the reasoning process. As described in Chapter 3, Wason and Shapiro (1971) used questions about travel involving two towns and two modes of transport, and for this there needed to be no natural association between either of the towns and either of the modes of transport.

[62] It could be argued that "worship leader" does possess a meaning. However, the phrase is so unusual and shares no context with the other words used, so that it is unlikely to have any effect on the participant's reasoning processes.

The advantage of the totally abstract approach, e.g the use of C1, C2, prop1, prop2 by Horridge et al. (2011), is that one can have complete confidence that no prior knowledge is being brought to bear. However, this approach may make it unnaturally difficult to remember the significance of the various terms used. Chapter 3 has already noted Wason and Shapiro's (1971) observation, in turn based on Wilkins (1929), that "concrete material is better remembered than abstract material". In reasoning research, an argument could be made that the approach taken should depend on whether one regards the memorisation as part of the reasoning, and hence part of what is being studied, or as an extraneous activity the effect of which one wishes to minimize.

For the studies reported here the goal is to understand and mitigate the difficulties experienced in using DLs, and as such the DL statements used should be as close as possible to those met in real-life. In real-life, names used in ontologies are generally chosen to be meaningful. Ideally, for the studies, names should be meaningful, familiar to all participants and yet not provide clues as to the correct response. In practice there are difficulties in choosing names, particularly to satisfy the last of these criteria, and a hybrid approach has been adopted in the three studies reported here. All property names are meaningful, e.g. *has_child*, as are some class names, e.g. *MALE* and *FEMALE*. Abstract names were used where it was felt that meaningfulness was less important, i.e. for individual names and in some cases for class names. Even here, where necessary, names are chosen to be mnemonic. In particular, in questions in the second and third studies where hierarchical relations need to be represented, top level classes are represented by single letters (e.g. *A*, *B*), second level classes by appending numbers (e.g., *A_1*, *A_2*), and third level classes by appending letters distinct from those used for the top-level classes (e.g. *A_1_X*, *A_1_Y*).

It is likely that, when working with real-life ontologies, people do draw on prior knowledge. They may, for example, reason from plausibility as discussed in Chapter 3. They may also make use of concepts not present in the questions but which can be constructed from the available concepts. For example, after completion of study 3, a participant mentioned that, in thinking about a question which used nested instantiations of the *has_child* property, the concept *grandfather* had been used, although it was not present in the question.

A related but different issue arises for the non-valid questions. In setting these questions it is necessary to create non-valid conclusions. Some potential conclusions will be more obviously non-valid than others. The conclusions should appear sufficiently plausible to necessitate appreciable analysis of the premises. Moreover, when comparing non-valid questions, e.g. with different levels of complexity in their premises, the conclusions need to be of comparable plausibility. In practice this may cause difficulties. One solution is to only perform analysis on the valid questions, using the non-valid purely as foils. This was the approach taken by Nguyen et al. (2012). This has the disadvantage of reducing the data available for analysis and for this reason the work described in later chapters makes as much use as possible of the data from non-valid questions.

4.6. Avoiding bias due to question ordering

It is reasonable to assume that the order of questions might affect both the accuracy of response and response time. On the one hand, we might expect that performance will improve as participants get used to the type of question. On the other hand, there could be a fatigue or boredom effect leading to worsening performance. In either case, such effects will distort comparisons between questions and hence bias the results.

Each of the studies had: an introductory section, where participants were asked to provide basic information about themselves; three or four sections with the questions, as described in Section 4.3; and a final section in which participants were invited to provide feedback. Thus the order of the questions within the study will depend on the order of the question sections and the order of the questions within the question sections.

In the first study, described in Chapter 6, there were three question sections and hence six possible permutations. There were 12 participants and hence each permutation was used twice. This compensates for the large-scale effect, e.g. a question is close to the beginning of the study for at most one-third of the participants. However, there was no compensation for the effect of ordering of questions within the sections. As this study was largely exploratory, this was regarded as acceptable.

In the second study, described in Chapter 7, there were four question sections and hence 24 possible permutations. Response time data was collected from 24 participants[63], and each participant was presented with a different section order. In this case, to compensate against a bias due to question order within sections, for each section half the participants saw the questions in one order, half in the reverse order.

As noted in section 4.3, the first two studies were implemented using a web-based tool. Consequently, the permutation of section and question order was a lengthy, manual process. However, study 3 used a tool which enabled automatic randomization of question section and question order within sections, and this enabled a controlled investigation of the effect of question position. The investigation revealed an effect due to the position of a question overall and also its position within its section. More detail is provided in Chapter 8.

4.7. Statistical significance

In the three studies reported here, sample sizes were inevitably quite small. Organizing and holding study sessions is necessarily a time-consuming process and will always limit study size. However, for these studies the greatest limitation was access to participants. It was not necessary that the participants be experts in ontology modelling or in logic, since the necessary details were explained to participants prior to the study. However, they did need to have sufficient background in computer science to be able to absorb those details. Very many of the participants came from the research community at the Open University, augmented by contacts within the relevant research communities; and, in the case of study 3, members of a research group at another university and an industrial research group. In practice, difficulties of obtaining appropriate participants did impose a limitation on study size.

Sample size, effect size and statistical significance are linked. For any given size of effect, the greater the sample size, the greater the likelihood of a small p-value. Thus, any given p-value is not a function of the size of the effect alone, but of the size of the effect and the sample size. In the work reported in this dissertation, $p < 0.05$ is regarded as implying significance. Small effects are only likely to give rise to significant differences when we have a sufficiently large sample size. Ideally, we need a sample size sufficient to make it likely that any effect large enough to be of interest to us will be significant. In practice, this is not always achievable at the level of individual questions, and as a result the majority of statistical analyses in subsequent chapters are based on aggregated questions.

[63] As explained in Chapter 7, accuracy data was collected from additional participants.

Related to this is the question of repeated statistical testing. Ioannadis (2005) is concerned about the validity of much reported scientific work. Part of this concern is with repeated statistical testing with a relatively low significance threshold. He provides a detailed analysis of the danger of this practice. The essence of his argument is that if you regard $p < 0.05$ as significant, then on average one time in twenty you will report a false positive, i.e. you will have a type I error. In some situations you may have very few, or even no, actual positives, so that given enough statistical tests most, or even all, the reported positive findings will be false.

When multiple comparisons are made, e.g. repeated t-tests, one approach to compensating for this problem is to make use of the Bonferroni correction. In the simple Bonferroni correction, if the generally-used threshold is α (here 0.05), and there are n tests, then each test uses the threshold α/n. McDonald (2014) provides a brief overview of multiple comparisons, whilst Shaffer (1995) discusses multiple comparisons in more detail. Both discuss the Bonferroni correction and refinements to it.

One particular situation which potentially leads to repeated significance testing occurs after an ANOVA, when it is required to determine which pairwise differences are significant. In the case of a one factor ANOVA with *n* levels there will be $n(n-1)/2$ pairwise comparisons. To compensate for the increased likelihood of false positives, Tukey's Honest Significant Difference (HSD) test (Tukey, 1949) can be used to adjust for the number of pairwise comparisons being used. This is provided as an optional feature with the ANOVA function in the statistical language, *R*, used for the work described in this dissertation.

Both McDonald (2014) and Shaffer (1995) stress the importance of deciding which tests constitute a 'family', i.e. which tests should be regarded together as a set of multiple comparisons. Shaffer (1995) provides a real-life example where an unnecessarily large definition of the family excessively reduced the power of the tests.

In the work reported here, where independent hypotheses are tested using individual pairwise comparisons, then no correction is applied. Where an ANOVA is used, then any subsequent pairwise comparisons are made using a Tukey HSD analysis.

More generally, it is important to bear in mind the order of magnitude of the p-value. In some of the tests reported, the p-values are very small indeed, and this clearly diminishes the force of the argument put forward by Ioannadis (2005). At the same time, much of the work reported here is inevitably exploratory and should be seen in this light. In particular, where results are close to significance this indicates the need for further study.

Finally, a comment needs to be made about reporting significance levels. As already noted, throughout this dissertation $p < 0.05$ is taken as implying significance. Other values of p are regarded as being not significant. In line with recommended practices of the American Psychological Association (2010), p values are normally reported to three decimal places, except that p values less than 0.001 are reported as '$p < 0.001$'. In certain cases where comparisons between p values are being made, small p values are quoted to a greater precision.

4.8. Understanding the reasoning processes

The analysis of accuracy and response time data provides a considerable insight into the reasoning processes of the participants. In psychological research, a number of attempts have been made to complement this type of data with more qualitative information.

One approach is the use of 'thinking aloud', i.e. participants are asked to speak their thoughts out loud as they perform a task. The researcher may note the key points as they are spoken, or the participants' comments may be recorded for subsequent analysis. A potential problem with this approach is that the act of speaking out their thoughts may alter the participants' thought processes. Ericsson and Simon (1998) suggest that this is only likely to be the case if the speech is directed at another individual. Hence, if the participant is merely talking to him or herself, then there may be little disturbance to thought processes. During the initial sessions in the first study, described in Chapter 6, participants were asked to think aloud. This was subsequently dropped; in part because participants forgot as they proceeded into the study. Thinking aloud may also affect response time data. Reminding participants to think aloud would be likely to even more affect such data, besides increasing working memory demands.

Another approach used in reasoning research is to provide pencil and paper for participants to use in arriving at an answer. Participants' writing and drawings are subsequently analysed. As described in Chapter 3, this approach was used by Bucciarelli and Johnson-Laird (1999). Again, the drawback from the point of view of the work described here is the impact this might have on the response time data.

In fact, Bucciarelli and Johnson-Laird (1999) also video-recorded the hands of the participants as they used pencil and paper. This provides insight into the dynamics of their behaviour. Horridge et al. (2011) used eye-tracking data to give insight into how participants thought about a reasoning task. This has the advantage that it provides information about the dynamics of user behaviour without affecting the response time data.

Another approach, which was adopted in this work, is the use of online feedback and post-study interviews. The interviews were conducted immediately after each study, but there is still the possibility that participants may have forgotten the details of their thought processes. There is also the danger that participants may indulge in post-hoc rationalization. Nevertheless, data from post-study intervals offers a useful complement to the quantitative accuracy and timing data. Easterbrook et al. (2008), in a survey of empirical methods for software engineering research, make the point that all methods have flaws, and that combining techniques may compensate for these flaws.

4.9. Threats to validity

Purchase (Chapter 6, 2012) describes *validity*, or *accuracy*, as "the extent to which the experiment correctly addresses the specified research questions". She further characterises validity as: *internal validity*, which is concerned with "the design of the experiment", and *external validity*, which is concerned with "the generalization of the results". Another aspect of validity (Blake et al., 2014) is *construct validity*, which "examines whether the independent and dependent variables yield an accurate measure to test our hypotheses". Each of these three aspects is considered briefly here and in more specific detail in each of the following four chapters.

4.9.1. Construct validity

For the survey in Chapter 5, the major threats to construct validity are the interpretation of the questions by the respondents and the interpretation of the responses by the researcher. To minimise the former, the questions were checked by two reviewers. To minimise the latter, as far as possible the responses were selected from a list, rather than being free format.

For the studies, in Chapters 6 to 8, the major threats to construct validity are the accurate recording of response and response time. In all three studies, the responses were recorded automatically. The technique for recording response times varied between the first two studies and the last; this is described in the appropriate chapters.

4.9.2. Internal validity

For the survey, the major threat to internal validity lies in the design of the questions. In particular, the range of possible answers needs to be comprehensive, in order to capture all the possibilities.

For the studies, the major problem is the effect of learning or fatigue, as the participant moves through the study. Each of the studies was divided into sections, hence these effects can be regarded as inter-section and intra-section. This is discussed in detail later.

4.9.3. External validity

For both the survey and the studies, threats to external validity come from two sources. Fistly, it is necessary to create appropriate questions and thereby gain relevant data. Secondly, it is necessary to ensure a representative sample of survey respondents or study participants. Again, the extent to which this was achieved is discussed in more detail later.

4.10. Discussion

As far as practically possible the work described here has attempted to compensate for sources of bias. The survey was distributed to a very wide audience, and responses were received from a wide range of disciplines, as reported in Chapter 5. The studies were designed to avoid bias due to prior knowledge and compensate for bias due to question order. Rigorous statistical testing was used in the analysis of the quantitative data, and this data was also complemented by qualitative data gained from online feedback and post-study interviews. However, the studies reported here should be seen as exploratory. Chapter 2 has described a limited amount of previous research into DL comprehension. Very little of this has been informed by insights from cognition or language studies. The work reported here has taken a broad view of how those insights might be applied. Certain of the results are clear-cut. However, it is in the nature of an exploratory study that some results will be more tentative.

5. The user experience

That all our knowledge begins with experience there can be no doubt.

Immanuel Kant, 'Critique of Pure Reason', Introduction, 1781

This chapter discusses the experience of using ontologies based on a survey of ontology users and a follow-up survey relating specifically to ontology patterns. In particular, it concentrates on the use of DLs as a background to the work to be reported in the next three chapters. More detail of the initial survey findings, including relating to other aspects of ontology use, are reported in Warren (2013). The survey identified commonly used DL features which were then included in the study described in Chapter 6, helping to ensure ecological validity for that study. The survey also identified that there was appreciable use of ontology patterns and a follow-up survey was undertaken to obtain more detail on this, e.g. the typical sizes of ontology patterns. The follow-up ontology pattern survey is reported in Warren (2014). Warren et al. (2014) provide an overview bringing together both surveys, with additional analysis.

Section 5.1 provides some information about the organisation of the surveys and about the respondents. Sections 5.2 to 5.7 report on findings from the first survey. Section 5.2 reports on the commonly used ontologies and section 5.3 on the commonly used ontology tools. Section 5.4 discusses the usage of ontology languages, including the various OWL profiles, whilst section 5.5 discusses the usage of specific DL language features. Section 5.6 describes some comments made by respondents about their experiences with ontology languages. Section 5.7 discusses ontology visualization, which can be viewed either as an alternative or a complement to representations using an ontology language. Section 5.8, which draws on both the original survey and the follow-on patterns survey, provides information about the use of ontology patterns and pattern libraries. This is important, in part because some ontology patterns are designed to overcome limitations in the ontology language. Ontology patterns are also important from the viewpoint of this dissertation because three ontology patterns, identified as commonly used by the work of Khan and Blomqvist (2010), form the basis of the questions used in the first study. This further helped to achieve ecological validity for that study. Section 5.9 discusses possible threats to validity. Finally, section 5.10 draws some conclusions.

5.1. The surveys and survey respondents

The original survey was conducted during the first three months of 2013 using the *Survey Expression*[65] tool. Responses were obtained using a number of contacts and relevant mailing lists. The latter included: the ontolog-forum[66]; the U.K. Ontology Network[67]; the Semantic Web for Life Sciences group and the Description Logic group, both on LinkedIn; lists maintained by the Open Knowledge Foundation[68], and the internal mailing list within the author's university department. Note that this does not necessarily constitute a random sampling of application areas. In particular, the only application-specific mailing list used was for the life sciences and this may have resulted in some bias towards the life sciences.

[65] www.surveyexpression.com

[66] This forum now uses Google Groups at https://groups.google.com/forum/#!forum/ontolog-forum. Historical archives are available at http://ontolog.cim3.net/.

[67] https://conferences.ncl.ac.uk/ukon2016/, for mailing list see also: https://groups.google.com/forum/#!forum/ontology-uk/join

[68] okfn-{en,Scotland,nl}@lists.okfn.org

The follow-on patterns survey was conducted during the second half of 2013, also using the *Survey Expression* tool and using a similar set of mailing lists. In addition, certain respondents to the first survey who indicated usage of ontology patterns were specifically asked to complete the second survey.

Subsection 5.1.1 describes the original survey questions and the respondents. Subsection 5.1.2 similarly describes the patterns survey and the respondents.

5.1.1. Ontology user survey

The questionnaire contained six sections:

- *You and your work* – information about the respondent and the work area.
- *Your ontologies* – information about the respondent's ontologies and how the respondent uses ontologies.
- *Tools and visualization* – which ontology tools are used, including visualization tools.
- *Ontology languages* – which ontology languages are used, and for those using DL languages, which language features are used; also for those using OWL, which OWL profiles are used.
- *Ontology patterns* – whether taken from a library or created by the respondent or respondent's colleagues.
- *Conclusions* – an opportunity for the respondents to make any final comments on their experience using ontologies.

There were 118 responses in all, although in general respondents did not answer all questions. All 118 respondents described their primary application area, giving the following breakdown: biomedical (31%); business (9%); engineering (19%); physical sciences (7%); social sciences (5%); and other (30%). The 'other' category included 12% (i.e. 12% of the original 118) with responses which could be classified as 'computing', indicating that this would have been a useful category to include in the questionnaire. A small number cited 'cultural heritage' and 'linguistics', indicating that, e.g. humanities, might also have been a useful category. This would have constituted 4% of the responses. Other responses included publishing, agricultural production and environment. Some of the respondents used 'other' to include a combination of categories or to specify particular subdisciplines within the predefined categories.

116 respondents described their sector, giving the following breakdown: academic (45%); from research institutes (25%); industrial (17%) and other (13%). 115 respondents provided information about the length of time they had worked with ontologies. 54% had over five years' experience and only 5% had less than one year. Figure 5-1 shows the response to this question, showing also the breakdown into application areas. As already noted, the distribution of application areas does not necessarily represent a random sampling of all ontology users.

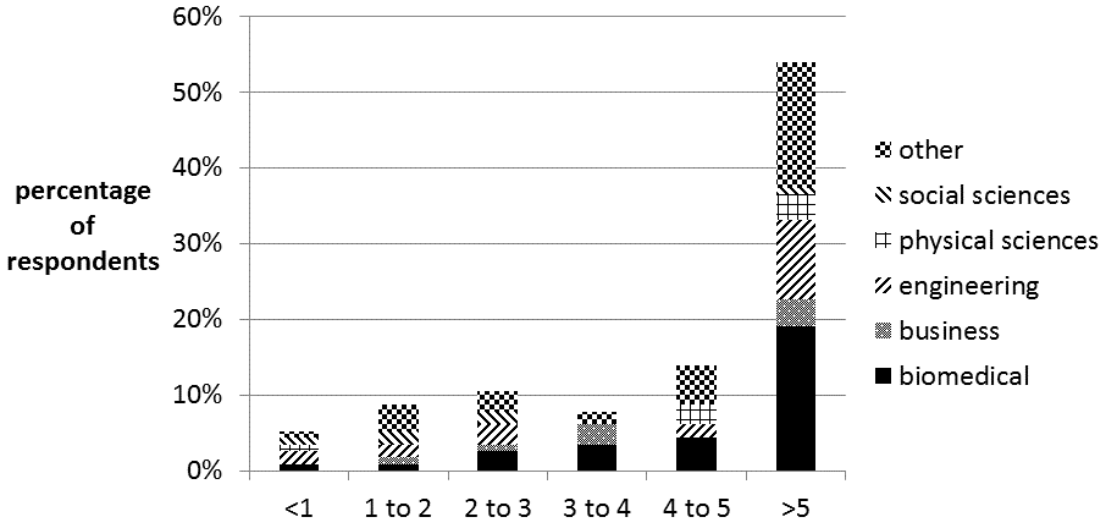

Figure 5-1 Breakdown of respondents by number of years working with ontologies

5.1.2. Ontology patterns survey

The questionnaire contained an introductory section describing the objectives of the survey, followed by four question sections:

- *You and your work* – information about the respondent and the work area.
- *Pattern use* – information about the patterns, e.g. the size of patterns being used and from where they are obtained.
- *Creating patterns* – information about how the need for a pattern is identified, how patterns are created, and how they are stored.
- *Conclusions* – an opportunity for respondents to make any final comments about their pattern use.

There were 13 respondents to this survey. As with the original survey, respondents were asked to indicate their application area. The same categories were used as before, with the addition of computing and humanities. There were five respondents from computing, three from engineering, two each from physical sciences and biomedical, and one from humanities. There were no respondents from business and social sciences.

Respondents were also asked to describe their sector, using the same categories as before. There were six responses from academia, five from research institutes and two from industry.

5.2. Ontology user survey: ontologies

Respondents were asked to specify their most frequently used ontologies, in order of frequency of use, up to a maximum of five. Of the 69 respondents who answered this question, 32 listed five ontologies and only 7 listed just one. Figure 5-2 shows the most popular five ontologies, showing the percentage of respondents who cited each ontology. Ten people indicated that they used their own ontologies, putting 'own' in sixth place, as shown. Note that in this figure, as multiple responses were permitted, the total of the percentages exceeds 100%; this applies to many other figures in this chapter. No respondent placed FOAF in their most frequently used category; the majority of respondents citing FOAF placed it in their second most frequently used category. All the other categories in Figure 5-2 were cited by some respondents as their most frequently used ontology.

Besides the ontologies shown in the figure there were a number relating to biology, medicine and chemistry, and also some generic ontologies. Amongst the latter were the W3C provenance ontology; the RDF data cube vocabulary; schema.org; upper level ontologies, e.g. DOLCE; and lexical databases, e.g. WordNet.

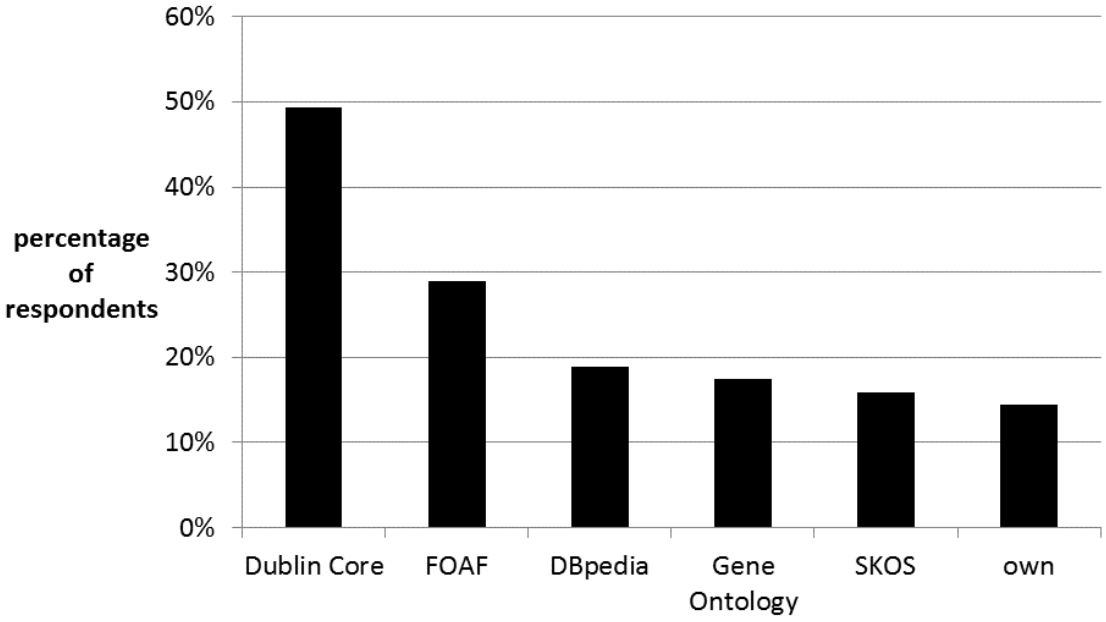

Figure 5-2 Usage of ontologies; percentage of 69 respondents

From the standpoint of the work described in later chapters, the relevant question is to what extent these ontologies make use of DLs. The most commonly used, Dublin Core, is defined in the RDF Schema Language (RDFS). FOAF and SKOS use an extended form of RDFS, see Allemang and Hendler (2011, Chapters 9 and 10)[69]. DBpedia is composed of RDF triples extracted from structured Wikipedia data (Bizer et al., 2009). The Gene Ontology (GO) is defined in a purpose-built language, but conversion into OWL DL is possible.

To generalize, the commonly used ontologies are not created in a general-purpose DL language, although conversion to OWL may be possible. Moreover, they are quite limited in the language constructs used. However, as the next section demonstrates, there is currently an appreciable usage of OWL editors.

5.3. Ontology user survey: ontology editors

Respondents were asked which ontology editors they used, from a choice of twelve plus an 'other' option. There were 65 respondents and multiple responses were permitted. Figure 5-3 shows all the tools for which there was more than one response. All the tools shown were among the predefined categories, except for OBO-Edit[70] and Neurolex, which are both editors for biomedical ontologies. Note that users of Protégé 3 were obliged to indicate whether they used the OWL or frame-based versions. They could, of course, indicate usage of both.

The dominance of DL, and specifically OWL, editors is striking. Considering the editors in decreasing order of usage, i.e. from the left in Figure 5-3, Protégé 4 is an OWL editor, as is

[69] The W3C comments that the FOAF ontology is "essentially compatible with OWL RL" except that it makes use of inverse functional datatype properties. See: https://www.w3.org/wiki/Good_Ontologies
[70] http://oboedit.org/

TopBraid Composer[71] and Neon[72]. On the other hand, Cmap[73], from the Florida Institute for Human and Machine Cognition (IHMC), is an editor for concept maps. Concept maps were originally developed to represent children's conceptual understanding but now have a wider range of knowledge management applications (Sánchez et al., 2010). Turning to the editors at the far right of the figure, SWOOP[74] is an OWL editor (Kalyanpur et al., 2006); Semantic MediaWiki is a wiki which enables the creation of informal semantic relations (Krötzsch et al., 2006)[75]; and OBO-Edit supports the Open Biomedical Ontologies (OBO) format, but also supports OWL.

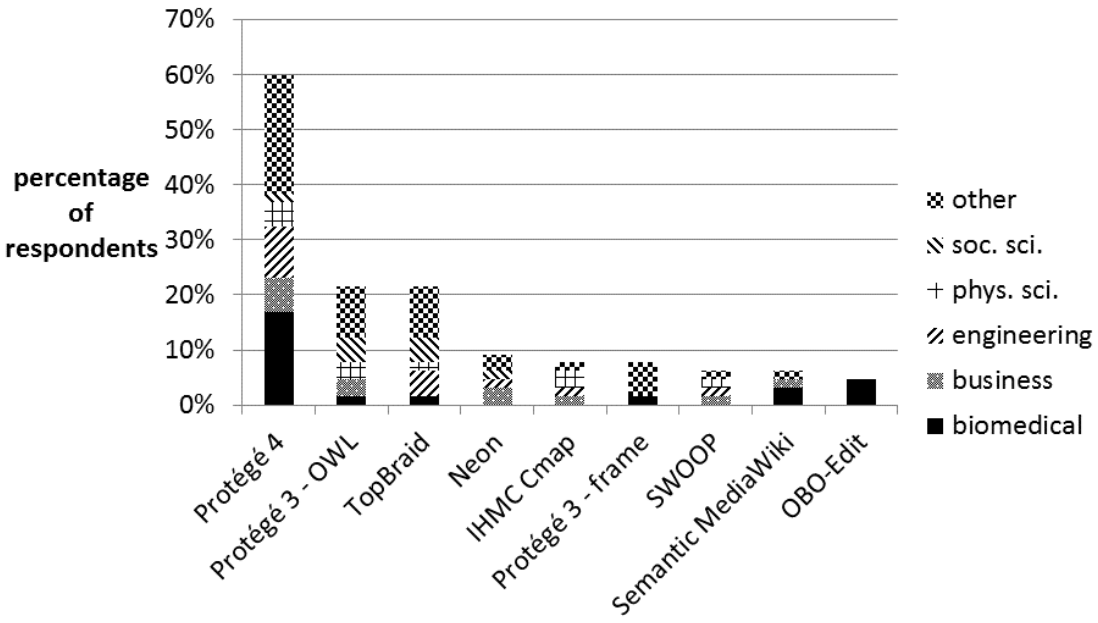

Figure 5-3 Usage of ontology editors; percentage of 65 respondents

Thus five of the nine most commonly used ontology editors are primarily OWL editors, including the four most commonly used, whilst one of the remaining four (OBO-Edit) supports OWL. Of course, the prevalence of OWL editors does not imply that people are using OWL features to any great extent. The editors may be being used for purely RDFS functionality, e.g. subsumption. The prevalence of OWL editors may reflect the fact that in recent years this is where most development has been concentrated, as a result of the standardisation of OWL by the W3C. To understand the extent to which the more sophisticated features of OWL are being used we need to understand which OWL profiles and which specific DL features are being used. These are the subjects of the next two sections.

5.4. Ontology user survey: ontology languages

Of the 65 respondents to a question about which languages they used, 58 indicated OWL, 56 RDF and 45 RDFS. The other two predefined options, OIL and DAML+OIL received no responses. 11 respondents indicated 'other' which included the OBO format, query languages, plus other more specialist languages.

[71] http://www.topquadrant.com/tools/modeling-topbraid-composer-standard-edition/
[72] http://neon-toolkit.org/wiki/Main_Page.html
[73] http://cmap.ihmc.us/
[74] https://github.com/ronwalf/swoop
[75] Two of the four respondents who indicated use of Semanic MediaWiki were using it with Neurolex (http://neurolex.org/wiki/Main_Page), a lexicon for neuroscientific information (Larson & Martone, 2013).

When asked which OWL profiles were used there were 133 responses from 54 respondents. Figure 5-4 shows the distribution of responses. The three profiles on the left are the first generation OWL profiles, the remainder are from the second generation of OWL. The relatively high number of responses for OWL Full and OWL 2 Full is surprising, given that these languages are not decidable. This may have resulted from a confusion about terminology or may represent people who did not consciously restrict the features they used, rather than actually using features beyond OWL DL or OWL 2 DL. On the other hand, these responses may represent applications which do not use reasoning, and hence are not concerned about the lack of decidability. Note also that the breakdown by application domain shows no obvious bias of a particular application domain towards a particular language profile.

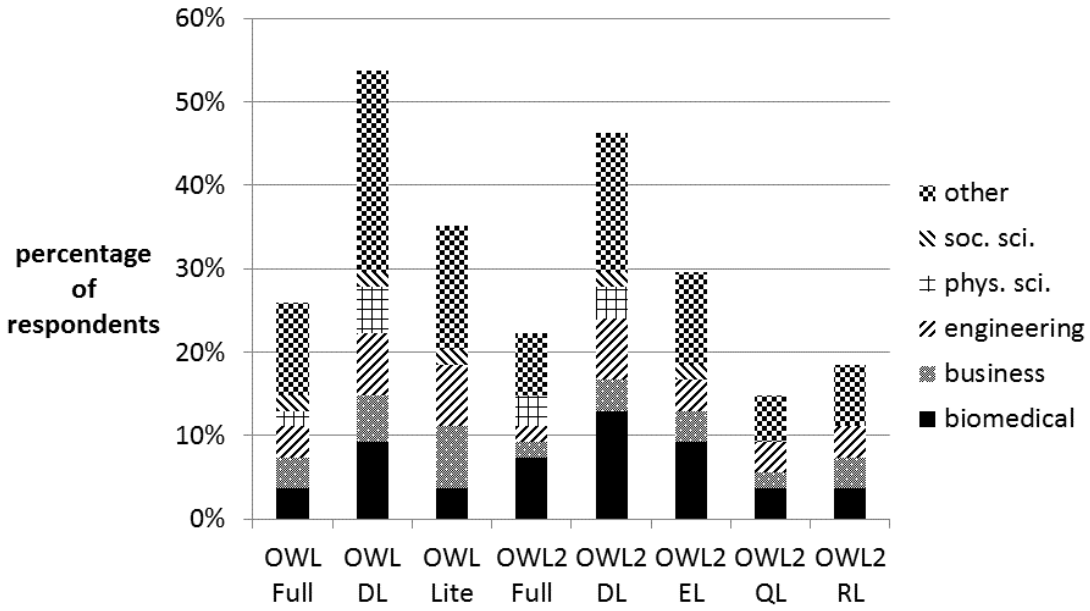

Figure 5-4 Usage of OWL profiles; percentage of 54 respondents

There were approximately 2.5 responses per respondent, indicating considerable usage of multiple profiles. This relatively high ratio may in part have been caused by people migrating from the first to second generation of OWL. Indeed, of the 32 respondents who indicated usage of more than one profile, there were 18 respondents using profiles from both generations. However, many of these 18 were using more than one profile from one or both generations; and in any case this still left 14 respondents using more than one profile entirely from one generation[76].

5.5. Ontology user survey: DL language features

Respondents using languages based on DLs were asked which language features they used from a list of 23 features. These features were identified from the relevant W3C sources, and the features available in Protégé. In hindsight, there a few omissions, in particular usage of class and object properties; although datatype properties were included. There were 47 respondents to this question. Table 5-1 shows the percentage of respondents citing each

[76] In fact, since the 18 respondents with responses in both generations represent exactly one-third of the total respondents, if multiple usage were caused entirely by migration to the second generation, this would lead to a ratio of 1.33 responses per respondent, well below the actual figure of 2.5. Note, though, that there were two who indicated all eight profiles, at least one of whom was a researcher in the area of the semantic web. No other respondent indicated more than five profiles.

language feature. The features are numbered from most commonly used (1: object property domain) to least commonly used (23: irreflexive object properties).

Table 5-1 Usage of DL features; percentage of 47 respondents

no.	feature	%age		no.	feature	%age
1	object property domain	79%		13	hasValue restrictions	60%
2	object property range	77%		14	cardinality restrictions	51%
3	disjoint classes	74%		15	symmetric object properties	51%
4	datatype properties	72%		16	functional datatype properties	51%
5	intersection of classes	70%		17	datatype subproperties	49%
6	transitive object properties	68%		18	complement of a class	47%
7	object subproperties	68%		19	qualified cardinality restrictions	43%
8	union of classes	66%		20	inverse functional object properties	36%
9	existential restrictions	66%		21	reflexive object properties	30%
10	inverse object properties	64%		22	asymmetric object properties	26%
11	functional object properties	64%		23	irreflexive object properties	17%
12	universal restrictions	60%				

Of the OWL property characteristics, there are four which over 50% of the respondents used: transitive object properties (68%); functional object properties (64%); symmetric object properties (51%); functional datatype properties (51%). The remaining four property characteristics constitute the four least commonly used features: inverse functional object properties (36%); reflexive object properties (30%); asymmetric object properties (26%); and irreflexive object properties (17%). Note that the last two of these, the asymmetric and irreflexive characteristics, were not present in the first generation of OWL. The work of the next three chapters has looked at human reasoning with transitive, functional, symmetric and inverse functional object properties, but has ignored the three least commonly used characteristics: reflexivity, asymmetry, and irreflexivity.

It is striking that even the less commonly used features were being used by an appreciable number of respondents. Even the least commonly used feature (irreflexive object properties) was used by around one in six of the question respondents. As a caveat it should be pointed out that respondents to this question only constituted 40% of the total respondents. Presumably these were the respondents with the more sophisticated knowledge of DL, who fully understood the question and were more likely to make use of a broad range of DL features. Hence it is possible that these percentages overestimate the usage of at least the less commonly used features. It should also be borne in mind that these percentages represent the proportion of respondents who use a particular feature, but give no indication of the extent to which that feature is used.

It might be thought that it is the same few people who are using all the less commonly used features. That this is not the case can be seen from Figure 5-4. Here the language features are numbered as in Table 5-1, from most commonly to least commonly used. The dashed line shows the percentage of question respondents indicating a particular feature, i.e. this is simply a graphical representation of the data shown in Table 5-1. The continuous line shows the percentage of respondents who used the particular feature or one of the subsequent ones (i.e. higher-numbered ones, or less commonly used ones). By definition this line must indicate 100% for language feature 1, since all question respondents use at least one

feature[77]. Also, by definition the two lines must coincide for the least commonly used feature, i.e. feature number 23.

It is striking that the continuous line takes high values for so much of the figure. The 'knee' does not occur until feature 17, where the value is 81%, indicating that 81% of the question respondents use at least one of the seven least commonly used features, i.e. features 17-23. If the less commonly used features were being used by the same small group of respondents, then the two lines would come close together, or even coincide, earlier as we move towards the right of the figure. In summary, it is not that only a small number of respondents use the less commonly used features but rather that very many respondents use at least one of these less commonly used features.

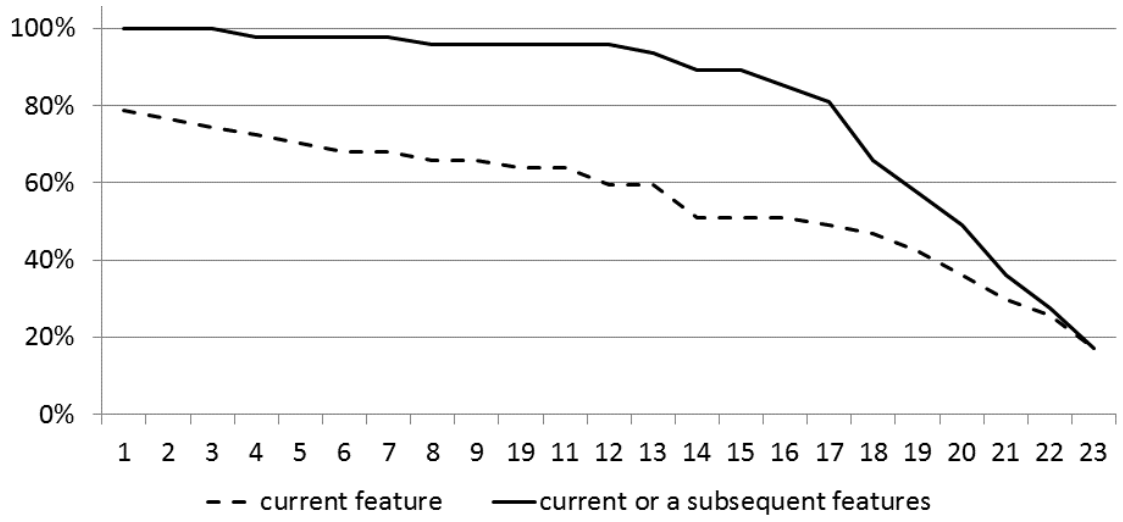

Figure 5-4 Use of DL language features; percentage of 47 respondents

5.6. Ontology user survey: ontology languages – respondents' comments

Respondents were asked for suggestions for additional ontology language features. One respondent wanted support for "acceptable" punning[78] and for the qualification pattern, as used in PROV-O[79]. Presumably what the respondent wanted was for these features to be embedded into the language. The same respondent commented on the difficulty of design decisions relating to classes vs. individuals and classes vs. properties. Another respondent asked for "easier ways to integrate and reuse vocabularies". Other requests for additional functionality related to modalities, time, arithmetic functionality and more sophisticated

[77] I.e. not specifying any language feature was taken as not responding to the question. In retrospect, it would have been useful to enable respondents to indicate that they do use DLs, but not any of the specified features. Additionally, it would have been useful to have had an option to enable all respondents to indicate any features used additional to the listed features. The most obvious omission from the list in question 17 is class subsumption.

[78] 'Punning' refers to the use of the same name to describe both a class and an individual. This is not permitted in OWL 1 DL but is permitted in OWL 2 DL. The use of the same name has no significance to the reasoner, only to the person undertaking the modelling. See https://www.w3.org/TR/owl2-new-features/, specifically section 2.4.1: https://www.w3.org/TR/owl2-new-features/#F12:_Punning The example given there is that of using *Eagle* to represent the class of all eagles and also an individual in the class of all animal and plant species.

[79] The qualification pattern uses a class to qualify an object property, i.e to associate additional information with an instantiation of the property, thereby overcoming the limitation of RDF to binary relations. It is similar to an N-Ary relation; for a discussion of both, see Dodds and Davis (2012). The PROV-O ontology is used "to represent and interchange provenance information". For a description of PROV-O, see https://www.w3.org/TR/prov-o/; for a discussion of the use of the qualification pattern in PROV-O, see specifically: https://www.w3.org/TR/prov-o/#description-qualified-terms.

property chain axioms. Additional functionality comes with a penalty, of course; two respondents noted that additional functionality can lead to problems of scalability.

Other comments related more generally to the use of DLs. One respondent wanted lightweight approaches to overcome the difficulty of characterising all information in a strongly semantic fashion; an example cited was the interpretation of 'temperature' as a physical phenomenon or a measurement depending on the system. This is related to another respondent's comment: "The rigor of the languages exceeds the rigor of the typical user by a wide margin". On the other hand, the rigour is required to permit meaningful reasoning; possibly some respondents did not make use of reasoning. Two respondents commented on the open world assumption; one noted the difficulty of grasping open world reasoning for those used to closed world reasoning; the other asked for partial support for closed world reasoning. Possible ways to satisfy this requirement might be syntax to embed the closure axiom (Horridge, 2011, section 4.13.1) into the language, or the ability to declare an entity to be distinct, without using, e.g. the *DisjointWith* or *DifferentFrom* statements. Another respondent noted the lack of discussion, in the survey, of "the issues of overlapping ontologies", quoting the OBO Foundry[80] ontologies as a collaborative set "that are open and can share terms".

One respondent stepped back from the specifics of ontology languages to make a general comment about ontology modelling and a call for better tools: "meta models … may not suit what you're doing … The complexity of it all is way beyond what we can hope to hold in our minds at any given time, but I have yet to use a tool that makes this complexity easily understood, or even easily workable with".

5.7. Ontology user survey: ontology visualization

The importance of ontology visualization is witnessed by the fact that a wide range of visualization tools has existed for some time, e.g. see Katifori et al. (2007). The number of visualization plug-ins available for the Protégé editor indicates that visualization is seen as an important support for the ontology development process. Indeed, the representation of ontologies visually is an alternative, or a complement, to their representation in terms of language.

In the survey, respondents were asked which visualization tools they used, from a choice of ten tools, plus an 'other' option. 47 respondents replied to the question, with multiple responses permitted. Figure 5-5 shows the visualization tools for which there was more than one response. TopBraid and OBO-Edit were amongst the 'other' option. The key point to note here is that these are chiefly tools for viewing the overall structure of an ontology, or part of an ontology. Some of the tools do permit properties to be displayed. However, whilst the display of overall structure supports reasoning about subsumption, none of these tools support reasoning about properties. Visual support for general DL reasoning is being developed, see the description of concept diagrams in Howse et al. (2011) and Stapleton et al. (2013). However, no generally available tools yet exist to offer visual support for reasoning.

[80] http://www.obofoundry.org/

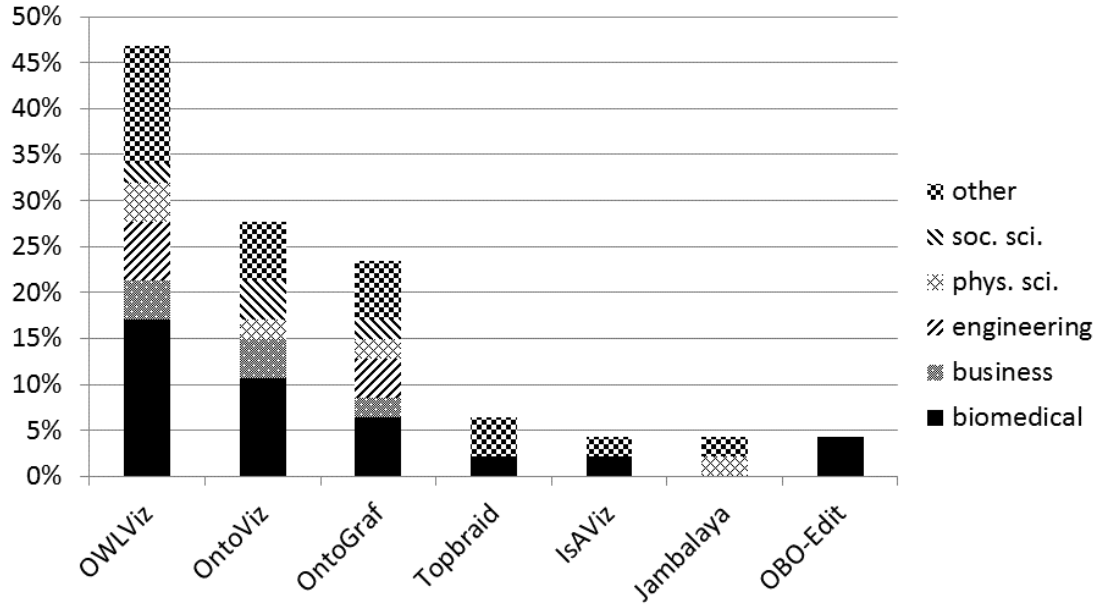

Figure 5-5 Usage of visualization tools; percentage of 47 respondents

Figure 5-6 shows the response to a question about how useful respondents find visualization. The dominant categories were 'useful to a small extent' and 'quite useful' - just over 60% of the responses were in these two categories. The categories 'very useful' and 'essential' accounted for around 34% of the respondents. This was from a sample of 56 respondents to the question, i.e. less than half the overall survey respondents. It is likely that many of those who did not respond to this question do not use ontology visualization tools, possibly because they do not find visualization useful, biasing the sample towards those who do find it useful.

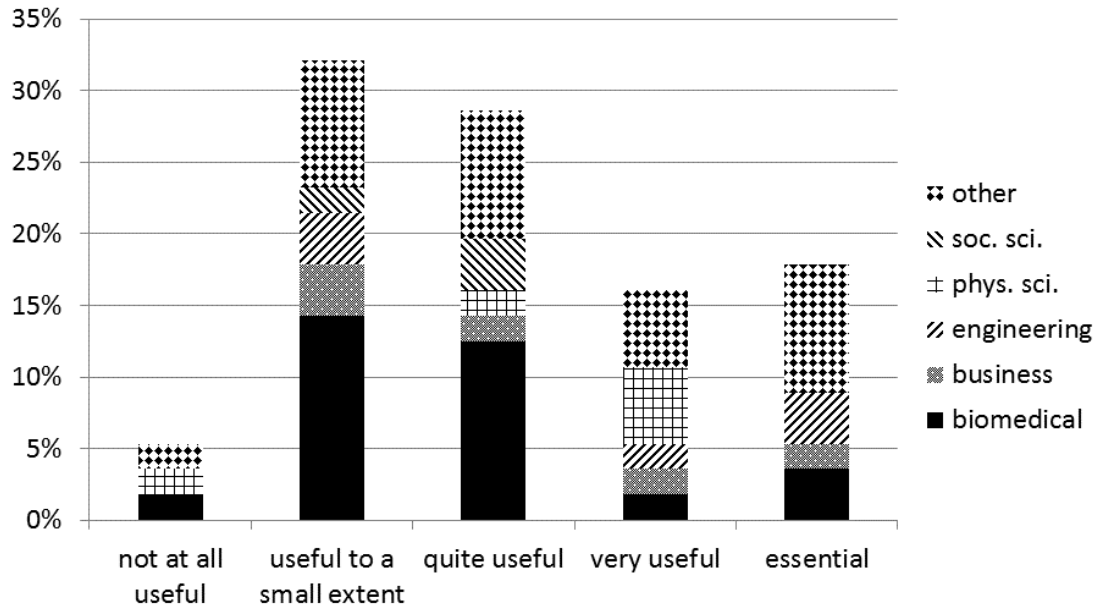

Figure 5-6 usefulness of visualization; percentage of 56 respondents

In response to a general question about the strengths and weaknesses of ontology development tools there were a few references to visualization. One respondent commented that "visualization is, especially for end-user, really hard and not task-specific". Another

wanted "proper visual editing features"; presumably so that changes could be made in the visualization.

In summary, ontology visualization was found useful to at least a small extent by 95% of the respondents, whilst many found it essential. However, given the tools available, its use was to understand the structure of the ontology rather than to support reasoning.

5.8. Ontology patterns

There are a variety of different types of ontology patterns. However, in the context of this work two are relevant: *logical patterns*, which are essentially language extensions; and *content patterns*, which are essentially mini-ontologies. Both types are explained in more detail in subsection 5.8.1. Content patterns are extensively used in ontologies (Khan & Blomqvist, 2010). To help achieve ecological validity, three of the most commonly occurring content patterns were used as the basis for the questions in the first study, as described in Chapter 6.

One concern was to know to what extent people were using pattern libraries. Subsection 5.8.1 describes the two main pattern libraries which have been developed, and also describes briefly how patterns can be categorised. Section 5.8.2, drawing on data from the ontology user survey, then describes how respondents were sourcing their patterns. Section 5.8.3, drawing on data from the ontology patterns survey, provides some information on the kinds of patterns being used.

5.8.1. Pattern libraries

The two main pattern libraries are the *Ontology Design Patterns (ODPs) Public Catalogue*[81], which was developed mainly by researchers at the University of Manchester, and the *Ontology Design Patterns.org*[82] portal, which was developed as part of the NeOn project[83]. The former contains a small number of generic patterns. The latter contains rather more patterns, some generic and some domain-specific.

The *Ontology Design Patterns.org* portal provides a classification into: structural, content, reasoning, presentation and lexico-syntactic patterns, and then a further subdivision of the structural, correspondence and presentation patterns. Falbo et al. (2013) discuss the classification of patterns, describing the classification at *Ontology Design Patterns.org* and relating this classification to the three phases of ontology development, i.e. the conceptual modelling, design and implementation phases. From the standpoint of our interest in DLs, the two most relevant categories are:

- *logical patterns* (a subcategory of structural patterns) As already observed, these can be regarded as language extensions. They do not contain class or property names, with the possible exception of generic entities such owl:Thing (⊤). A logical pattern is analogous to a macro. It may contain prototypical classes, to be instantiated when the pattern is used. An example is the 'N-Ary Relation' Pattern, designed to overcome the limitation of OWL to binary properties, and which is present in both the *Ontology Design Patterns (ODPs) Public Catalogue* and the *Ontology Design Patterns.org* portal. Another example, from the *Ontology Design Patterns.org* portal is the 'negative property

[81] http://www.gong.manchester.ac.uk/odp/html/index.html
[82] http://ontologydesignpatterns.org
[83] http://www.neon-project.org/nw/Welcome_to_the_NeOn_Project

assertions' pattern, which enables the assertion that two given individuals cannot be connected by a given property[84].

- *content patterns* These are mini-ontologies, i.e. they can define classes, properties and individuals. Many of the content patterns in the library are generic. Some are specific to the application domains investigated in the NeOn project. Examples of content patterns, taken from *Ontology Design Patterns.org*, are the 'part of' pattern and the 'ClimacticZone' pattern. The former enables modelling of part-whole relations, and contains the mutually inverse object properties, *hasPart* and *isPartOf*. The latter is specific to the fishery domain, and includes classes *ClimacticZone* and *AquaticResource*, and properties *hasResource* and *hasClimacticZone*. Three of the content patterns from *Ontology Design Patterns.org* were used in the study described in Chapter 6.

In the ontology users' survey, and in the follow-up ontology patterns survey (Warren, 2014) there was no attempt to define or categorize ontology patterns.

5.8.2. Ontology user survey: pattern sources

In the ontology user survey (Warren, 2013), respondents were asked from where they obtained ontology patterns, with multiple responses permitted. There were 35 respondents and the percentage of respondents in each category is shown in Figure 5-7. Note that the figure moves from formal sources at the left, i.e. pattern libraries, to informal sources, i.e. own mental models, at the right. The latter were described in the question as patterns which were "not written down". Note that the phrase 'mental models' used here is to be interpreted more generally than the specific usage in Chapter 3 and in later chapters. The final category, 'other', included two references to OBO, one to a small pattern library, and one to the W3C.

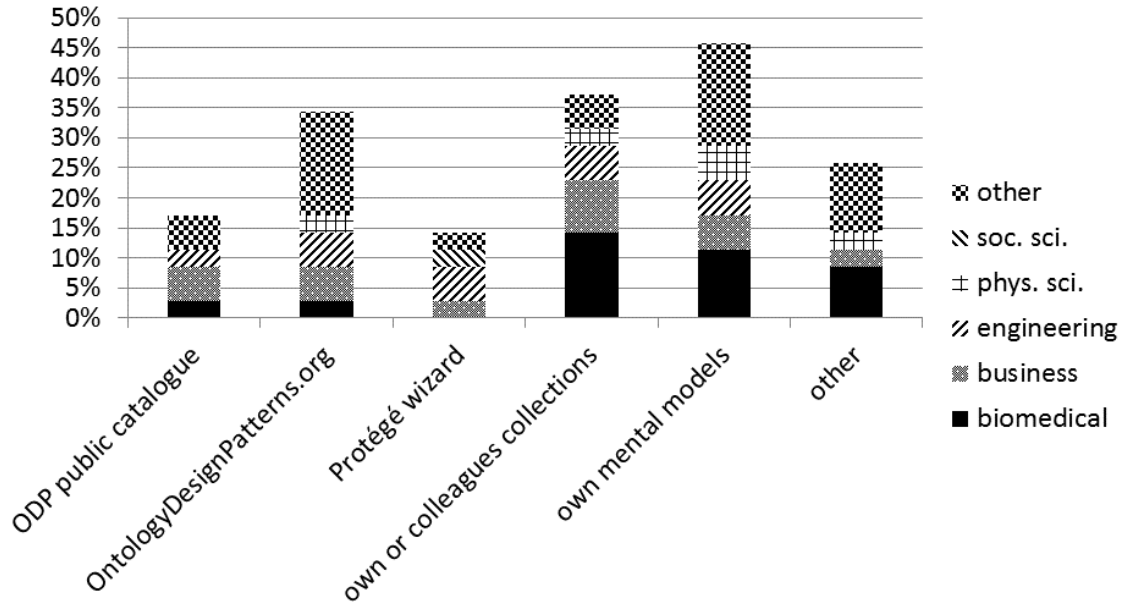

Figure 5-7 Sources of patterns; percentage of 35 respondents

It is noticeable that the biomedical respondents were much more prevalent at the right-hand half of the figure. From the biomedical respondents there was one response for each of the libraries, from the same respondent, and no responses for the Protégé wizard. The

[84] A constructor for achieving this directly is available in OWL 2. This pattern is useful for less rich modelling languages such as OWL 1.

biomedical respondents were chiefly using their own patterns, including from OBO, rather than generic patterns.

The dominant paradigm of informal reuse of patterns was also apparent in response to a question about how patterns were being used. The possible responses were: "Import patterns, e.g. as OWL files"; "Use patterns as examples and recreate them (possibly modified) in your own ontologies"; and "Other". Of the 32 respondents to this question, 28% imported patterns and 75% recreated patterns[85].

Overall, the picture which emerges regarding pattern creation and usage is a relatively informal one, with many patterns being personal to the ontology developer, or borrowed from colleagues, some not even written down, and with only a relatively small amount of downloading from pattern libraries.

5.8.3. Ontology patterns survey: pattern types

In the follow-up ontology patterns survey (Warren, 2014), respondents were asked about the size of their patterns, i.e. the number of classes, properties and individuals, and also how specific their patterns were. For the latter, respondents were able to indicate one of four categories. In order of increasing specificity, these were:

- generic
- suited to a broad discipline, e.g. anatomy
- suited to a specific discipline, e.g. human anatomy
- specific to my work

Respondents were able to specify up to five pattern types, although in fact no respondent specified more than four. Note that these were pattern types, not individual patterns. One respondent claimed to be using over thirty patterns of each of three pattern types. Another respondent claimed to be using over ten patterns of a particular type[86]. All other respondents used no more than ten patterns of a particular type

Overall, 'generic' and 'broad discipline' patterns were quite small, with at most ten classes and at most five properties; although one respondent did report using a 'broad discipline' pattern with over one hundred individuals. Some of the 'specific-discipline' and 'work-specific' patterns were also quite small, i.e. with no more than ten classes or properties, although one respondent reported use of a 'specific discipline' pattern with tens of classes and tens of properties and another reported use of 'work-specific' patterns with over a hundred classes and tens of properties. On the basis of Spearman's rank correlation, there was a significant correlation between specificity and number of classes ($\rho = 0.58$, $p = 0.015$), but not between specificity and number of individuals ($\rho = 0.47$, $p = 0.055$) or between specificity and number of properties ($\rho = 0.38$, $p = 0.136$).

It seems highly likely that the larger pattern types represent concept patterns, in general specific to respondents' disciplines or work areas, and are effectively mini-ontologies. Many of the smaller pattern types may also represent content patterns, used to represent commonly occurring structures within ontologies. However, it is possible that some of the smaller pattern types are logical patterns with prototypical classes and properties which are defined

[85] There were five responses in the 'other' category, including "Fully integrated in the tool". Some responses in this category may have arisen because of ambiguity in the question.

[86] Some respondents may have been confused by the question, possibly interpreting 'pattern' as 'pattern instantiation'.

when the pattern is instantiated, e.g. the N-Ary relation pattern mentioned in subsection 5.8.1.

5.9. Threats to validity

Section 4.8 described the three types of threats to validity and briefly interpreted them in the context of the survey. This section examines these threats in more detail.

5.9.1. Construct validity

Subsection 4.8.1 noted that the questions were reviewed prior to the survey being issued, to remove any ambiguity. It may be that some respondents had difficulty with some of the questions. For example, it has already been noted (see footnote to section 5.8.3) that, in the patterns survey, there may have been confusion between 'pattern' and 'pattern' instantiation. However, in general the responses appeared reasonable and indicative that the questions were properly understood. The use of multiple-choice responses left little scope to misinterpret the responses. Where free form responses were supplied, there was very rarely any ambiguity.

5.9.2. Internal validity

With hindsight, some of the questions could have been better designed. Section 5.1.1 has already noted that 'computing' would have been a useful category to include in the question about applications. Its absence is probably due to the author subconsciously including it within the category 'engineering'. Similarly, the question relating to DL features should have included class subsumption; property subsumption was included.

5.9.3. External validity

Subsection 4.8.3 noted that threats to external validity come from the nature of the questions and the profile of the respondents.

The questions represented a wide range of issues relating to ontology use. This section has only reported on that part of the surveys relevant to DLs and the studies reported in the next three chapters. With hindsight, more could have been done to obtain data useful for the design of those studies. In particular, the question on language features discussed in Section 5.5, could have included an option for those who did not use any of the features listed. Furthermore, the fact that the question referred to DLs might have put some respondents off, e.g. those who used only the basic RDFS features; the question could have been phrased more generally. It would also have been useful to ask participants to gauge how much they used each feature, e.g. on a scale from 'very rarely' to 'very often'. More significantly, a question or questions to query respondents on their difficulties in using DLs would have been useful; although this kind of information might be better obtained from interviews or focus groups. It is important to avoid too lengthy questionnaires; the danger is that respondents will give up before the end. One strategy to mitigate the effect of this might be to order the questions in decreasing importance to the researcher.

The profile of respondents was quite broad. They included computer scientists, e.g. from the author's university, but they also included domain experts reached via the mailing lists. Because the only domain-specific mailing lists used were concerned with biological applications, there is likely to have been some bias towards experts in the biological domain. Moreover, those who are on these mailing lists and respond to such surveys are likely to be the more committed users of ontologies. Hence the responses are unlikely to be representative of the more casual users.

5.10. Conclusions

The ontology users' survey and the follow-up ontology patterns survey were designed to gain an overview of how ontologies are being created, edited and used. This chapter has concentrated on those results which throw most light on how DLs are being used to create ontologies. OWL appears to be the dominant ontology language. Whilst many users may only be using a small subset of the OWL features, there are an appreciable number who do use the more unusual features. Amongst the property characteristics, the transitive and functional, and to a less extent the symmetric characteristics are the most commonly used, whilst the inverse functional, reflexive, asymmetric and irreflexive are the least four commonly used features overall. The use of the inverse object property feature, i.e. describing two object properties as inverses, is also relatively common.

Ontology visualization is relatively widely used, with varying degrees of enthusiasm. However, the tools currently being used, although some do allow the display of object properties, do not support reasoning about object properties. Ontology patterns were used by some respondents, frequently rather informally. However, there are two pattern libraries on the Web and these do enjoy some usage. Generic patterns from one of these libraries provides the background for the questions in the first study, see Chapter 6.

Finally, respondents made a range of general comments. These included suggestions for language extensions but also a comment on the difficulty created by the rigour of the languages and a comment on the need for better tools to help make sense of the complexity. Finally, to end the chapter on a positive note, one respondent commented of ontologies "I couldn't build what I build without them."

6. The cognitive difficulties of some common DL constructs

To err is the price for thinking.

Émile Chartier ('Alain'),
'Sketches of Man', 1927

This chapter describes a preliminary study into the difficulties experienced with some common DL features. The features were chosen carefully to achieve ecological validity, i.e. to use commonly occurring language features in commonly occurring contexts. Section 6.1 describes how this choice was made. Section 6.2 then describes how the study was organised. In the analysis of the response time data, statistical tests were used which required that the data be approximately normal[87]. Section 6.3 discusses the statistics of the response time data and the use of the log transformation to render it approximately normal. The study contained three question sections, and sections 6.4 to 6.6 describe each of these sections.

One of the research questions in Chapter 1 asked how the difficulties people experience with DLs could be interpreted in terms of theories of reasoning. As far as possible section 6.4 to 6.6 explain participant performance in terms of the theories of reasoning described in Chapter 3. Section 6.7 describes feedback from participants which provides some further insight into their difficulties. Section 6.8 analyses the relationship between performance and the self-declared level of knowledge of the participants. Section 6.9 analyses the effect of question position on performance.

Section 6.10 discusses, for questions with valid putative conclusions, the effect of the number of reasoning steps on accuracy and response time. Section 6.11 discusses the effect on response time of three interrelated factors: validity of the putative conclusion; the nature of the participant's response; and the accuracy of that response. Finally, section 6.12 draws some conclusions. The study's main findings are also described in Warren et al. (2014).

6.1. Identifying commonly used DL features

Four sources were used to establish the most commonly used DL features: the survey described in Chapter 5; an analysis of functor[88] usage made by Power and Third (2010); an analysis of axiom patterns made by Power (2010); and an analysis of content pattern usage made by Khan and Blomqvist (2010).

Power and Third (2010) provide a list of the most commonly used OWL functors based on an analysis of ontologies in the TONES ontology repository[89]. Table 6-1 shows the result of their analysis. Note that, as in Power and Third (2010), the table is ordered by the number of ontologies using the functor at least once and also shows the number of occurrences of the functors. The ordering is to guard against overestimating the importance of functors which are used considerably in a few large ontologies. Examples of these are the *ObjectPropertyAssertion* and *DataPropertyAssertion* functors which would be considerably higher in the table were the actual number of occurrences to be used[90].

---

[87] Specifically, the t-test, ANOVA and regression analysis.

[88] The term *functor* is used by Power and Third (2010) to represent statements in OWL. This stems from the use of prefix notation such that the statements appear analogous to mathematical functions with arguments. *ClassAssertion*, for example, is a functor with two arguments, the class and the individual being asserted to be in the class. *ObjectPropertyAssertion* is a functor with three arguments, the property and the two individuals related by the property.

[89] http://rpc295.cs.man.ac.uk:8080/repository/

[90] At first sight, it might appear anomalous that *ObjectPropertyRange* and *ObjectPropertyDomain* occur in so many more ontologies than *ObjectPropertyAssertion*. Possibly many of the ontologies were T-Box ontologies,

Table 6-1 Frequencies of OWL functors, from Power and Third (2010)

Functor	**No. ontologies using functor at least once**	**% age**	**No. axioms using functor**	**% age**
SubClassOf	190	94%	468,812	74.0%
EquivalentClasses	94	46%	6,082	1.0%
ObjectPropertyRange	92	45%	2,275	0.4%
ObjectPropertyDomain	91	45%	2,176	0.3%
DisjointClasses	88	43%	94,390	14.9%
SubObjectPropertyOf	75	37%	2,511	0.4%
InverseObjectProperty	63	31%	1,330	0.2%
TransitiveObjectProperty	59	29%	221	0.0%
FunctionalObjectProperty	56	28%	1,129	0.2%
DataPropertyRange	52	26%	2,067	0.3%
ClassAssertion	49	24%	12,798	2.0%
DataPropertyDomain	47	23%	2,019	0.3%
FunctionalDataProperty	37	18%	931	0.1%
ObjectPropertyAssertion	22	11%	19,524	3.1%
DataPropertyAssertion	14	7%	17,488	2.8%
SubDataPropertyOf	6	3%	12	0.0%
TOTAL	203	100%	633,791	100%

There are strong similarities between Table 6-1 and the first half of Table 5-1 in the previous chapter; although it should be borne in mind when comparing the two tables that Table 5-1 shows the proportion of respondents who use a certain feature and provides no information about the extent to which the feature is used. Moreover, the survey described in Chapter 5 did not include some of the functors shown in Table 6-1[91]. Conversely, Table 6-1 does not include certain features prominent in the survey, specifically intersection, union, and the existential and universal restrictions. It is not clear why these features were excluded; possibly they were not regarded as functors. However, Power and Third (2010) also provided a list of 'functor argument patterns', i.e. commonly occurring patterns of functor arguments, which showed that the existential restriction did occur in about 25% of the patterns whilst the universal restriction occurred in around 0.8%. Intersection occurred in 0.3% of the patterns; whilst union occurred to a negligible degree[92].

Both sources point to the high usage of object property domain and range statements, disjoint classes, object subproperties and inverse object properties. Table 6-1 contains only two object property characteristics: transitivity and functionality. Table 5-1 shows that these two were the most commonly reported in the user survey.

In another study, Power (2010) analysed 48 ontologies from the TONES repository to determine which axiom patterns frequently occurred. Table 6-2, taken from Power (2010), shows the commonly occurring axiom patterns. With the exception of the last, these are not patterns in the sense of Chapter 5. They are language features, as shown by the description

i.e. pure ontologies intended to be used in conjunction with a knowledgebase which would contain *ObjectPropertyAssertions*.

[91] Specifically: *SubClassOf*, *EquivalentClasses*, *SubObjectProperty*, *DataPropertyRange*, *ClassAssertion*, *DataPropertyDomain*, *ObjectPropertyAssertion*, *DataPropertyAssertion*. In these respects the survey was not comprehensive.

[92] If it occurred at all, it was in the 1.8% of patterns classified as 'other patterns'.

column added by this author in Table 6-2, but not present in Power (2010). They all occur in one or both of the set of features identified by the survey, in Table 5-1, or by Power and Third (2010) in Table 6-1. The final pattern in Table 6-2 can be regarded as a logical pattern, in the sense described in Chapter 5.

Examining the features present in Table 6-2 supports the work of Power and Third (2010) in identifying the importance of class subsumption, disjointness, object and data property assertion, and class assertion. Besides this, Table 6-2 confirms that the existential restriction is commonly used. Intersection appears to be present to an appreciable amount via the disjointness pattern. However, this is likely to have been achieved by using the DisjointClass statement, which Power and Third (2010) had already identified as occurring in 14.9% of the axioms, see Table 6-1. Intersection will have occurred explicitly in the final statement (intersection of atomic and anonymous class), which is consistent with the limited use found by Power and Third (2010).

Neither the universal restriction nor union were present in any of Power's (2010) commonly occurring patterns. This is consistent with Power and Third's (2010) finding very limited usage of the universal restriction and no usage of union. These results contrast with the survey in Chapter 5, where there was appreciable usage of intersection and union, and of both the existential and universal quantifiers. Again, the observation needs to be made that the survey was looking at the extent of respondents who used a certain feature, not the extent to which that feature was used.

Table 6-2 Axiom pattern frequencies, taken from Power (2010)

Pattern	**Description**	**Frequency**	**% age**
$C \sqsubseteq C$	subsumption	18,961	42.3%
$C \sqcap C \sqsubseteq \bot$	disjointness	8,225	18.3%
$C \sqsubseteq \exists P.C$	existential restriction	6,211	13.9%
$[I, I] \in P$	object property assertion	4,383	9.8%
$[I, L] \in D$	data property assertion	1,851	4.1%
$I \in C$	class assertion	1,786	4.0%
$C \equiv C \sqcap \exists P.C$	intersection of atomic and anonymous classes	500	1.1%
Other		2,869	6.4%
TOTAL		44,786	100%

N.B. (i) C, P, D, I and L refer to atomic classes, object properties, datatype properties individuals and literals. The occurrence of a symbol more than once in a given pattern does not imply that it refers to the same entity. E.g. the two occurrences of C in the first pattern could refer to different classes. (ii) Description column added by author.

There were clearly only a limited number of questions which could be asked in this and the subsequent studies, and hence only a limited number of DL features which could be examined. In order to limit these features, it was decided to exclude data properties. The rationale for this is that data properties are not likely to pose any logical properties additional to those posed by the analogous object properties. All the language features associated with data properties in Tables 5-1, 6-1 and 6-2[93] also occurred in their object property form and in that form were included in the set of features investigated in the study.

[93] Namely, data subproperties, functional data properties, data property range, data property domain, and data property assertion.

The list of features used in the study is shown in Table 6-3. With the exception of features relating to data properties, this includes all the language features in Tables 6-1 and 6-2 plus all those in Table 5-1 except cardinality and the last four in Table 5-1: inverse functional object properties, reflexive object properties, asymmetric objective properties, and irreflexive object properties. As discussed in more detail shortly, Khan and Blomqvist (2010) have identified the 20 most commonly used content patterns. Table 6-3 includes all the features found in those patterns, with the exception of features relating specifically to data properties, as already discussed, and cardinality restrictions. Cardinality restrictions were excluded on the grounds that the inclusion of the universal and existential restrictions tested understanding of the general concept of a restriction in DLs, whilst cardinality restrictions merely added considerations of arithmetic.

Power and Third (2010) do not mention *complement of a class*. Nor does it occur in the common axiom patterns found by Power (2010). However, it was employed by 47% of the respondents to the survey question about DL feature usage. Consequently, complement is used in this and the two subsequent studies. In addition it has been known for some time that negation is difficult for human reasoners (Wason, 1959) and, since complement is the analogue in DL of negation in PL, it seems useful to evaluate performance with complement. Moreover, it may be the case that a less frequently used feature causes a disproportionate amount of difficulty and it seems useful to test out a feature which, although not common, could be a source of difficulty. Similarly, *union* occurs neither in Power and Third's (2010) list nor Power's (2010) list, although it was employed by 66% of the respondents to the survey question. It was included in the study in part to enable a comparison with *intersection*.

Table 6-3 Commonly used DL features used in study 1

	language feature	**Manchester OWL Syntax**
Class features	subsumption	SubClassOf
	class equivalence	EquivalentTo
	disjoint classes	DisjointWith
	class assertion	Type
	conjunction	and
	disjunction	or
	complement	not
Property features	property range	Range
	property domain	Domain
	property hierarchy	SubPropertyOf
	inverse object properties	InverseOf
	transitive object property	Characteristics: Transitive
	functional object property	Characteristics: Functional
	symmetric object property	Characteristics: Symmetric
Restrictions	existential restriction	some
	universal restriction	only

Besides ensuring that the studies test out the most frequently used language features, another aspect of ecological validity is using those features in a realistic context, i.e. combining them in the ways in which they are commonly combined in real applications. One way to do this is to make use of commonly occurring content patterns. Evidence for the usage of content patterns comes from Khan and Blomqvist (2010). They searched 682 online ontologies to

determine the frequency of occurrence of the 76 content patterns in the *OntologyDesignPatterns.org* library (see subsection 5.8.1), and list the twenty most frequently detected[94].

The five most frequently detected patterns were, in decreasing order of usage: *constituency*, *participation*, *componency*, *coparticipation*, and *types of entities*. The first two of these were ruled out for use in the study because of their simplicity. The remaining three were used in the study, and are illustrated later. *Coparticipation* and *types of entities* were extended to enable all the features of Table 6-3 to be used in at least one pattern in the study.

6.2. The study

The study was administered using the *SurveyExpression*[95] tool, which displayed the questions and recorded the responses, and was conducted under laboratory conditions. *Camtasia*[96] was run on the PC to record screen activity and the recordings were subsequently analysed to obtain timing data. The study consisted of five sections. In the first section participants were asked to provide information about themselves, e.g. they were asked to indicate their knowledge of logic. The next three sections contained the questions, each section being based on one of the content patterns mentioned in Section 6.1. Each section began with a page displaying the content pattern. The following pages then contained the questions. Each question page contained a set of axioms and a putative conclusion. Participants were required to indicate whether the conclusion was a logical consequence of the pattern and the axioms listed in the question. Each question page also repeated the pattern, so there was no need for the pattern to be memorized by participants. In all there were 21 questions, 13 with valid conclusions and 8 where the conclusions were not valid. In the final section of the study participants were invited to give feedback. This is discussed in section 6.6. Throughout the study participants were free to move at their own pace. Note in particular that merely indicating an answer did not move the study on to the next page. To do that participants were required to click on 'Next'. Participants were therefore free to change a response before moving on.

The patterns, axioms and putative conclusions were expressed in MOS[97]. Participants were provided with a handout, which explained the syntax and semantics of exactly that part of MOS required for the questions. They were asked to read the handout before commencing the study and kept the handout for reference whilst completing the study. As with the patterns, there was no requirement to memorize the syntax or semantics of the language. Participants were also asked not to use pen and paper.

All the participants had some background in computer science, and were from the Knowledge Media Institute or the Centre for Research in Computing, both at the Open University. There were 12 participants, three of whom were female. To reduce any effect of question order, each of the six permutations of section order was used twice. The order of questions within each section was not varied. Section 6.9 considers the effect of question position on participant performance.

[94] 'detected' because the search process was not 100% comprehensive.
[95] www.surveyexpression.com
[96] https://www.techsmith.com/camtasia.html
[97] As explained in section 2.3, the syntax was slightly simplified by removing colons from the end of keywords, e.g. 'Class:' was simplified to 'Class'. Also, 'Types' was replaced with 'Type'.

6.3. Response time data

In the analysis of response time data reported later in this chapter, use has been made of statistical tests which require that the data be at least approximately normal. However, as can be seen from Figure 6-1, the response time data for the study showed a positive skew. Moreover, a Shapiro-Wilk test for normality gave a p-value of 1.2×10^{-15}, indicating that the data deviated considerably from the normal. The data was transformed using appropriate transformations from Tukey's ladder of powers, as explained in section 4.3, working from the identity transformation to the left along the ladder of powers, to correct for the positive skew. In each case the Shapiro-Wilk test was applied. Table 6-4 shows the p-values obtained for each transformation. As can be seen, the log transformation provides the best transformation to the normal[98]. Figure 6-2 shows the density distribution of the transformed response time data.

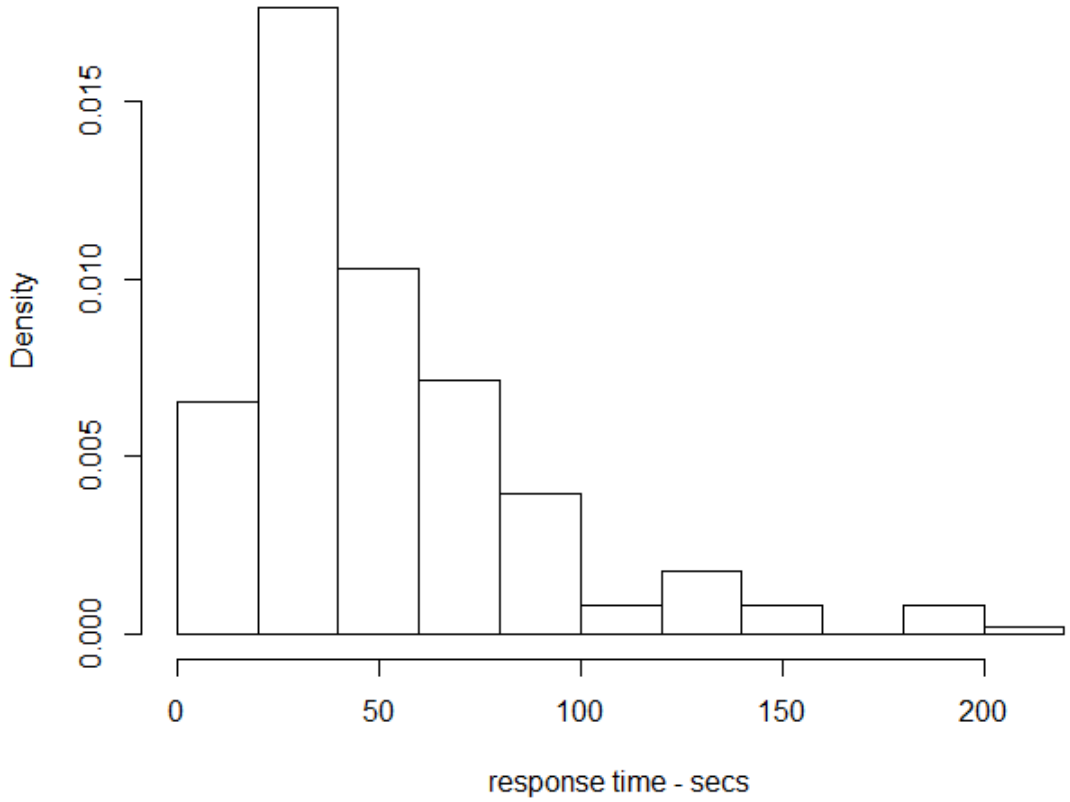

Figure 6-1 Density distribution for response time; area under histogram normalized to 1

[98] Note that when the log transformation is applied, the base of the log will make no difference to the degree of normality of the resultant data. However, all log transformations of time in this and subsequent chapters are to base 10 to aid interpretation of data plots such as in Figure 6-2.

Table 6-4 Shapiro-Wilk test applied to response time data

Transformation	**p-value**
time (i.e. identity transformation)	1.2×10^{-15}
square root (time)	6.2×10^{-8}
log (time)	0.50
1 / square root (time)	3.1×10^{-6}

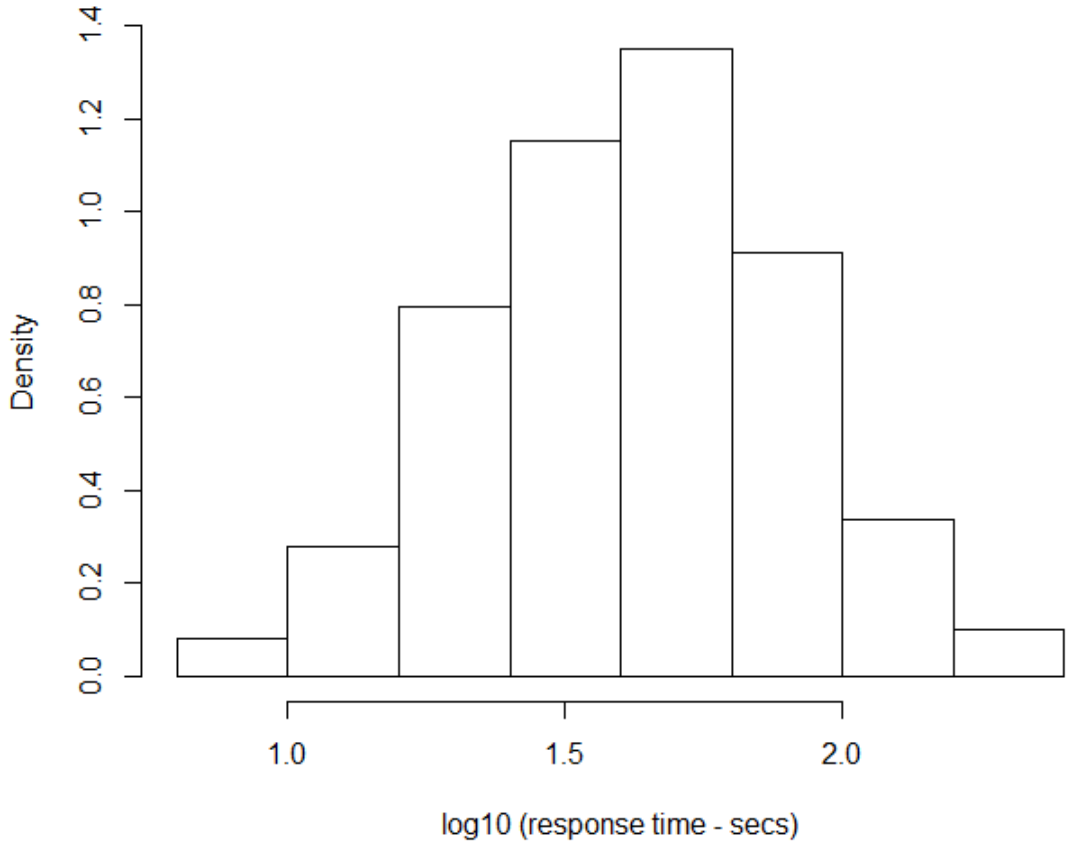

Figure 6-2 Density distribution for log10 (response time);
area under histogram normalized to 1

Frequently, the requirement is to compare questions, and hence it is the normality of data from individual questions which is important, rather than the normality of the data overall. One approach is to identify a transformation which minimizes the maximum deviation from normal across each of the 21 questions. Table 6-5 shows the minimum Shapiro-Wilk p-value obtained across all the questions for an appropriate range of transformations. This suggests that the reciprocal of the square root of time might be slightly preferable to the log transformation.

However, Figure 6-3 shows the results of applying Shapiro-Wilk tests to each question, for both the reciprocal square root and log transformation. From the figure it can be seen that there are ten questions for which the log transformation gives a better approximation to normality, compared with seven questions for which the reciprocal square root is better. For the other four questions there is little appreciable difference between the two transformations. In fact, it can be seen that, with one exception, the p-values obtained with the Shapiro-Wilk test after log transformation are greater than 0.032 ($10^{-1.5}$); the exception

is the value of 0.00753 shown in Table 6-5. With three exceptions, the Shapiro-Wilk p-values after the log transformation are greater than 0.1. Hence, the log transformation was used on time data prior to statistical testing. This also maintains consistency with the approach in the other two studies where, as will be seen in the next two chapters, the log transformation gave the best approximation to normality.

Table 6-5 Minimum p-value for Shapiro-Wilk test applied to response time data across all questions

Transformation	**p-value**
time (i.e. identity transformation)	0.00006
square root (time)	0.00097
log (time)	0.00753
1 / square root (time)	0.00966
1 / time	0.00030

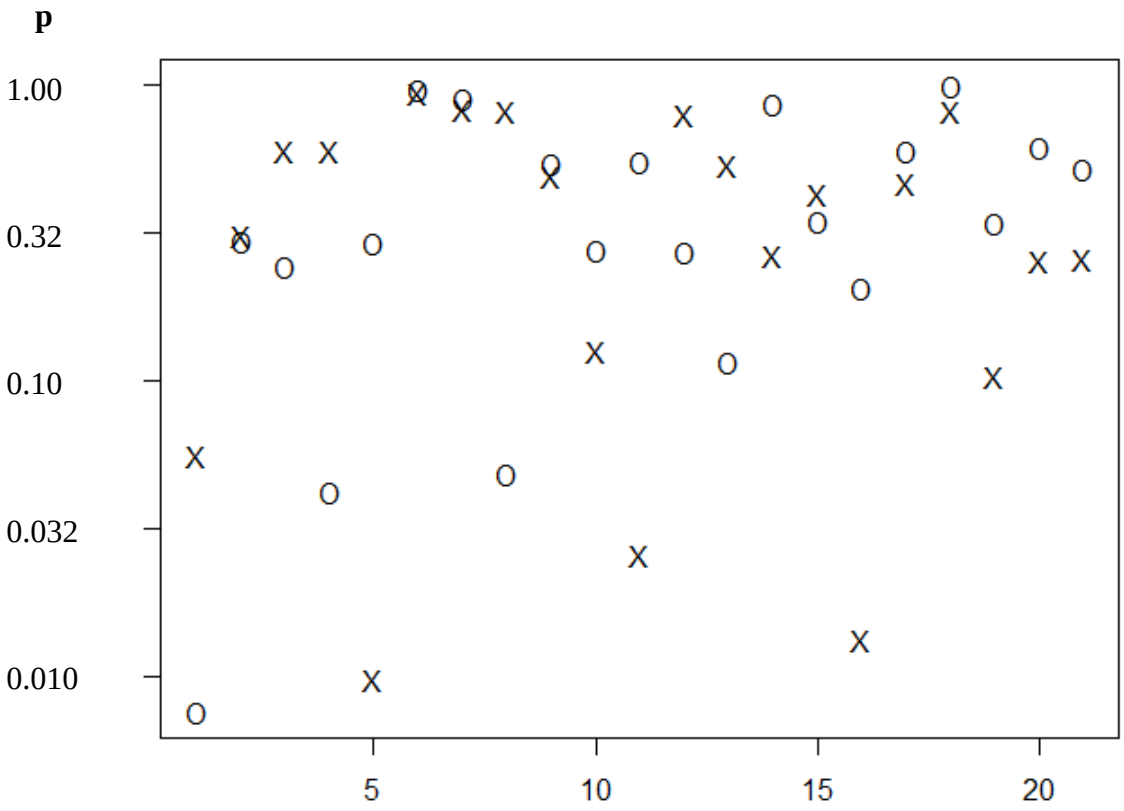

Figure 6-3 Shapiro-Wilk p-values for each of 21 questions, after log transformation (circle) and reciprocal square root transformation (x). Shown on a $\log_{10}$ scale.
N.B. $10^{-0.5} \approx 0.32$; $10^{-1.5} \approx 0.032$

6.4. The componency pattern

There were ten questions based on the componency pattern. Table 6-6 shows the pattern and Table 6-7 shows the questions, i.e. the axioms and putative conclusions. Table 6-7 shows whether the conclusion is valid or non-valid, the number of correct responses and the mean and standard deviation of the response time for each question. For the valid questions, the table also shows the number of reasoning steps required; Appendix A provides a justification for the number of reasoning steps associated with each valid question in this study. The questions all involve reasoning about object properties, i.e. relations, and involve

transitivity, inverse properties, property subsumption, as well as the universal restriction and class assertion.

Table 6-6 Componency pattern

Class Object	SubClassOf has_component only Object SubClassOf is_component_of only Object
Property has_part	Characteristics Transitive
Property is_part_of	Characteristics Transitive
InverseOf	has_part
Property has_component	SubPropertyOf has_part
Property is_component_of	SubPropertyOf is_part_of InverseOf has_component

Table 6-7 Questions associated with the componency pattern

No.	Question	validity	no. steps	%age correct	mean time (s.d.) - secs		
					overall	correct	incorrect
1	A is_part_of B; C is_part_of B ⇒ A is_part_of C	not-valid	n/a	100%*	62.3 (53.0)	62.3 (53.0)	n/a
2	A is_part_of B; B is_part_of C ⇒ A is_part_of C	valid	1	100%*	20.3 (9.7)	20.3 (9.7)	n/a
3	B is_part_of C; A is_part_of B ⇒ A is_part_of C	valid	1	100%*	30.7 (34.0)	30.7 (34.0)	n/a
4	A has_component B B has_component C ⇒ A has_component C	not-valid	n/a	33%	62.8 (36.2)	77.5 (55.9)	55.4 (23.0)
5	A has_component B B has_component C ⇒ A has_part C	valid	3	83%*	29.0 (15.8)	29.7 (17.0)	25.5 (10.6)
6	A has_component B B is_part_of C ⇒ A has_part C	not valid	n/a	83%*	57.9 (31.0)	49.1 (16.9)	102.0 (58.0)
7	A has_component B C is_part_of B ⇒ A has_part C	valid	3	100%*	37.4 (37.7)	37.4 (37.7)	n/a
8	A Type Object A has_component B C Type not Object ⇒ B DifferentFrom C	valid	2	100%*	49.9 (18.4)	49.9 (18.4)	n/a
9	A Type Object; A has_part B C Type not Object ⇒ B DifferentFrom C	not valid	n/a	83%*	47.5 (25.4)	48.4 (24.2)	43.0 (42.4)
10	A has_component B C is_component_of B ⇒ C is_part_of A	valid	4	100%*	54.2 (30.5)	54.2 (30.5)	n/a

** answered significantly better than chance ($p < 0.05$)*

In general, these questions were answered very well. Six of the questions were answered correctly by all participants. Given the sample size, 100% correct responses for a question is significantly better than chance ($p < 0.001$, Fisher's Exact Test, one-sided). Another three questions were answered correctly by 83% of the participants, which is also significantly better than chance ($p = 0.019$, one-sided test). This shows that respondents were able to cope well with reasoning about transitivity, inverse properties and property subsumption. Moreover, in question 8 they were able to combine class assertion with reasoning about a universal restriction.

The exception to this high level of accuracy is question 4, where only 33% of the participants responded correctly, i.e. worse than chance although not significantly so ($p = 0.194$, one-sided test). It appears that the participants who answered incorrectly have erroneously assumed that *has_component* is transitive. There are a number of reasons why this might be:

1. The layout of the question, *A ... B; B ... C*, encourages the elimination of the shared term, B, to arrive at *A ... C*.
2. On reaching question 4, the participants have completed three questions involving the use of *is_part_of*, which is transitive, and are therefore in the habit of assuming transitivity[99].
3. They are aware that *has_component* is a subproperty of *has_part*, which is transitive and may have falsely assumed that transitivity is 'inherited'.
4. Participants may take account of their everyday understanding of what it means to be a component and assume transitivity, e.g. as in 'the engine is a component of the car and the piston is a component of the engine, hence the piston is a component of the car'[100].

Whilst it is not possible to know to what extent these, or possibly other, factors influenced the participants, the third suggests a potential confusion regarding property subsumption. As explained in subsection 2.3.3, in general subproperties do not inherit characteristics from their superproperties. This was made clear in the handout, which gave transitivity as an example of a characteristic which is not inherited. This fact may have been forgotten by study participants under the cognitive stress of answering the question. In particular, the assumption of inheritance of characteristics is a natural error for those used to object oriented programming.

The fourth factor indicates the need for careful choice of names, in patterns and in ontology design generally. It is perfectly possible for an entity name to have associated semantics which are not intended to be assumed in the context of the ontology. In the case of the componency pattern it might have been better to have used *has_direct_component* or *has_immediate_component* to avoid the assumption of transitivity.

Finally, a comment needs to be made regarding questions 2 and 3. These questions are identical, except for the order of the two axioms. Both questions received 100% correct responses. However, the mean time for question 3 is appreciably longer than question 2. This might suggest that participants found the order of question 3 harder than question 2, perhaps having to reverse the order of axioms before reaching the conclusion. Whilst this may have been a common strategy, there is in fact no significant difference between the times to respond on the two questions ($t(11) = 0.91319$, $p = 0.381$, paired two-sided test). In fact, the relatively large difference in the two mean times arises almost completely from the existence of one response time of 132 seconds for question 3. When this datapoint is

[99] In retrospect the order of the questions could have been arranged to avoid this, or the order of the questions in each section could have been randomized to compensate for any inter-question interaction. The issue of the effect of question position is consider later in this chapter and in later chapters.

[100] The handout did state that where meaningful names were used they should not affect any conclusions to be drawn. However, such caveats can be easily forgotten when answering actual questions.

removed, the mean time for question 3 becomes 21.5 seconds, i.e. very little greater than for question 2, which was 20.3 seconds[101].

6.5. The modified coparticipation pattern

There were six questions based on a modification of the coparticipation pattern. The modified pattern is shown in Table 6-8. Table 6-9 shows the questions and data regarding the questions.

Table 6-8 Modified coparticipation pattern

Class Event	EquivalentTo has_participant some Object
	DisjointWith Object
Class Object	*DisjointWith Event*
Class Player	*SubClassOf Object*
Class Game	*SubClassOf has_participant some Player*
Property coparticipates with	Domain Object, Range Object
	Characteristics Symmetric, Transitive
Property has_participant	Domain Event, Range Object
	InverseOf is_participant_in

N.B. the statements in italics were added to the original pattern.

Table 6-9 Questions associated with modified coparticipation pattern

No.	Question	validity	no. steps	%age correct	mean time (s.d.) - secs		
					overall	correct	incorrect
1	A coparticipates_with B ⇒ A Type not Event	valid	2	92%*	54.9 (35.5)	55.5 (37.2)	48.0 (n/a)
2	A is_participant_in B C coparticipates_with D ⇒ A DifferentFrom C	not valid	n/a	92%*	68.8 (49.3)	66.4 (50.9)	96.0 (n/a)
3	A is_participant_in B C is_participant_in B ⇒ A is participant in C	not valid	n/a	100%*	43.6 (31.4)	43.6 (31.4)	n/a
4	A has_participant B C is_participant_in D ⇒ B DifferentFrom D	valid	4[102]	92%*	44.6 (20.9)	44.3 (21.8)	48.0 (n/a)
5	B coparticipates_with A B coparticipates_with C ⇒ C coparticipates with A	valid	2	100%*	34.8 (15.9)	34.8 (15.9)	n/a
6	A Type Game ⇒ A Type Event	valid	3	67%	47.6 (25.6)	39.0 (23.5)	64.8 (23.0)

** answered significantly better than chance ($p < 0.05$)*

As with the questions in the previous section, these questions were also answered in general very well. Two were answered correctly by all participants, which is significantly better than chance ($p < 0.001$, one-sided test) and three were answered correctly by 92% of the participants, which is also significantly better than chance ($p = 0.003$, one-sided test). As with the previous questions, participants were able to reason well about class subsumption, transitivity and inverse properties. Added to these features, they could also reason well about symmetry and disjointness.

[101] Note that a clue to this behaviour is given by the relatively large standard deviation for question 3. This can be indicative of one or more outlying datapoints. In this case, when the outlier is removed, the standard deviation for the remaining eleven points is 12.4 seconds.

[102] This number was erroneously give as 3 in Warren et al. (2014).

The one exception was question 6, which was not answered significantly better than chance ($p = 0.194$, one-sided test). A distinguishing feature of this question is the use of the existential restriction. It is not possible to state definitively whether the problem arises because of the complexity of the question or whether there is an inherent difficulty with the existential restriction. As Table 6-9 shows, the number of steps (3) is not greater than for several questions in the study which were answered appreciably more accurately. Moreover, in question 8 of the previous section, all participants answered correctly a question involving the universal restriction. This might suggest that the existential restriction is harder to reason with than the universal restriction, although other differences between the questions could be responsible for the difference in accuracy. In any case, as discussed in section 2.2, the use of restrictions to define classes will be unfamiliar to those accustomed to conventional logic and the closed world convention. On the other hand, the questions involving only object properties will involve reasoning processes quite familiar from logic and mathematics, and even everyday life. One remedy for this is more emphasis on, and practice with, restrictions during training.

6.6. The types of entities pattern

There were five questions based on a modification of the types of entities pattern. Table 6-10 shows the modified pattern. Table 6-11 shows the questions and data regarding the questions. Performance on these questions was in general worse than on the questions described in the preceding two sections. Only two questions were answered significantly better than chance: question 4 ($p = 0.003$, one-sided test) and question 1 ($p = 0.019$, one-sided test). Question 5 was answered better than chance, but not significantly so ($p = 0.073$, one-sided test). Question 3 was answered exactly at chance. Question 2 was answered worse than chance, but not significantly so ($p = 0.073$, one-sided test). Questions 2, 3 and 5 are considered in the next three subsections.

Table 6-10 Modified types of entities pattern

Class Entity	EquivalentTo Event or Abstract or Quality or Object
Class Event	SubClassOf Entity
	DisjointWith Abstract, Quality, Object
Class Abstract	SubClassOf Entity
	DisjointWith Event, Quality, Object
Class Quality	SubClassOf Entity
	DisjointWith Event, Abstract, Object
Class Object	SubClassOf Entity
	DisjointWith Event, Abstract, Quality
Class Nonconceptual	*EquivalentTo Event or Object*
Class Nontemporal	*EquivalentTo Abstract or Quality or Object*
Property represents	*Characteristic Functional*

N.B. the statements in italics were added to the original pattern.

Table 6-11 Questions associated with modified types of entities pattern

No.	Question	validity	no. steps	%age correct	mean time (s.d.) - secs		
					overall	correct	incorrect
1	A represents B; C represents D ⇒ A DifferentFrom C	not valid	n/a	83%*	91.5 (61.7)	80.3 (57.0)	147.5 (71.4)
2	A Type Entity A Type not (Event and Quality) ⇒ A Type (Abstract or Object)	not valid	n/a	25%	75.1 (48.1)	64.7 (29.7)	78.6 (53.9)
3	A represents B; C represents D B Type Object; D Type Event ⇒ A DifferentFrom C	valid	2	50%	75.8 (31.0)	82.8 (20.5)	68.7 (39.6)
4	A Type Entity A Type not (Event or Quality) ⇒ A Type Abstract or Object	valid	2	92%*	44.0 (19.1)	42.4 (19.1)	62.0 (n/a)
5	A Type (Nonconceptual and Nontemporal) ⇒ A Type Object	valid	3	75%	63.1 (32.4)	65.1 (37.6)	57.0 (7.9)

** answered significantly better than chance ($p < 0.05$)*

6.6.1. Negated conjunction – question 2

Question 2 should be contrasted with the more accurately answered question 4; the difference being that question 2 uses conjunction in the second premise; question 4 uses disjunction. The difference in accuracy is significant (Fisher's Exact Test, p = 0.003). Question 2 also took significantly longer than question 4 (t(11) = 2.7898, p = 0.018, paired test).

One way to reason about question 2 is to note that, since *Event* and *Quality* are disjoint, *Event and Quality* must be *Nothing* (⊥) and hence *not (Event and Quality)* is *Thing* (⊤). Hence the second premise is tautological and provides no information. The first premise alone cannot limit *A* to the union of *Abstract* and *Object*.

It is interesting to compare the difference in performance between questions 2 and 4 with the difference Khemlani et al. (2012a) found between negated conjunction and negated disjunction. As noted in Chapter 3, subsection 3.3.1, they found that 18% of their participants were able to reason accurately about negated conjunction, compared with 89% for negated disjunction. They interpreted this in terms of mental model theory, in particular a failure to create all three mental models associated with negated conjunction. They suggest that, expressed in the context of question 2, people represent *not (Event and Quality)* as one mental model: *not Event and not Quality*.

Another contributory factor in both cases may be the ambiguity of the word *and*, as also discussed in Chapter 3. Knowing that *Entity* consists of four subclasses, and misinterpreting *and*, they might simply delete *Event* and *Quality* to arrive at the union of *Abstract* and *Object*.

Note that both these failure mechanisms lead to a mental model for question 2 which is identical to that for question 4, for which the correct response is 'valid'. One way of reducing this difficulty is to make students learning DLs aware of De Morgan's Laws[103].

6.6.2. Functional object property – question 3

The correct reasoning for question 3 is that, since Object and Event are disjoint, B and D must be different. The functionality of *represents* then ensures that A and C are different.

[103] *not (X and Y)* ≡ *not X or not Y* and *not (X or Y)* ≡ *not X and not Y*

There are a number of possible reasons why the accuracy of response on this question is low, in particular compared with questions relating to transitivity in the other sections of the study. It may be that participants were influenced by their inherent understanding of the words being used. The transitive property names used in the previous two patterns in sections 6.4 and 6.5, e.g. *has_part*, may appear inherently transitive. Whereas *represents* may not appear inherently functional; it is perfectly possible for something to represent more than one entity, e.g. the colour blue can represent the sky and the sea, besides a mental state. Note that the property *represents* was not present in the original types of entities pattern but was added in order to enable a question about functionality. However, the use of meaningful names is likely to be ecologically more valid than abstract names; this is generally the practice in ontologies, with the exception of some in the life sciences. As in section 6.4, this highlights the importance of careful choice of names in real-life ontologies; *represents_only* might have been a better choice of name. Related to this, there is also a possible confusion between functionality and inverse functionality. Note that this does not happen with transitivity; the inverse of a transitive function is also transitive[104]. One remedy for this confusion is suggested in Chapter 8.[105]

Another source of difficulty with this question may arise from its inherent complexity. Because one reasoning step requires the manipulation of four entities, the question has a relational complexity of four. The study reported in Chapter 7 investigated performance with a functional property in a less complex situation, making possible a comparison with a similarly complex question employing a transitive property.

6.6.3. Conjunction and disjunction – question 5

Question 5 was answered better than chance, but not significantly so. In fact, two participants answered incorrectly whilst one participant did not provide an answer[106]. The classes *Nonconceptual* and *Nontemporal* are the unions of two and three subclasses respectively. All of these subclasses are mutually disjoint, hence the conjunction of *Nonconceptual* and *Nontemporal* must be the one subclass they have in common, i.e. *Object*. It is possible that participants who had difficulty with this question were confused by the ambiguity of *and*, as with the question discussed in subsection 6.6.1, and interpreted the operator as union rather than intersection.

Another difficulty might have been holding the four subclasses in working memory and remembering which ones were subclasses of each of the two superclasses. Effectively the participant is looking for the overlap between the two superclasses. This might be regarded as holding two mental models, one with two subclasses, the other with three, in working memory at the same time. This can be avoided by a syntactic approach in which each of the subclasses from one superclass (e.g. *Event* and *Object* from *Nonconceptual*) is in turn

[104] Let T be a transitive property and assume that $x\ T^{-1}\ y$ and $y\ T^{-1}\ z$. Then $x\ T^{-1}\ y \Rightarrow y\ T\ x$ and $y\ T^{-1}\ z \Rightarrow z\ T\ y$. Then $y\ T\ x; z\ T\ y \Rightarrow z\ T\ x \Rightarrow x\ T^{-1}\ z$. I.e. $x\ T^{-1}\ y$ and $y\ T^{-1}\ z$ imply $x\ T^{-1}\ z$ and hence T^{-1} is transitive.

[105] A fundamental question is the extent to which, in real-life situations, ontology users draw on their inherent understanding of the characteristics of properties, rather than making use of characteristics stated in the ontology. It is possible that this happens to a large extent. The declaration of property characteristics may be required chiefly for the benefit of software reasoners. In this case, using abstract names in studies such as these would deviate completely from normal usage. In any case, the choice of property names in real-life ontologies should definitely avoid any inappropriate suggestion of property characteristics. On the other hand, the choice of a name which conveys a required characteristic might support user reasoning. However, the latter strategy can raise problems if the characteristics of a property are changed.

[106] Participants were permitted to skip a question if they felt unable to answer it. The skipped question was treated as an incorrect answer. In fact, question 5 in Table 6.8 was the only question skipped. The software used for the third study, described in Chapter 8, did not permit questions to be skipped.

conjunctively combined with each subclass from the other superclass (*Abstract*, *Quality* and *Object* from *Nontemporal*). Each conjunction will either result in *Nothing* (⊥) or, in one case, *Object*. Thus the most subclasses that need to be held in working memory are three, the two subclasses currently being conjunctively combined and, possibly, *Object*[107]. This approach requires an, at least intuitive, understanding that conjunction distributes over disjunction.

6.7. Participant feedback

After completing the questions, participants were able to provide written feedback about what they found difficult and what they found easy, and to make general comments. Some participants also made comments verbally. The most common theme was the use of intuition, in particular relating to names. Here there were conflicting views. One participant (p1) commented that "using named individuals instead of capital letters would have been easier", whilst another (p2) held the opinion that it was "easy to reason with anonymous things", since this safeguarded against the danger of using intuition rather than relying on formal axioms. These contrasting views were also present when considering class and property names. One participant (p3) commented that because class and property names were familiar it was necessary to check whether the meaning in the OWL expression was similar to the normal English usage; another (p4) stated that "the axioms were realistic so one could rely to some extent on common sense". Of course, as has already been pointed out in Chapter 5, in real-world situations it is likely that ontology users do make use of their intuitions, based on the names used for entities. It is the software, not the human, reasoners which work only with the axiom systems. On a related, but different, theme participant p1 also commented favourably on the lack of use of formal logic symbols, which is a feature of MOS.

Four participants commented on the value of diagrams. Here there were no conflicting views but a consensus that diagrams are useful, e.g. participant p3 stated: "perhaps I would have done better if I'd drawn diagrams on paper" and another participant (p5) commented: "a pictorial representation of the relationships would have been easier to use". One participant (p6) expressed a related view that colour-coding for OWL entity types and font weights and styles for keywords would be useful.

There were some interesting comments about OWL features, including the difficulty of using the existential and universal quantifiers (participant p1); confusion between *and* and *or* (participant p3); and the effect of users' legacy, e.g. that of a database background (participant p7). These all echo observations made in previous sections.

6.8. Participant performance

At the beginning of the study participants were asked to rank themselves on their knowledge of formal logic and on their knowledge of OWL or another DL formalism. Subsection 6.8.1 reports on how these rankings correlated with participant performance. Participants were also asked about their usage of DL. Subsection 6.8.2 discusses an analysis of the relationship between DL usage and participant performance.

[107] Although, if the subclasses are dealt with in the order shown in the two *Equivalent* statements in the pattern, then *Object* only occurs as the result of the conjunction at the very end of the process.

6.8.1. Knowledge of logic and DL

Participants were asked to rank themselves on their knowledge of formal logic and their knowledge of OWL or another DL formalism, in both cases using the same categories:

- No knowledge at all
- A little knowledge
- Some knowledge
- Expert knowledge

Table 6-12 shows for each of the two questions, a breakdown by category giving: the number of participants in each category, the mean percentage of correctly answered questions, plus mean and standard deviation times to complete the study. Note that in both cases there were no participants in the 'no knowledge at all' category. Note also that the time here includes the time reading the on-screen study preamble plus the preamble for each section, i.e. it is greater than the sum of the times for each question. As can be seen, there was increasing performance, in terms of increasing percentage correct and reducing mean time, with increasing knowledge both of logic and of DL. It should be noted, though, that in the case of knowledge of logic, there was a very uneven distribution of participants, with the majority being in the 'some knowledge' category.

Table 6-12 Participant performance by knowledge of logic, and by knowledge of OWL or another DL formalism

	Knowledge of logic				**Knowledge of OWL or other DL formalism**		
	no. participants	percentage correct	mean (s.d.) time - secs		no. participants	percentage correct	mean (s.d.) time - secs
A little knowledge	2	64%	1935 (968)		3	78%	2289 (384)
some knowledge	8	87%	1762 (796)		5	81%	1771 (922)
expert knowledge	2	88%	1107 (346)		4	90%	1113 (254)

There was a significant Spearman's rank correlation between knowledge of formal logic and accuracy of response ($\rho = 0.53$, $p = 0.038$, one-sided test)[108] and between knowledge of OWL or DL and accuracy ($\rho = 0.54$, $p = 0.036$, one-sided test). This should not be taken to imply that, in general, knowledge of formal logic and knowledge of DL equally influence accuracy for these questions. For this particular set of participants, knowledge of formal logic and knowledge of DL were significantly correlated ($\rho = 0.58$, $p = 0.024$, one-side test). In fact, seven of the twelve participants rated their knowledge of the two topics equally, and for the other five participants there was never more than a difference of one place on the scale. Hence from this particular set of participants it is not possible to separate out the effects of knowledge of logic and knowledge of DL.

There was a significant Spearman's rank correlation between knowledge of DL and total time to answer the questions ($\rho = -0.65$, $p = 0.011$, one-sided test). However, the Spearman's rank correlation between knowledge of formal logic and total time was not significant ($\rho =$

[108] In this and subsequent usages of Spearman's rank correlation in this chapter, the p-value could not be calculated exactly because of ties. It is assumed that this does not materially affect the results.

-0.29, p = 0.178). Here, the small number of participants in two of the categories makes reliable analysis difficult[109].

6.8.2. Usage of DL

Participants who used OWL or another DL formalism were asked whether they were:

- learning about the language
- using the language in their work
- researching the language

There was also an 'other' option. One of the participants did not answer the question; the participant was assumed to make no use of DL. Another participant used the 'other' option to indicate no use of DL. No participants were learning about the language. Six participants used DL and four were researching DL. Table 6-13 shows a breakdown by category giving: the number of participants in each category, the percentage of correctly answered questions, plus mean and standard deviation times to complete the study.

Table 6-13 Participant performance by usage of DL

	no. participants	**percentage correct**	**mean (s.d.) time - secs**
no use	2	76%	2500 (168)
use in work	6	83%	1638 (877)
researching	4	88%	1364 (438)

One might expect the performance of those participants who do not use DLs to be worse than for the other participants. This is indeed the case, although the sample of those not using DLs is very small. A logistic analysis of deviance revealed no significant difference in percentage correct across the three categories (p = 0.229). Similarly, an ANOVA revealed no significant difference in time across the three categories ($F(2,9) = 1.908$, $p = 0.204$). These results are not surprising. Given that almost all the participants came from a research environment, there is probably little real difference in background between those who indicated they used the language in their work and those who indicated they were researching the language. Moreover, the number of participants who had no experience of using DLs was too small to achieve a statistically significant result.

6.9. Effect of question position

As noted in section 6.2, each permutation of the study section order was used twice. This compensates for any effect of inter-section learning which would occur if the sections of the study were in a fixed order. In the latter case we might expect performance to be better on the second and third sections, after participants had gained experience. It also has an effect, to an extent, of randomizing the position of individual questions in the study. For example, the first question in section 6.4 (based on the componency pattern) will be in 1st position if that section is used first and 12th position if the section is used last; whilst if the section is used second the question will be in the 6th or 7th position, depending on which section is used first. When comparing the performance on each question, this will at least partially compensate for any effect of question position.

[109] In particular, the mean time of 1935 seconds for 'a little knowledge' of logic is the mean of 2619 seconds and 1250 seconds. The former is the longest of all the participants and the latter is one of the lower times. An indication of the wide discrepancy between these two values can be seen from the large standard deviation shown for the 'a little knowledge' of logic category in Table 6.12.

In principle, change in performance during the study might be positive or negative. The former because of gaining experience, the latter because of fatigue or boredom. In fact, a logistic regression of accuracy against position showed no significant effect ($p = 0.889$). Similarly, a regression of the log[110] of question response time against position showed no significant effect ($F(1, 250) = 0.5143$, $p = 0.474$).

The effect of position within a section was also investigated, for each of the three sections. In each case a logistic regression showed no significant dependence of accuracy on position within the section ($p >$ at least 0.1). Regression of log time against position within section showed a more varied picture. For the questions described in section 6.4, based on the componency pattern, there was a significant effect ($F(1,118) = 4.954$, $p = 0.028$), such that response times became longer moving through the section. There was no significant effect for the questions described in section 6.5, based on the modified coparticipation pattern ($F(1, 70) = 2.713$, $p = 0.104$), nor for the questions described in section 6.6, based on the modified types of entities pattern ($F(1, 58 = 3.925$; $p = 0.052$). Section 6.4 is an anomaly, not only in that there is a significant effect, but also in that the effect is to increase the response time with passage through the section. It may also be relevant that this section had the most questions. For each of the other sections, and overall, the regression analysis, although not significant, indicated a reduction in time. Since the questions were not randomized within each section, the effect of the order of questions may be being confounded with the effect of differences between the questions.

The effect of question position is investigated in a controlled way for study 3, described in Chapter 8. This is made possible because question position in that study was completely randomized. At this point, for study 1 we can say that there was no significant effect on accuracy caused by question position, neither overall nor within individual sections. For time the situation is confused and it is not possible to come to even a tentative conclusion.

6.10. Effect of number of reasoning steps

It seems likely that the number of reasoning steps to reach a valid conclusion will influence the probability of reaching that conclusion and the time taken to reach it. Indeed, Johnson-Laird et al. (1989) cite Oshershon (1975) as reporting a correlation between the number of steps and participants' accuracy. Section 6.10.1 investigates, for this study, what effect number of reasoning steps had on accuracy, whilst section 6.10.2 investigates the effect on reasoning time.

6.10.1. Accuracy and number of reasoning steps

It was hypothesized that the number of reasoning steps might have some effect on accuracy of response for the valid questions. Table 6-14 shows the mean percentage of correct responses for the valid questions, categorized by the required number of reasoning steps. A logistic analysis of deviance showed a significant dependence on number of steps ($p = 0.020$). However, a logistic regression did not give a significant result ($p = 0.469$), i.e. there was no indication of a trend. Table 6-14 shows that there was anomalously high accuracy on the questions with four steps, of which there were only two. A logistic regression with those two questions removed showed a significant decrease in accuracy with number of steps ($p = 0.038$). This suggests that accuracy of response does decrease with number of steps. To investigate this phenomenon in more detail would require comparison between questions with varying numbers of steps of equal complexity.

[110] As discussed in section 6.3, all statistical tests involving time which assume normality use log time.

Table 6-14 Percentage of correct responses to valid questions, categorized by number of reasoning steps

	number of reasoning steps				overall
	1	**2**	**3**	**4**	
no. questions	2	5	4	2	13
percentage correct	100%	87%	81%	96%	88%

6.10.2. Response time and number of reasoning steps

For those questions with valid conclusions which were correctly answered, it was hypothesized that the response time might be dependent on the number of reasoning steps required to arrive at the correct conclusion. Table 6.15 shows the mean time and standard deviation for correct valid responses divided into categories by number of steps. The table also shows the number of data points, i.e. the number of correct responses in each category.

Table 6-15 Mean time and standard deviation for correct valid responses, by no. of steps

	number of reasoning steps				overall
	1	**2**	**3**	**4**	
number of data points	24	52	39	23	138
mean time – secs	25.5	49.8	42.6	54.2	43.4
standard deviation - secs	25.0	26.7	30.3	30.5	29.2

Inspection of the mean times suggests, perhaps surprisingly, that whilst there is an appreciable difference between questions requiring one step and those requiring more, there is little difference amongst the latter. This is confirmed by an ANOVA of log time against number of steps, and subsequent Tukey HSD analysis. The ANOVA showed a significant dependence on number of steps ($F(3, 134) = 10.11$, $p < 0.001$). The Tukey HSD analysis showed a significant difference between the time for one step and for each of the other three categories ($p < 0.001$ for a comparison with two and four steps; $p = 0.001$ for three steps), whilst there was no significant pairwise difference between the other three categories ($p >$ at least 0.1).

In summary, the dominant features of the data are the difference between the questions which require only one reasoning step and those requiring more than one step, and the lack of difference between the latter, more complex, questions. This suggests that most of the time is not spent in executing the reasoning steps, but in ancillary activities. A certain amount of time, averaging around 25 seconds for these particular questions, is required for all questions. Then, for questions requiring more than one step, a very approximately similar amount of time is additionally required.

6.11. Response time – effect of validity, response and accuracy

Three factors which might be thought to influence response time are:

- whether the putative conclusion is valid or non-valid;
- the nature of the participant's response, i.e. valid or non-valid;
- the accuracy of that response.

These three factors are interdependent. Pick any two and the third will be determined. ANOVAs of log time against each of the three pairs of factors were conducted. Since these ANOVAs were unbalanced, the results potentially differ depending on the order of factors

in the analysis[111]. Hence two ANOVAs were conducted with each pair of factors, with the two different orders, making six ANOVAs in all.

The results of the ANOVAs are shown in Table 6-16. Note that, whilst the top-level F and p-values depend on the order of factors, the interaction effect is independent of their order (Shaw & Mitchell-Olds, 1993). Note also that all the top-level effects are significant, except for response when it follows validity. The interaction between accuracy and validity is also not significant; the other two interactions are significant. The two non-significant effects are shown in italics in the table. In summary, when we take any pair from the three factors, then both members of the pair are significant. In the one situation where this is not the case at the top-level, i.e. the factor response when it follows validity in the analysis, then the former factor has a significant interaction effect on the latter.

Table 6-16 Dependence of log time on validity, response and accuracy: result of ANOVAs

top-level effects				**interaction effect**
first factor		**second factor**		
validity	F(1, 248) = 17.833 p < 0.0001[112]	response	*F(1, 248) = 0.281* *p = 0.597*	F(1, 248) = 9.031 p = 0.003
response	F(1, 248) = 9.746 p = 0.002	validity	F(1, 248) = 8.368 p = 0.004	
validity	F(1, 248) = 17.833 p < 0.001	accuracy	F(1, 248) = 8.803, p = 0.003	*F(1, 248) = 0.508* *p = 0.477*
accuracy	F(1, 248) = 13.409 p < 0.001	validity	F(1, 248) = 13.228, p < 0.001	
response	F(1, 248) = 9.746 p = 0.002	accuracy	F(1, 248) = 11.981 p < 0.001	F(1, 248) = 5.418 p = 0.021
accuracy	F(1, 248) = 13.409, p < 0.001	response	F(1, 248) = 8.318 p = 0.004	

N.B. (1) Each row represents the result of a two-factor ANOVA. Columns 1 and 2 (i.e. the leftmost columns) show the first factor and the resultant data. Columns 3 and 4 show the second factor and resultant data. Column 5 shows the interaction effect, which does not depend on the order of factors. (2) Non-significant effects shown in italics

Table 6-17 shows the mean times for each of the four possible situations. In this case the two independent factors, validity and accuracy have been used. Any two of the three factors could have been used; the resultant values would be the same but arranged differently. A Tukey HSD was carried out after each of the ANOVAs to enable a pairwise comparison between each of the four quadrants. The results of the Tukey HSDs are identical for each of the ANOVAs. This is to be expected, given that the same six pairwise comparisons between the four quadrants are being made.

[111] In fact, there are a number of different approaches to multi-factor ANOVA. In one approach any difference in top-level effect due to order of factors is avoided by repeating the ANOVA with each factor in first position. The approach used here is the default 'R' method, which does give results depending on order. See Shaw and Mitchell-Olds (1993) for a discussion of unbalanced ANOVA.

[112] These F and p-values for validity are repeated in the table, under validity * accuracy. This occurs because the values associated with validity will be the same when it is used as the first factor, regardless of which of the other two factors is the second factor in the ANOVA. The same phenomenon occurs for the other two factors.

Table 6-17 shows that the correct responses to the questions with valid putative conclusions are, in general, the most quickly answered. On the basis of the Tukey HSD, these responses are significantly faster than the incorrect responses to the questions with non-valid conclusions ($p < 0.001$) and significantly faster than correct responses to the non-valid questions ($p = 0.002$), but not significantly faster than the incorrect responses to the valid conclusions ($p = 0.055$)[113]. Pairwise comparisons between the other three categories of responses shows no significant differences ($p >$ at least 0.3).

Table 6-17 Mean times (standard deviation, number of data points) for responses - secs[114]

	correct response	**incorrect response**	**correct and incorrect**
valid putative conclusion	43.4 (29.2, 138)	58.4 (27.5, 18)	45.1 (29.3, 156)
non-valid putative conclusions	59.5 (42.4, 72)	76.3 (48.8, 24)	63.7 (44.4, 96)
valid and non-valid putative conclusions	48.9 (35.1, 210)	68.6 (41.6, 42)	52.2 (36.9, 252)

In general, correct responses are significantly faster than incorrect responses. This is confirmed by a t-test comparison between the times for the correct and incorrect questions ($t(62.177) = 3.7946$, $p < 0.001$). This may be because the incorrect responses will be biased towards the harder questions, which are likely to take longer.

More interestingly, shown by the Tukey HSD, the correct responses to the non-valid questions took significantly longer than the correct responses to the valid questions. This may give some insight into how participants answer these questions. It is consistent with an approach in which participants attempt to prove the putative conclusion and, if they are unable to do so eventually give up and either decide the correct response is non-valid or seek to construct a counter-example. An alternative approach, of first seeking to construct a counter-example would be likely to lead, for the correct responses, to the non-valid conclusions being faster than the valid conclusions.

Even here, care must be exercised in interpreting the result. The valid and non-valid questions were not balanced for difficulty. It may be that the valid questions were easier and hence took less time to answer correctly than the non-valid questions. Indeed, overall the percentage of valid questions answered correctly was 88%, whilst for non-valid questions the percentage was 75%. Thus the valid questions appear easier than the non-valid, and this may have had an effect in reducing the response time for valid questions.

A better comparison of response times between valid and non-valid questions would be between questions of similar structure which were correctly answered by all participants. For example, in Table 6-7 question 1 is the only non-valid question correctly answered by all participants. Questions 2 and 3 are similarly structured valid questions, also correctly

[113] Note that the incorrect responses to the valid questions formed the smallest of the four categories, with 18 data points.

[114] Note that this table is consistent with the independence of the two factors. Since the ANOVA used log time, we should expect a multiplicative effect. For example, taking the correct responses to the valid questions (top left) as base, the multiplicative effect of moving in two dimensions to the incorrect response to the non-valid questions should be approximately equal to the product of each of the effects of movements in one dimension. This can be confirmed when the ratios of the mean times are computed:

1. incorrect responses to valid questions / correct responses to valid questions = 58.4/43.4 = 1.35
2. correct responses to non-valid questions / correct responses to valid questions = 59.5/43.4 = 1.37
3. incorrect responses to non-valid questions / correct responses to valid questions = 76.3/43.4 = 1.76

Then 1.35 x 1.37 = 1.85, i.e. around 5% greater than the figure computed in (3) above.

answered by all participants. The mean times for the two valid questions are both appreciably less than for the non-valid question. A paired t-test between question 1 and 2 shows that question 1 response times differed significantly from question 2 response times ($t(11) = 5.8163$, $p < 0.001$, paired test). Similarly question 1 response times differed significantly from question 3 response times ($t(11) = 4.1465$, $p = 0.002$, paired test). Questions 3 and 5 of Table 6-9 offer another opportunity to compare non-valid and valid questions. Again, both questions were correctly answered by all participants and they are relatively similar in structure. In this case, although the mean time for the non-valid question (question 3) is greater than for the valid question (question 5), the difference is not significant on a paired t-test ($t(11) = 0.72254$, $p = 0.485$)[115].

In summary, whilst by no means conclusive, there is evidence to support the hypothesis that participants start by attempting to prove the conclusion, rather than attempting to show that it need not follow from the premises. A more conclusive study would require carefully designed questions, on the one hand representative of a range of reasoning tasks and on the other hand paired to ensure that, as far as possible, the valid and non-valid questions were of equal difficulty.

Finally, when considering participants' responses to the questions with non-valid conclusions, it should be noted that in only one case is it possible to prove the negation of the conclusion. This is question 3 in Table 6-9[116]. For the other questions with non-valid conclusions, the most one can do is to prove that the negation of the conclusion is consistent with the axioms, e.g. by creating a counter-example.

6.12. Discussion

This chapter has described a preliminary study, investigating participants' difficulties with common DL constructs in the context of commonly-used content patterns. Five difficulties have been highlighted, reviewed in subsection 6.12.1. Besides these difficulties, some other factors influencing participants' performance have been investigated. These are reviewed in subsection 6.12.2. Finally, subsection 6.12.3 discusses threats to the validity of the study.

6.12.1. The difficulties

One difficulty may have been partly caused by the assumption that if a property is transitive, then this will also be true of its subproperty. This may in part be a result of some participants having a background in object-oriented design and programming. However, it may also be more fundamental, relating to the normal usage of the prefix 'sub' in English. A subspecies, for example, is a specialization of species. Any animal in a subspecies will possess the characteristics of the species, plus certain other characteristics specific to the subspecies.

In general, the entities in a subclass possess the characteristics of entities in the superordinate class, since they are themselves members of that superordinate class[117]. Whereas the instantiations of a subproperty do not in general possess the characteristics of the superproperty. In retrospect, the choice of the keyword *SubPropertyOf*, whilst perfectly

[115] The mean time for question 3 is influenced by two outliers, as is indicated by the relatively large standard deviation. Removal of just the largest outlier (121 seconds) reduces the mean time for question 3 to 36.5 seconds.

[116] The reasoning here is that *is_participant_in*, being the inverse of *has_participant*, must have domain *Object* and range *Event*. Hence, from the second axiom, *C* must be in *Object*. However, for the conclusion to hold, *C* must be in the range of *is_participant_in*, i.e. *Event*. Since *Object* and *Event* are disjoint, *C* cannot be in both, and the conclusion cannot hold.

[117] For example, restrictions are inherited. If *A*, *A1* and *B* are classes, with *A1* a subclass of *A*, and *P* is some object property, then *A SubClassOf P {some | only} B* implies *A1 SubClassOf P {some | only} B*.

logical, may be misleading because of the prefix *sub*. It may be that the hierarchical concept of subproperties is confusing and unnecessary. If *S* is a subproperty of *P*, then this means precisely that *a S b* implies *a P b*[118]. An alternative way of specifying this might be to use a keyword which avoids the implicature of hierarchy, e.g. *S Implies P*[119]. Looked at this way, a subproperty hierarchy is no more than a chain of inferences. Indeed, in everyday English, whilst we talk about subclasses and subsets, we rarely talk about subproperties; e.g. 'being a sister is a subproperty of being a sibling' sounds much less natural than 'being a sister implies being a sibling'. In practice, more attention needs to be given in training to the relation between property characteristics and subproperty characteristics.

The use of the existential restriction to define a class caused a difficulty. The use of restrictions to define classes is central to DLs but quite unfamiliar to those coming from a database background. A great deal of emphasis, and indeed repetition of examples, is needed during training. Diagrammatic representation of DLs might also help, as in the work of Stapleton et al. (2014). The existential restriction, and also the universal restriction, are further investigated in the two subsequent studies.

The use of a functional object property also caused a difficulty. The context of the question was relatively complex, and it is not possible on the basis of this study to determine to what extent the problem is inherent to functionality and to what extent it was caused by question complexity. This is examined in the next study, whilst the final study also investigates inverse functionality.

The interaction of conjunction and disjunction, also caused some participants a difficulty, although not to the same extent as the previously discussed problems. The ambiguity of *and* may have contributed here, as may have the complexity of the question. The use of the keywords *and* and *or* is discussed in more detail in Chapter 8.

The greatest difficulty was experienced with negated conjunction, where the accuracy was less than chance, albeit not significantly so, and significantly less than for negated disjunction. This agrees with the findings of Khemlani et al. (2012a), who were working with study participants not trained in logic. They attributed the problem to not creating the necessary set of mental models. This may well be the case. The ambiguity of *and* may have been a factor, perhaps contributing to the failure to create all the necessary mental models.

6.12.2. Other factors

Participants' previous knowledge of logic and DL, as declared by the participants themselves, did affect performance. Knowledge of DL was both significantly positively correlated with response accuracy and negatively correlated with response time, whilst knowledge of logic was significantly positively correlated with accuracy but not significantly correlated with response time. However, the very uneven distribution of participants in the categories for knowledge of logic, and the high correlation for these participants between knowledge of logic and knowledge of DL, means that these results must be treated with caution. The nature of the sample of participants did not permit any conclusions as to the effect of experience of DL usage on performance.

---

[118] See Motik et al. (2012) which states "if an individual x is connected by OPE1 [the subproperty] to an individual y, then x is also connected by OPE2 [the superproperty]". They do state that "Object subproperty axioms are analogous to subclass axioms" without specifying the extent of the analogy.

[119] Indeed Motik et al. (2012) use the word 'implies' in an English expansion of a subproperty statement:
"SubObjectPropertyOf(a:hasDog a:hasPet) Having a dog implies having a pet."

The position of a question might be expected to have an effect on the accuracy with which it is answered and on the response time. In fact, overall position of a question within the study had no significant effect on either accuracy or response time. Similarly, the effect of position within each section had no significant effect on accuracy. The effect of question position within each section had an effect which varied between the sections and was only significant for one section. However, these results also need to be interpreted with caution as the position of the questions was not fully randomized.

For the valid questions, an increasing number of reasoning steps appears to lead to reduced accuracy of response; although a more controlled experiment is required to verify this. An analysis of time for the correct responses to valid questions showed that questions which only required one reasoning step were answered significantly faster than those requiring more than one step. On the other hand, there was no significant difference in response time between questions requiring two, three and four steps. The number of questions was relatively small. However, this does suggest that an appreciable amount of time is spent in a (perhaps trial and error) process of arriving at the reasoning strategy.

Finally, response time is significantly dependent on the three interdependent factors: accuracy, validity of putative conclusion, and nature of response. Mean response time is significantly less for the correct questions than for the incorrect ones. This could be because the incorrect responses are biased towards the harder questions. Mean response time for correct valid responses is significantly less than for correct non-valid responses. This suggests that, in general, participants are attempting to prove the putative conclusion and providing a response of non-valid if they are unable to do so in a reasonable time.

6.12.3. Threats to validity

As discussed in Section 4.8, threats to validity are divided into three categories, each of which is discussed here.

Threats to construct validity

Subsection 4.8.1 has already noted that the responses were recorded automatically. The measurement of response time was potentially more error prone, since it was done manually. Screen activity was recorded by *Camtasia* and then reviewed by the experimenter. The timepoints were taken from the time data displayed by *Camtasia* at the bottom of the screen. In principle, this enabled timepoints to be measured to the nearest second, i.e. ±0.5 second, and hence response times to be calculated ± 1 second. In practice, there might have been some difficulty in recording the precise timepoint, giving an error of ±1 second. This would lead to a potential error of ±2 seconds in the calculation of response times. This should be compared with a minimum response of 9 seconds; most response times were much greater than this.

Threats to internal validity

Subsection 4.8.2 noted that the major threats to internal validity for each of the studies were inter-and intra-section effects caused by, e.g. learning or fatigue. The permutation of section order described in Section 6.2 was designed to compensate for any inter-section effect and for any overall effect of position in the study. However, the position of questions in each section was fixed and thus there was no compensation for any intra-section effect. Inspection of Tables 6-7, 6-9 and 6-11 shows no obvious effect on accuracy of position within section. However, in Tables 6-7 and 6-9, the first question had the second longest mean response time, whilst in Table 6-11 the first question had the longest mean time. Indeed, inspection

of Tables 6-9 and 6-11 suggests that there may be a tendency for response time to decrease as the participant moves through a section. In study 2, an attempt was made to compensate for this, as discussed in Section 7.1.

Threats to external validity

Subsection 4.8.3 noted that the generalisability of the study may be limited by two factors: the nature of the questions and the profile of the participants. As described in Section 6.1, the questions were designed to be ecologicaly valid, i.e. to employ the kind of constructs actually used, and in a realistic context. In this way, it was intended to identify difficulties which had a real effect on DL users.

All the participants had some background in computer science, as noted in Section 6.2. This is also illustrated in Table 6-12, which shows that all the participants had at least a little knowledge of logic and a little knowledge of OWL or another DL formalisms, and the great majority of participants had some knowledge of both logic and DL. Consequently, the participants were not representative of the kind of domain experts who might have a limited background in computing. To an extent, the second study compensates for this, as will be described in the next chapter.

7. Further investigations – controlled comparisons

In seeking to ascertain the 'laws of association of ideas', which are psychological 'laws of thought', the psychologist may find the fallacies into which the average human mind is prone to fall an even more instructive study than the rigidly correct intellectual processes of the soundest scientific thinker.

D. G. Ritchie, 'The Relation of Logic to Psychology', 1896

The previous chapter identified a number of particular difficulties which people experience in understanding and reasoning with DLs. This chapter investigates in more detail three of those areas of difficulty: functional object properties; negated conjunction; and existential and universal restrictions.

The study described in the previous chapter presented DL features in the context of commonly used ontology patterns, in the interest of ecological validity. The study described in this chapter has been designed to enable comparison between DL features, e.g. between functionality and transitivity in object properties. Nonetheless, the DL features investigated are those commonly used by ontology developers, as identified by the survey of Chapter 5 and the work cited in Chapter 6.

The first section of this chapter describes the organisation of the study. Section 7.2 then describes the statistics of the response time data and justifies the use of the log transformation, as in Chapter 6. The following four sections describe each of the four sections of the study. Section 7.3 is concerned with further investigation into the difficulty with functional object properties. Section 7-4 looks again at negated disjunction and conjunction. Sections 7.5 and 7.6 are concerned with restrictions. The former looks at the interaction between negation and restriction; the latter at nested restrictions. Section 7.7 reports on participant feedback, providing further insight into the difficulties experienced. Section 7.8 describes the effect on performance of the self-declared level of knowledge of the participants. Section 7.9 looks at the effect of question position on accuracy and response time. Section 7.10 investigates the effect on response time of three interrelated factors: validity of the putative conclusion; the nature of the participant's response; and the accuracy of that response. Finally, Section 7.11 draws some conclusions. The study's main findings are also described in Warren et al. (2015).

7.1. Organisation of the study

Two sets of participants were used, under different conditions. For the first set, the study was administered under laboratory conditions, using the same approach as for the first study, described in Chapter 6. Then, after the laboratory study was complete, an online study was administered to enable additional participants to take part and thus obtain further data. The analyses of response time described in later sections make use of the data from the laboratory study. The analyses of accuracy make use of the data from the laboratory study, enhanced where possible with data from the online study.

7.1.1. The laboratory study

The *SurveyExpression* tool was used to display the questions and record the responses. *Camtasia* was run on the PC to record screen activity and the recordings were analysed to obtain timing data. As with the first study, participants were drawn from the Knowledge Media Institute and the Centre for Research in Computing at the Open University. In addition, there were participants from other universities and research institutes who were

identified through their membership of the UK Ontology Network[120], plus participants from industry[121]. All the participants had experience of working with ontologies. Almost all the participants were primarily computer scientists; a few had a background in biology and could be regarded as bioinformaticians. Most of the sessions took place using the author's computer and with the author and participant physically co-located. In a few cases the sessions were held remotely. The participant accessed the *SurveyExpression* pages on the participant's own computer and a Skype link was set up with the author's computer. This enabled *Camtasia* to be run on the author's computer so that a recording could be made for later analysis to obtain the timing data.

As there were four question sections, the intention was to obtain 24 participants so that each permutation of section order would be represented once. For technical and other reasons, on four occasions there was a failure to obtain a screen recording of the whole session. Consequently, only accuracy data was available for four of the participants. As a result, there were 24 participants for whom accuracy and timing data were available and another 4 for whom only accuracy data were available. Thus all the analysis of timing data for the laboratory study was based on 24 participants, whilst the analysis of accuracy data was based on 28 participants from the laboratory study enhanced by additional participants from the online study as described below.

The permutation of section order was intended to compensate for inter-sectional effects and to partially randomize the position of questions in the study. In addition, two mutually reverse orderings of questions within each section were also used, to mitigate intra-sectional effects and further randomize the overall position of questions in the study.

As with study 1, there was a preliminary section where participants were invited to provide information about themselves, e.g. their knowledge of logic, and a final section where participants could provide feedback. A five-page handout, similar to the handout used in study 1, contained all the necessary information about the syntax and semantics of the subset of MOS used in the study. The handout was provided to participants to read before beginning the study and was available for reference during the study.

7.1.2. The online study

As discussed in Chapter 8, it is difficult to obtain statistical significance, even for quite large effects, with the relatively modest participant sizes available. This is particularly the case for the accuracy data. Consequently, the online study was undertaken to obtain more accuracy data. With a few exceptions, described in the appropriate sections below, the study used exactly the same questions as the laboratory study. The online study was launched after the latter was completed so that no participants in the laboratory study were able to view the questions before taking part.

Because it was felt that remote online participants might not be prepared to spend as long on the study as laboratory participants, the online study was divided into each of the four sections, which could be undertaken separately. As a result, not all the sections were completed. Only data from completed sections was used and consequently the number of additional data points varied from 12 for the questions discussed in Section 7.4 to 37 for the questions discussed in Section 7.3.

[120] https://groups.google.com/forum/#!forum/ontology-uk

[121] Specifically, several from BT (http://www.btplc.com/) Research Labs at Ipswich, U.K., and one from iSOCO (http://www.isoco.com/en/) in Madrid, Spain.

As noted in the previous subsection, in the laboratory study a certain degree of randomization was achieved by permuting the order of the sections and presenting the questions within each section in two different orders. In the online study, the *surveygizmo*[122] web tool was used, which enabled randomization of the question order.

At the beginning of each of the study sections, participants were invited to download the handout, which was exactly as used for the laboratory study.

Whilst the online study was a useful way of increasing the sample size for the accuracy data, the data might not be of the same quality as the laboratory data. In particular, although participants were requested not to use pen and paper, there was no control on this. The fact that there were a number of partially completed responses, i.e. that some participants had given up, might imply that the completed responses were biased towards those participants with more facility in this kind of reasoning. Although the analysis is chiefly concerned with comparisons between questions, the self-selecting nature of the participants might still introduce a bias.

7.1.3. The questions

There were ten questions in the section relating to object properties, described in Section 7.3. All the other sections contained eight questions, making a total of 34. As with the previous study, each question consisted of a set of axioms and a putative conclusion. Participants were required to state whether the conclusion was valid or non-valid. For all the questions with non-valid putative conclusions used in this study, it is not possible to prove the converse, i.e. the conclusion is consistent with the axioms. For the questions in Section 7.4, consistency does require that certain classes be empty.

7.2. Response time data

A Shapiro-Wilk test on the response time data available from the online study gave a p-value of less than 2.2×10^{-16}, whilst a plot of the data revealed a positive skew. Consequently, the data was transformed using transformations from Tukey's ladder of powers. The resultant Shapiro-Wilk p-values are shown in Table 7-1. This suggests that, to use tests such as ANOVA which require approximate normality, a log transformation would be appropriate.

Table 7-1 Shapiro-Wilk test applied to response time data

Transformation	**p-value**
time (i.e. identity transformation)	$< 2.2 \times 10^{-16}$
square root (time)	2.8×10^{-14}
log (time)	0.11
1 / square root (time)	$< 2.2 \times 10^{-16}$

However, as was noted in section 6.3, since comparisons are generally between questions, it is important that response time data for each question be approximately normal. Table 7-2 shows the minimum p-value across all questions for each of the relevant transformations. This minimum p-value is greatest for the log transformation. Moreover, although the value shown (0.0077) is relatively small, there was only one other value below 0.05 across the 34 questions. In fact, the log transformation resulted in the other 32 Shapiro-Wilk p-values

[122] https://www.surveygizmo.co.uk/

being greater than 0.1. Consequently, the log transformation has been applied to response time data prior to all statistical tests in this chapter.

Table 7-2 Minimum p-value for Shapiro-Wilk test applied to response time data across all questions

Transformation	**p-value**
time (i.e. identity transformation)	2.5×10^{-7}
square root (time)	2.5×10^{-5}
log (time)	0.0077
1 / square root (time)	0.00019

7.3. Functional object properties

In the first study, participants experienced difficulty with a question involving a functional object property. As explained in subsection 6.6.2, the question had a relational complexity of four and it was not clear whether the difficulty arose entirely from the question's complexity or whether participants were experiencing an inherent difficulty with functionality. Functionality might, for example, be inherently more difficult than transitivity, which was shown in Chapters 5 and 6 to be another commonly used object property characteristic.

The question in the first study required a reasoning step of the form:

$$a\ F\ b;\ c\ F\ d;\ b\ \mathit{DifferentFrom}\ d \Rightarrow a\ \mathit{DifferentFrom}\ c \qquad (1)$$

Here, *F* is a functional object property and *a*, *b*, *c*, *d* are individuals. The step has RC 4 because it requires the concurrent attention to four individuals.

Another inference involving functionality is:

$$a\ F\ b;\ a\ F\ c \Rightarrow b\ \mathit{SameAs}\ c \qquad (2)$$

This has RC 3 because it only requires the concurrent attention to three individuals. If there is no inherent difference in difficulty between functionality and transitivity, then we might expect this reasoning step to display the same difficulty as the following one, where T represents a transitive object property:

$$a\ T\ b;\ b\ T\ c \Rightarrow a\ T\ c \qquad (3)$$

This also has RC 3 because it requires the concurrent attention to three individuals.

These two RC 3 inferences are the basis of the comparison between functionality and transitivity in this study. Because these two inferences are, individually, relatively easy, it was thought that human reasoners might perform so well on them both as to make discrimination difficult. Therefore, an inference was used which involves functionality and requires two applications of the RC 3 reasoning step:

$$a\ F\ b;\ a\ F\ c;\ b\ F\ d;\ c\ F\ e \Rightarrow d\ \mathit{SameAs}\ e \qquad (4)$$

The first two axioms imply that *b* is the same as *c* and the third and fourth axioms imply that *d* is the same as *e*. Note that between the two RC 3 reasoning steps, there is an additional step in which *c* is substituted for *b* in *b F d*, or *b* is substituted for *c* in *c F e*. In either case, this means the inclusion of a reasoning step which only requires concurrent manipulation of two individuals (the individual being substituted and the individual being replaced) and

hence has RC 2. Thus the inference consists of three steps, with RC 3, 2 and 3. This inference is shown in question 3 of Table 7-3 below. The table shows all the eight question used in this section of the study, whether the putative conclusion is valid and, for the questions with valid putative conclusions, the relational complexity of each step in the reasoning. Note that, for brevity, *T* represents a transitive object property, and *F* represents a functional object property. In the actual questions, *greater_than_or_equal_to* was used for the former, *has_nearest_neighbour* for the latter. However, as with the previous study, it was made clear, e.g. in the handout, that the choice of names used should not affect any conclusions to be drawn. Note also that, as with subsequent tables in this and the following chapter, the table does not show the necessary declarative statements which were included in the questions. Moreover, in places the tables use a semicolon to delimit the axioms. In the actual questions, each axiom was on a separate line.

Table 7-3 Questions employing transitive and functional object properties

No.	**axioms**	**putative conclusion**	**valid / non-valid**	**RC**
1	a T b; b T c; c SameAs d; d T e	a T e	valid	3,2,3
2		d T b	non-valid	n/a
3	a F b; a F c; b F d; c F e	d SameAs e	valid	3,2,3
4		a SameAs e	non-valid	n/a
5	a F b; a F c; d F b; e F f; c DifferentFrom f	d DifferentFrom e	valid	3,2,4
6		a DifferentFrom d	non-valid	n/a
7	a F b; c F d; b Differentfrom d; e F a; f F g; c SameAs g	e DifferentFrom f	valid	4,2,4
8		a DifferentFrom f	non-valid	n/a

Question 1 of Table 7-3 was designed to mirror question 3, this time using transitivity. The first two axioms imply that *a T c*, constituting a reasoning step of RC 3, similar to inference (3) above. The next axiom, *c SameAs d* is deliberately included to mirror the reasoning step of RC 2 in question 3. The axiom leads to *d* being substituted for *c* in *a T c*, giving *a T d*. The final axiom then gives *a T e*. Thus we arrive at a chain of reasoning steps, with the same relational complexity (3, 2, 3) as question 3. This enables a controlled comparison between functionality and transitivity.

Questions 2 and 4 have the same axioms as questions 1 and 3, but non-valid putative conclusions. Since there is no valid reasoning chain associated with these questions, there are no associated RC values.

The remaining questions investigate the effect of increasing complexity on reasoning with functionality. Question 5 starts with the same two axioms as question 3, leading to *b* being the same as *c*, in a reasoning step of relational complexity 3, isomorphic to inference (2) above. This, together with the last axiom (*c DifferentFrom f*) means that *b DifferentFrom f*; this is arrived at in a reasoning step of relational complexity 2 by substituting *b* for *c* in the last axiom. This gives *d F b; e F f; b DifferentFrom f*, which in turn leads to *d DifferentFrom e* in a reasoning step of RC 4, isomorphic to the inference (1) above. Thus, question 5 comprises three reasoning steps of RC 3, 2, 4.

In a similar way, question 7 uses two reasoning steps of RC 4, both isomorphic to inference (1) above, and a middle step, substituting *g* for *c*, of RC 2. This gives a chain of RC 4, 2, 4. Thus questions 5 and 7, along with question 3 already discussed, employ functional object properties to create questions of increasing complexity, enabling an investigation of the

effect of such increasing complexity. Finally, questions 6 and 8 have the same axioms as questions 5 and 7 but with non-valid putative conclusions.

For each of the questions with a non-valid putative conclusion, the axioms and the conclusion are consistent, i.e. it is not possible to prove the converse of the conclusion, given the particular characteristic associated with the object property. For example, in question 2 if we assume that *T* corresponds to *has_sibling* then we can see that the axioms and the conclusion can all be true. The failure of the axioms to infer the conclusion can be seen by the construction of a counter-example. For example, if we take *T* to be *has_descendant*, this satisfies the transitivity requirement. However, it is clear from the second and third axioms that *b T d*, i.e. *b has_descendant d*. Consequently, we cannot also have the conclusion, *d has_descendant b*.[123]

To summarise the objectives of this section, there are two research hypotheses:

> H7.1 Reasoning about functionality is inherently more difficult than reasoning about transitivity, after controlling for relational complexity.
>
> H7.2 Reasoning about functionality becomes increasingly difficult with increasing relational complexity.

Here, 'difficulty' is measured by percentage of correct responses and time to respond.

The following two subsections provide more information on the laboratory and online studies, whilst the rest of the section discusses the hypotheses.

7.3.1. Functional object properties – the laboratory study

For this section, the questions were presented in two mutually reverse orders. Table 7.4 shows the orders used, identifying the questions by the numbering used in Table 7-3. In each ordering, all four different axioms were used for the first four questions, and then repeated for the second four, but in such a way that there are never more than two valid or two non-valid questions together. Moreover, each pair of questions with identical axioms are separated by three other questions. Of the 24 participants for whom response time information was available, 12 saw the questions in one order, 12 in the other order. Of the additional four participants for whom only accuracy information was available, three saw the questions in order 1 and one saw the questions in order 2. This strategy of dividing the participants between two mutually reverse orderings was also used in the other three sections, with participants assigned to order 1 and order 2 consistently in each section.

[123] To generalize, if we take *T* to be symmetric, as well as transitive, e.g. *has_sibling*, we can prove the conclusion in question 2. If we take *T* to be antisymmetric, e.g. *has_descendant*, we can prove the converse of the conclusion. With only the assumption of transitivity, the conclusion is simply consistent with the axioms.

Table 7-4 Functional object properties - question ordering

order 1			**order 2**	
question no.	valid / non-valid		question no.	valid / non-valid
2	non-valid		7	valid
3	valid		6	non-valid
5	valid		4	non-valid
8	non-valid		1	valid
1	valid		8	non-valid
4	non-valid		5	valid
6	non-valid		3	valid
7	valid		2	non-valid

The results of the laboratory study are shown in Table 7.5. Note that, with the exception of the two questions relating to transitivity, for each pair of questions with the same axioms, the non-valid questions were answered more accurately than the valid questions. Moreover, this difference in accuracy increases with increasing complexity. This may be indicative of a bias towards responding with 'non valid' when confronted with a difficult question. Note also that question 5 was not answered significantly better than chance ($p = 0.172$, one-sided test). Question 7 was answered worse than chance, although not significantly so ($p = 0.286$, one-sided test). All other questions were answered significantly better than chance ($p < 0.05$, one-sided test). Thus the two questions with valid putative conclusions and involving functionality with RC 4 were not answered significantly better than chance.

Table 7-6 shows the data separately for the two orderings. The two questions which were answered first were questions 2 and 7. In the case of question 7, there is an appreciable difference between the mean response times for the two orderings, suggesting that when it was posed first it incurred a considerable time penalty. This is not the case for question 2, possibly because it is an appreciably easier question. In fact, the mean time for question 2 when it was the first question is slightly less than when it occurred last. There also seems to be an appreciable penalty for question 6 when it occurred in the second position. Again, there seems no such penalty for the easier question 3 which also occurred second.

Table 7-5 Questions employing transitive and functional object properties

No.	**axioms**	**putative conclusion**	**valid / non-valid**	**RC**	**%age corr N=28**	**mean time (s.d.) – secs N = 24**		
						overall	**corr**	**incorr**
1	a T b; b T c; c SameAs d; d T e	a T e	valid	3,2,3	96%*	34 (14)	34 (14)	40 (n/a)
2		d T b	non-valid	n/a	86%*	48 (34)	42 (14)	87 (95)
3	a F b; a F c; b F d; c F e	d SameAs e	valid	3,2,3	75%*	52 (36)	56 (39)	36 (9)
4		a SameAs e	non-valid	n/a	96%*	61 (46)	62 (47)	35 (n/a)
5	a F b; a F c; d F b; e F f; c DifferentFrom f	d DifferentFrom e	valid	3,2,4	61%	84 (67)	90 (66)	73 (70)
6		a DifferentFrom d	non-valid	n/a	79%*	92 (66)	86 (62)	111 (80)
7	a F b; c F d; b Differentfrom d; e F a; f F g; c SameAs g	e DifferentFrom f	valid	4,2,4	43%	109 (79)	96 (55)	119 (96)
8		a DifferentFrom f	non-valid	n/a	71%*	96 (47)	93 (50)	101 (43)

N.B. (i) In this and subsequent similar tables in this chapter, the overall mean time for each question may not correspond to the weighted sum of the times for the correct and incorrect responses, using the percentage correct given, since the latter is calculated on the basis of a larger sample.
*(ii) * answered significantly better than chance ($p < 0.05$)*

Table 7-6 Object property questions - performance broken down by ordering

No.	**valid / non-valid**	**position**		**%age correct**		**mean time (s.d.) - secs**	
		order 1	**order 2**	**order 1 N = 15**	**order 2 N = 13**	**order 1 N = 12**	**order 2 N = 12**
1	valid	5	4	100%	92%	28 (11)	40 (14)
2	non-valid	1	8	87%	85%	47 (16)	49 (47)
3	valid	2	7	67%	85%	49 (33)	54 (39)
4	non-valid	6	3	100%	92%	42 (26)	79 (56)
5	valid	3	6	47%	77%	80 (79)	88 (55)
6	non-valid	7	2	73%	85%	73 (65)	111 (64)
7	valid	8	1	40%	46%	86 (69)	132 (85)
8	non-valid	4	5	73%	69%	91 (47)	101 (48)

7.3.2. Functional object properties – the online study

There were 37 completed responses to this section obtained from the online study. The questions were identical to those used for the laboratory study except for a slight change to questions 1 and 2, i.e. the two questions which employ transitivity. The revised questions are shown in Table 7-7. In question 1 in the laboratory study the two transitive reasoning steps were such that the shared term was at the end of the first axiom and the beginning of the second (i.e. a T b; b T c). In question 1 in the online study, the shared term is at the beginning of the first axiom and the end of the second (a T b; c T a). Question 2 in Table 7.6 uses the same axioms as question 1, but with an invalid putative conclusion.

It was hypothesized that this revised order might seem less natural and be more difficult for participants. Specifically:

H7.3 Reasoning about transitivity with the shared term at the end of the first axiom and beginning of the second axiom will lead to more accurate responses than when the shared term is at the beginning of the first axiom and end of the second axiom.

Comparisons of the two versions of questions 1 and 2 enable this hypothesis to be tested. However, a caveat should be made regarding the different invalid putative conclusions in the two versions of question 2. No systematic process was available for creating invalid conclusions of equal 'difficulty' and the effect of the different order of axioms may be confounded with differing difficulties associated with the putative conclusions.

Table 7-7 Revised questions 1 and 2 for online study

No.	**axioms**	**putative conclusion**	**valid / non-valid**	**RC**
1	a T b; c T a; b SameAs d; e T c	e T d	valid	3,2,3
2		a T c	non-valid	n/a

Table 7.8 shows the percentage of correct responses for the online study and the two studies combined.

Table 7-8 Percentage of responses correct for laboratory study, online study, and combined studies

No.	**valid / non-valid**	**object property characteristic**	**RC**	**percentage correct**	
				online study N = 37	**combined studies N = 65**
1	valid	transitive	3,2,3	95%*	95%*
2	non-valid	transitive	n/a	95%*	91%*
3	valid	functional	3,2,3	78%*	77%*
4	non-valid	functional	n/a	95%*	95%*
5	valid	functional	3,2,4	57%	58%
6	non-valid	functional	n/a	86%*	83%*
7	valid	functional	4,2,4	73%*	60%
8	non-valid	functional	n/a	54%	62%*
TOTAL				79%	78%

** question answered significantly better than chance ($p < 0.05$)*

7.3.3. Ordering of axioms in transitive inference – H7.3

It is useful to start by considering hypothesis H7.3, since this clarifies the difference between the laboratory and the online study. It is clear from inspection that there is no significant difference between the accuracy of responses to question 1 in the two forms. For question 2, the difference is contrary to the expected direction, but is in any case not significant ($p = 0.390$, Fisher's Exact Test, two-sided).[124]

In summary, there is no support for hypothesis H7.3. The ordering of the axioms in the transitive inference in question 1 makes no significant difference to accuracy or response time.

[124] In fact, with the exception of question 7, there is no significant difference between the proportion of correct responses in the laboratory and the online situations ($p > 0.1$ in all cases). For question 7, the difference is significant ($p = 0.021$, Fisher's Exact Test, two-sided). There was also no significant difference between the two studies when the data was aggregated across all questions ($p = 0.397$, Fisher's Exact Test, two-sided).

Because of this lack of significant difference between questions 1 and 2 in the two studies, hypotheses H7.1 and H7.2 will be analysed using the combined accuracy data from both studies.

7.3.4. Functionality and transitivity – hypothesis H7.1

Comparison of questions 1 and 3 enables hypothesis H7.1 to be investigated. The difference in percentage correct responses between the two questions is significant ($p = 0.004$, Fisher's Exact Test, two-sided). Moreover, the difference in time between the two questions is significant ($t(23) = 3.0843$, $p = 0.005$, paired test). These results are consistent with hypothesis H7.1, i.e. reasoning about functionality is harder than reasoning about transitivity, after controlling for complexity.

Comparisons between questions with non-valid putative conclusions are less clear-cut to interpret. In particular, we do not have a defined set of reasoning steps, with known complexity, which participants need to follow. As suggested in Chapter 6, many participants are likely to arrive at a non-valid response by failing to prove validity. Other participants may construct a counter-example. The difficulty of constructing such counter-examples may or may not vary appreciably across the questions and may or may not be correlated with the difficulty of answering the analogous questions with valid putative conclusions.

Inspection of Table 7-8 shows that question 4, the non-valid question employing functionality, was actually answered more accurately than question 2, which employed transitivity, counter to hypothesis H7.1. However, this difference was not significant ($p = 0.492$, Fisher's Exact Test, two-sided). The mean response time for question 4 was greater than for question 2. However, this difference was also not significant ($t(23) = 1.262$, $p = 0.220$, paired test).

In summary, as regards hypothesis H7.1, the laboratory study confirmed that reasoning about functionality is more difficult than reasoning about transitivity when required to reach a valid conclusion. However, when required to refute a non-valid conclusion, there was no evidence of any difference in difficulty between functionality and transitivity.

7.3.5. Increasing complexity – hypothesis H7.2

Comparison of questions 3, 5 and 7, all with valid putative conclusions and of increasing relational complexity, enable hypothesis H7.2 to be investigated. A logistic analysis of deviance reveals a significant variation in proportion of correct responses across the questions ($p = 0.044$). A subsequent Tukey HSD analysis showed no significant pairwise differences (question 3 versus 5: $p = 0.067$; question 5 versus 7: $p = 0.983$; question 3 versus 7: $p = 0.099$). A logistic analysis of deviance for questions 4, 6 and 8, with non-valid putative conclusions also shows a significant variation ($p < 0.001$). A subsequent Tukey HSD analysis showed no significant difference between questions 4 and 6 ($p = 0.082$), a significant difference between questions 6 and 8 ($p = 0.019$) and a significant difference between questions 4 and 8 ($p < 0.001$).

Care needs to be exercised in interpreting the response time data. As already noted, question 7 was the first question to be answered by half the participants. Table 7.6 suggests that this had an appreciable effect on mean response time. As a result, it is not valid to analyse the times for question 3, 5 and 7, e.g. using ANOVA, on the full set of 24 data points. One alternative is to perform an ANOVA on data from the 12 participants who saw question 7 last, i.e. in order 1. This did not give a significant result ($F(2, 33) = 1.889$, $p = 0.167$). However, using the full dataset there was a significant difference in response time between

questions 3 and 5 ($t(23) = 3.0226$, $p = 0.006$, paired test)[125]. This does indicate a significant difference in reasoning about functionality between a situation of RC 3 and a situation of RC 4.

When comparing the mean response time for questions 4, 6 and 8, the problem arises that question 6 was answered second by half the participants and, from Table 7.6, appears to have incurred a time penalty which may distort the overall mean response time. To avoid this problem, an ANOVA was performed on data from the participants who saw the questions in order 1. This revealed that the response time varied significantly across the questions ($F(2, 33) = 4.8097$, $p = 0.015$). A subsequent Tukey HSD analysis revealed that there was a significant difference between questions 4 and 8 ($p = 0.011$) but no significant difference between questions 4 and 6 ($p = 0.195$) and between questions 6 and 8 ($p = 0.394$).

In summary, the data confirms hypothesis H7.2. Complexity significantly reduces accuracy and significantly increases response time.

7.3.6. Conclusions

Table 7-9 summarises the conclusions relating to hypotheses H7.1, H7.2 and H7.3. Confirming a valid inference was significantly less accurate and took significantly longer with functionality than with transitivity. On the other hand, there was no significant difference in accuracy or response time between transitivity and functionality when refuting a non-valid inference. Increasing complexity significantly reduced accuracy and increased response time; this effect was discernible in both valid and non-valid questions. Finally, changing the order of the axioms in a transitive inference has no effect on the accuracy of responses.

[125] Note that question 3 occurred second for half the population and seventh for the other half; question 5 occurred third and sixth. For both questions, the difference between the mean response times for the two orderings was relatively small, i.e. compared to the difference overall between the questions.

Table 7-9 Object properties – summary of hypotheses

		accuracy	**response time**
H7.1 – functionality vs. transitivity	valid conclusions	functionality significantly less accurate	reasoning about functionality takes significantly longer
	non-valid conclusions	no significant difference	no significant difference
H7.2 – effect of relational complexity on reasoning about functionality	valid conclusions	complexity significantly affects accuracy overall; no significant pairwise differences	significant increase in response time between RC 3, 2, 3 and RC 3, 2, 4
	non-valid conclusions	complexity significantly affects accuracy overall; significant pairwise differences between RC3, 2, 3 / RC4, 2, 4 and RC3, 2, 4 / RC4, 2, 4	complexity significantly affects response time overall; significant pairwise difference between RC3, 2, 3 / RC4, 2, 4
H7.3 – effect of order of axioms in transitive inference	valid conclusions	no significant difference	no data available
	non-valid conclusions	no significant difference	no data available

7.4. Negation, disjunction and conjunction

The objective of this section was to further investigate the interaction between negation and disjunction, and the interaction between negation and conjunction; in particular to investigate the difficulties with negated conjunction identified in the first study. In order to ground the questions in an ecologically valid context, reference was made to Rector's (2003) discussion of the need to define exceptions. Rector (2003) noted that frame systems, originating from Minsky (1975), had developed the idea of defining defaults, and then exceptions to those defaults. This approach raised problems of computational tractability, and in any case was replaced with logic-base approaches. Attempts to introduce default mechanisms into the logic-based systems also ran into problems of tractability. The details of this discussion are not important here. What is important is that, in this discussion, Rector (2003) gives examples of defining exceptions in OWL, including two making use of negated conjunction. The first defines two exceptions:

"Drug type A and not subtype B and not subtype C" (1)

The second example extends this to define an exception to an exception:

"Drug type A and not (subtype B and not subtype B1) and not subtype C" (2)

The questions used are shown in Table 7-10. The table also indicates whether each question had a valid or non-valid putative conclusion, shows the RC of each valid question (see discussion later in this subsection) and provides a representation of the axioms and of the conclusion as one or more mental models. Note that these are the questions as used for the

laboratory study. The questions used in the online study varied slightly, as explained in subsection 7.4.2 below.

Table 7-10 Boolean operator questions – laboratory study

No.	axioms	putative conclusion	valid / not-valid	RC	mental model(s)
1	Z EquivalentTo (TOP_CLASS and not A and not B) TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	valid	3	TC ¬a ¬b ≡ c
2	Z EquivalentTo (TOP_CLASS and not (A or B)) TOP_CLASS DisjointUnionOf A, B, C		valid	3	
3	Z EquivalentTo (TOP_CLASS and not (A and not A_1)) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A1, A2	Z EquivalentTo (B or A_1)	valid	3	TC ¬a TC a1 ≡ b a1
4		Z EquivalentTo B	not-valid	n/a	
5	Z EquivalentTo (TOP_CLASS and (not A or A_1)) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2	Z EquivalentTo (B or A_1)	valid	2	
6		Z EquivalentTo A_1	not-valid	n/a	
7	Z EquivalentTo ((TOP_CLASS and not A) or (TOP_CLASS and A_1)) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2	Z EquivalentTo (B or A_1)	valid	3	
8		Z EquivalentTo A_2	not-valid	n/a	
9	Z EquivalentTo (TOP_CLASS and not (A and not (A_1 and not A_1_X))) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2 A_1 DisjointUnionOf A_1_X, A_1_Y	Z EquivalentTo (B or A_1_Y)	valid	4	TC ¬a TC a1 ¬a1x ≡ b a1y
10		Z EquivalentTo A_1_Y	not-valid	n/a	

Question complexity

The questions are divided into three groups: questions 1 and 2; questions 3 to 8; and questions 9 and 10. These represent increasing order of complexity. Here complexity can be measured in a variety of ways. The simplest may be to consider the number of levels in the class hierarchy, which can be regarded as one plus the number of DisjointUnionOf axioms. Thus, questions 1 and 2 have two levels:

(i) TOP_CLASS; (ii) A, B, C.

Questions 3 to 8 have three levels:

(i) TOP_CLASS; (ii) A, B; (iii) A_1, A_2.

Questions 9 and 10 have four levels:

(i) TOP_CLASS; (ii) A, B; (iii) A_1, A_2; (iv) A_1_X, A_1_Y

The same complexity categorisation can be arrived at by considering the mental models. This will be discussed later.

Alternatively, a syntactic approach can be taken by counting the levels of nesting in the axioms. Question 1 has no nesting, i.e. no brackets except for those which are used to contain

the whole expression after the *EquivalentTo* keyword[126]. On the same basis, questions 3 to 8 have one level of bracketing and questions 9 and 10 have two levels of bracketing. Using this approach question 2 is an anomaly, having one level of bracketing and thus being in the same category as questions 3 to 8.

Another approach, is to take account of the number of reasoning steps needed for each question, and the relational complexity of each step. This approach requires making assumptions about how a participant would mentally tackle a question, which might not be the same way someone trained in Boolean algebra and provided with pen and paper would tackle the question. Appendix B describes an analysis of the questions with valid putative conclusion, i.e. questions 1, 2, 3, 5, 7, 9. With the exception of question 5, there is an increase in the number of reasoning steps as we move from the lower to medium to higher level of complexity as determined by the number of levels in the class hierarchy or by the number and complexity of the mental models. Also with the exception of question 5, the questions at the lower and medium levels of mental model complexity have maximum RC 3. Question 5 has maximum RC 2. The fact that question 5 is not answered appreciably more accurately than the questions 3 and 7 (see Table 7-15 below) suggests that it is the complexity of mental models which is the main determinant of difficulty. Finally, question 9 is clearly more complex, not only in having more steps but in having a maximum RC 4.

Negated disjunction

Question 1 was motivated by statement (1) above, taken from Rector (2003). The axioms for question 2 are semantically identical, but the first axiom uses a negated disjunction. Both questions have the same valid putative conclusion. It was known from the first study, and from Khemlani et al. (2012a), that negated disjunction is easier than negated conjunction. The objective of questions 1 and 2 was to determine whether participants would find it harder to reason about negated disjunction, as in question 2, than the fully expanded form, as in question 1.

The hypothesis to be tested is:

> H7.4 Reasoning with a negated disjunction is inherently more difficult than reasoning with the expanded form.

As the axioms are semantically equivalent they have the same mental model representation. Table 7-10 shows the mental model representation, in two forms. The first form (*TC* ¬*a* ¬*b*) is created directly from the first axiom. The second form (*c*) is equivalent to the valid conclusion and arises when the second axiom, i.e. the axiom employing *DisjointUnionOf*, is taken account of. In both cases the representation only requires one mental model. Although the two questions are represented by the same mental models, it could be that the form of one question lends itself more easily to the creation of the models. As pointed out in Chapter 3, subsection 3.3.2, for axioms using Boolean operators, the mental model representation corresponds to DNF, with each model being one of the disjuncts. On this basis, the first axiom of question 1, which is in DNF, may be easier to translate into a mental model representation than the first axiom of question 2.

Negated conjunction

The remaining questions are designed to further investigate negated conjunction. Questions 3 to 8 are motivated by statement (2) above, from Rector (2003). In particular, the first

[126] These brackets are not required by the rules of syntax, nor are they required to clarify the syntax. Their use was purely to help readability.

axiom of questions 3 and 4 is similar to statement (2), in that it includes a nested expression, and thus an exception to an exception. For simplicity, because it does not include negated conjunction, the final part of statement 2 ("*and not subtype C*") has been omitted. In the first axiom of questions 5 and 6, *not (A and not A_1)* has been expanded to *not A or A_1*. In questions 7 and 8 the whole axiom has been expanded to DNF. Thus the axioms of questions 3 to 8 are all semantically equivalent. Questions 3, 5 and 7 have the same valid putative conclusion. Questions 6, 4 and 8 have different non-valid putative conclusions.

Since the axioms for questions 3 to 8 are semantically equivalent they have the same mental model representation. As with questions 1 and 2, the representation is shown in two forms. The first (*TC ¬a; TC a1*) arises immediately from the first axiom. The second (*b; a1*) results from taking account of the two remaining axioms, employing *DisjointUnionOf*. In both cases there are two alternative mental models, compared with the one mental model required for questions 1 and 2.

Questions 9 and 10 are more complex still, involving exceptions to exceptions to exceptions. The mental model representation for these two questions has two models, as with questions 3 to 8. However, whilst the first model shown (TC ¬a) is identical for both sets of questions, the second model is more complex for questions 9 and 10 (TC a1 ¬a1x) than for questions 3 to 8 (TC a1).

Negated conjunction and DNF

The first axioms of questions {3, 4}, {5, 6}, and {7, 8} represent a transition from the form used in Rector (2003) to DNF. Since DNF corresponds to the mental model representation, one might expect the order of reducing difficulty for the questions with valid conclusions to be 3, 5, 7, and for the questions with non-valid conclusions to be 4, 6, 8. This leads to the hypothesis:

> H7.5 Reasoning with a Boolean axiom in MOS becomes more difficult the further that axiom is syntactically removed from its DNF.

Increasing complexity

On the basis of mental model theory, and on the basis of the depth of the class hierarchy, the questions are arranged into three levels of complexity. This offers an opportunity to investigate the following hypothesis:

> H7.6 Reasoning with Boolean axioms becomes more difficult with increasing complexity, e.g. as measured by the complexity of the corresponding mental model.

It would be surprising were these hypotheses not to hold in the general case, e.g. at very high levels of complexity. The real question of concern is whether this effect makes itself felt at the levels of complexity being studied.

Questions 1, 3 and 9 from Table 7-10 can be used to test this hypothesis. These three questions are in the general form shown in Rector (2003), with increasing levels of nesting. In the cases of question 1 and question 3 there are semantically equivalent alternative questions. However, these would confound the effect of complexity with the effect of different syntactic formats.

There is no question with a non-valid putative conclusion at the lowest level of complexity. However, comparison of questions 4 and 10, both with non-valid putative conclusions, permits another comparison of the effect of complexity. As with the questions with valid conclusions discussed previously, questions 4 and 10 have the same syntactic form.

However, as discussed in the previous section, the effect of the differing complexity of the axioms will be confounded with any effect of the different putative non-valid conclusions[127].

7.4.1. Negation, disjunction and conjunction – the laboratory study

As with the questions in the previous section, for the laboratory study the questions were presented in two mutually reverse orders, as shown in Table 7-11.

Table 7-11 Negation, disjunction and conjunction - question ordering

order 1			order 2	
question no.	**valid / non-valid**		**question no.**	**valid / non-valid**
3	valid		4	non-valid
1	valid		2	valid
8	non-valid		7	valid
10	non-valid		9	valid
5	valid		6	non-valid
6	non-valid		5	valid
9	valid		10	non-valid
7	valid		8	non-valid
2	valid		1	valid
4	non-valid		3	valid

The results of the laboratory study are shown in Table 7-12. As a reminder, the table also shows whether the putative conclusion was valid or non-valid and shows the division into the three levels of complexity.

Table 7-12 Boolean operator questions – results from laboratory study

No.	**valid / non-valid**	**%age corr N=28**	**mean time (s.d.) – secs, N = 24**		
			overall	**corr**	**incorr**
1	valid	82%*	39 (26)	36 (23)	56 (37)
2	valid	86%*	43 (29)	36 (24)	78 (27)
3	valid	61%	96 (56)	99 (64)	89 (37)
4	non-valid	64%	105 (78)	112 (78)	89 (79)
5	valid	64%	65 (38)	61 (33)	74 (48)
6	non-valid	79%*	58 (33)	54 (32)	70 (40)
7	valid	68%*	70 (45)	57 (31)	109 (59)
8	non-valid	89%*	65 (26)	66 (26)	59 (32)
9	valid	54%	90 (48)	91 (49)	89 (51)
10	non-valid	68%*	94 (47)	94 (50)	95 (45)

** answered significantly better than chance ($p < 0.05$)*

Table 7-13 shows the accuracy and timing data for the questions in the two orders of presentation. Inspection of the table suggests that questions 3 and 4, when presented first had appreciably longer response times. This difference was significant in both cases (question 3: $t(22) = 3.9248$, $p < 0.001$; question 4: $t(21.928) = 7.8683$, $p < 0.001$). No ordering effect is apparent for the accuracy data. For example, the difference between the

[127] In retrospect, this problem could have been mitigated by making the putative conclusion for question 10 the same as for question 4: *Z EquivalentTo B*.

percentage of correct responses for question 4 in the two orderings (53% and 77%) was not significant (p = 0.254, Fisher's Exact test, two-sided).

Table 7-13 Boolean operator questions – performance by ordering

No.	valid / non-valid	position		%age correct		mean time (s.d.) - secs	
		order 1	order 2	order 1 N = 15	order 2 N = 13	order 1 N = 12	order 2 N = 12
1	valid	2	9	80%	85%	49 (31)	30 (15)
2	valid	9	2	100%	69%	22 (11)	64 (26)
3	valid	1	10	60%	62%	128 (57)	63 (31)
4	non-valid	10	1	53%	77%	42 (17)	167 (62)
5	valid	5	6	67%	62%	60 (33)	69 (43)
6	non-valid	6	5	80%	77%	36 (16)	79 (32)
7	valid	8	3	80%	54%	39 (13)	101 (44)
8	non-valid	3	8	93%	85%	67 (29)	63 (24)
9	valid	7	4	60%	46%	71 (31)	110 (56)
10	non-valid	4	7	60%	77%	94 (44)	95 (51)

7.4.2. Negation, disjunction and conjunction – the online study

There were 12 completed responses to this section from the online study. The questions used were changed in three ways from the questions used for the laboratory study. Firstly, the first two questions were omitted. It was hoped that reducing the section from 10 to 8 questions, i.e. the same size as the other sections, might encourage completed responses. Secondly, questions 4 and 8 were modified to have the same non-valid putative conclusion as question 6, i.e. Z EquivalentTo A_1. This better enables a comparison between questions 4, 6 and 8. Thirdly, the first axiom in questions 9 and 10 was modified to DNF. The intention was to enable a comparison between the form used in the laboratory study and DNF. Table 7.14 shows the questions, with the amendments in italics. For comparison and aggregation with the laboratory study, the same numbering is used, i.e. from 3 to 10, omitting questions 1 and 2.

Table 7.15 shows the percentage of correct responses for the online study and the two studies combined. Note that, for the pairs of questions with the same axioms and valid and non-valid putative conclusions (i.e. {3, 4}, {5, 6}, {7, 8} and {9, 10}), the question with a non-valid conclusion was always answered more accurately than the question with a valid conclusion. Based on both the laboratory study data and the aggregated data, the two questions which were not answered significantly better than chance (i.e. questions 3 and 9), were the valid questions, at the medium and high level of complexity, using the form motivated by Rector (2003).

Table 7-14 Boolean operator questions – online study

No.	axioms	putative conclusion	valid / not-valid	mental model(s)
3	Z EquivalentTo (TOP_CLASS and not (A and not A_1)) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A1, A2	Z EquivalentTo (B or A_1)	valid	TC ¬a TC a1 ≡ b a1
4		*Z EquivalentTo A_1*	not-valid	
5	Z EquivalentTo (TOP_CLASS and (not A or A_1)) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2	Z EquivalentTo (B or A_1)	valid	
6		Z EquivalentTo A_1	not-valid	
7	Z EquivalentTo ((TOP_CLASS and not A) or (TOP_CLASS and A_1)) TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2	Z EquivalentTo (B or A_1)	valid	
8		*Z EquivalentTo A_1*	not-valid	
9	*Z EquivalentTo ((TOP_CLASS and not A) or (TOP_CLASS and A_1 and not A_1_X))* TOP_CLASS DisjointUnionOf A, B A DisjointUnionOf A_1, A_2 A_1 DisjointUnionOf A_1_X, A_1_Y	Z EquivalentTo (B or A_1_Y)	valid	TC ¬a TC a1 ¬a1x ≡ b a1y
10		Z EquivalentTo A_1_Y	not-valid	

Table 7-15 Boolean operator questions – results of online study and two studies combined

No.	valid / non-valid	percentage correct	
		online study N = 12	combined studies N = 40
3	valid	50%	58%
4	non-valid	58%	63%
5	valid	58%	63%
6	non-valid	92%*	83%*
7	valid	50%	63%
8	non-valid	83%*	88%*
9	valid	58%	55%
10	non-valid	75%	70%*

* *answered significantly better than chance ($p < 0.05$)*

7.4.3. Negated disjunction – hypothesis H7.4

Comparison of questions 1 and 2 enables hypothesis H7.4 to be investigated. Note that this is based entirely on the laboratory study, since these questions did not occur in the online study. There was no significant difference in accuracy between the two questions ($p = 1.000$, Fisher's Exact Test, two-sided), nor in response time ($t(23) = 0.24693$, $p = 0.807$, paired test, two-sided). This analysis does not suffer from any difficulties of interpretation arising from question ordering, as the questions were counterbalanced. Half the participants saw question 1 in position 2 and question 2 in position 9; the other half saw question 1 in position 9 and question 2 in position 2. In summary, reasoning with negated disjunction (question 2) appears no harder than reasoning with the expanded form (question 1).

In summary, there was no support for hypothesis H7.4. Reasoning with negated disjunction is not significantly less accurate or significantly slower than reasoning with the expanded form.

7.4.4. Negated conjunction and DNF – hypothesis H7.5

Comparison of questions 3, 5 and 7 enabled an investigation of hypothesis H7.5. All three questions have semantically equivalent axioms and all three have a valid putative conclusion. A logistic analysis of deviance revealed no significant difference across the three questions ($p = 0.87$). This was based on the aggregated laboratory and online studies, since the questions were identical in both studies.

Inspection of the mean response time for these three questions reveals that there is little difference between questions 5 and 7, but that the response time for question 3 is appreciably longer than for questions 5 and 7. However, inspection of Table 7-13 suggests that this is because half the participants saw question 3 as their first question. Consequently, an ANOVA was conducted using data from the 12 participants who saw question 3 in position 10, i.e. the participants who saw the questions in order 2. This indicated no significant difference across the questions ($F(2,33) = 3.22$, $p = 0.053$).

Questions 4, 6 and 8, all with non-valid putative conclusions, offer another opportunity to investigate hypothesis H7.5. Here the putative conclusions were different for each of the three questions in the laboratory study, and therefore a confounding factor, but the same for each of the questions in the online study. Hence the online study offers a more controlled comparison of the effect of complexity. On the other hand, the small sample size restricts the statistical power of the test. Using the online study data, a logistic analysis of deviance showed no significant variation across the three questions ($p = 0.126$). When data from the two studies was aggregated, ignoring the difference in putative conclusions, then a logistic analysis of deviance did show a significant variation across the questions ($p = 0.020$)[128]. A subsequent Tukey HSD analysis revealed a significant difference between questions 4 and 8 ($p = 0.035$) but no significant difference between questions 4 and 6 ($p = 0.120$) and questions 6 and 8 ($p = 0.806$). Thus, accuracy with the DNF form was significantly better than with the Rector (2003) form.

Inspection of the mean response times for questions 4, 6 and 8 shows the same difficulty as occurred for questions 3, 5 and 7. Here, question 4 is the first question to be answered in order 2, and there is a wide difference between the mean response times for this question in the two orderings. Consequently, an ANOVA was conducted using data purely from order 1. This indicated that time varied significantly across the questions ($F(2, 33) = 5.801$, $p = 0.007$). A subsequent Tukey HSD analysis revealed a significant difference between question 6 and question 8 ($p = 0.007$) but no significant difference between question 4 and question 8 ($p = 0.055$) or question 4 and 6 ($p = 0.659$). This suggests that, contrary to hypothesis H7.5, DNF (question 8) took longer to interpret, at least than the form intermediate between DNF and the Rector (2003) form (question 6).

A further comparison of the DNF with the form derived from Rector (2003) is available by comparing the accuracy of questions 9 and 10 in the laboratory study with questions 9 and

[128] Aggregation of data from the two studies means that for part of the data all three questions had the same putative conclusions and for part of the data all three questions had different putative conclusions. Analysing separately the two questions for which the putative conclusion had been changed in the online study (questions 4 and 8), there was no significant difference between the proportion of correct responses in the laboratory and online study (Fisher's Exact Test, two-sided; question 4: $p = 0.736$; question 8: $p = 0.627$).

10 in the online study. Here, using Fisher's Exact Test, there was no significant difference for the questions individually (question 9: $p = 1.000$; question 10: 0.725), nor for questions combined ($p = 0.802$). The power of this test is necessarily limited by the small sample size in the online study.

In summary, there is no evidence to support hypothesis H7.5 for the questions with valid conclusions. For the questions with non-valid conclusions, there is evidence that the DNF was significantly more accurate than the Rector (2003) form. On the other hand, the DNF took significantly longer than the form intermediate between the DNF and the Rector (2003) form.

7.4.5. Increasing complexity – hypothesis H7.6

Earlier in this section it was noted that questions 1, 3 and 9 can be used to investigate the effect of increasing complexity. As question 1 did not appear in the online study, all comparisons are limited to data from the laboratory study. Inspection of Table 7-12 shows a reduction in percentage of correct responses with increasing complexity. However, a logistic analysis of deviance revealed no significant variation in accuracy across the three questions ($p = 0.055$).

Comparison of the response times for questions 1, 3 and 9 suffers from the difficulty that question 3 was answered first by the order 1 participants and, to a lesser extent, that question 1 was answered second by the same participants. To avoid these problems, an ANOVA was conducted using order 2. This indicated a significant variation with question ($F(2, 33) = 24.41$, $p < 0.001$). A subsequent Tukey HSD analysis indicated significant differences between all three questions (Q1 vs. Q3: $p < 0.001$; Q3 vs. Q9: $p = 0.016$; Q1 vs. Q9: $p < 0.001$).

Questions 4 and 10, both with non-valid putative conclusions, provide another opportunity to study the effect of complexity, although the different putative conclusions provide a confounding effect. Because of the change to question 10 in the online study, the comparison was restricted to the laboratory study. Inspection of Table 7-12 shows that there was very little difference in the accuracy with which these questions were answered, with the more complex question (question 10) being answered slightly more accurately. A Fisher's Exact Test showed no significant difference in accuracy between the two questions ($p = 1.000$, two-sided).

Question 4 was answered first in order 2, and inspection of Table 7-13 suggests this has appreciably increased the mean response time for this question. Consequently, a comparison was made between questions 4 and 10 solely for those participants who saw the questions in order 1. This indicated a significant difference in response time between the two questions ($t(11) = 5.5308$, $p < 0.001$), with question 10 taking longer.

In summary, there is partial support for hypothesis H7.6. Complexity, at the level studied, significantly affects response time but has no significant effect on accuracy.

7.4.6. Conclusions

Table 7-16 summarises the conclusions for hypotheses H7.4, H7.5 and H7.6. There are two noteworthy effects. Firstly, syntactic differences in the way the first axiom was expressed do make a difference, but apparently only for the questions with non-valid conclusions. Secondly, an increase in complexity does significantly increase response time.

Table 7-16 Negation, disjunction and conjunction – summary of hypotheses

		accuracy	**response time**
H7.4 – negated disjunction vs. expanded form	valid conclusions	No significant difference	
	non-valid conclusions	Not investigated	
H7.5 – reasoning with Boolean axioms more difficult the further the expression is from DNF	valid conclusions	No significant difference	
	non-valid conclusions	some evidence that DNF significantly more accurate than Rector (2003) form	DNF took significantly longer than form intermediate between DNF and Rector (2003) form
H7.6 – reasoning with Boolean axioms more difficult with increasing complexity, e.g. of mental model	valid conclusions	No significant difference	Significant increase in response time at each increase in complexity.
	non-valid conclusions	No significant difference between medium and high levels of complexity. Difference between low and medium levels of complexity not tested.	Significant increase in response time between medium and high levels of complexity. Difference between lowest and medium level of complexity not tested.

7.5. Negation and restriction

This section is concerned with investigation of the existential and universal restrictions, including with the interaction between negation and these restrictions. Table 7-17 shows the questions, as used in both the laboratory and the online study. Note that all questions have the same putative conclusion: *X DisjointTo Y*. Consider first questions 1 to 4. They share the same second axiom, which uses the universal restriction to constrain the class Y. For the first axiom, which constrains class X, questions 1 and 3 use the existential restriction, whilst questions 2 and 4 use the universal restriction. Questions 1 and 2 have the negation immediately before the named class (*MALE*), whilst questions 3 and 4 have the negation immediately before the anonymous class formed from the object property *has_child* and an existential or universal restriction. Thus, all four possible variants of type of restriction and position of negation are used. The first axioms of questions 1 and 4 are semantically equivalent, as are the first axioms of questions 2 and 3[129]. For this reason, the first axiom of

[129] The equivalence between the first axioms are based on the following MOS identities, where *R* is an object property and *C* is a class:

R some not C ≡ not R only C

R only not C ≡ not R some C

questions 1 and 4 are shown in a distinctive typeface, and the first axiom of questions 2 and 3 in normal typeface. Questions 5 to 8 repeat the first axiom from the first four questions. Questions 5 to 8 then share the same second axiom, which uses the existential restriction to constrain the class Y. In summary, we have three binary factors which characterize the eight questions: nature of restriction in first axiom; position of negation in first axiom; nature of restriction in second axiom.

Table 7-17 Negation and restriction questions

No.	First axiom – constraining X	Second axiom – constraining Y	valid / non-valid
1	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child only MALE	valid
2	X SubClassOf has_child only (not MALE)		non-valid
3	X SubClassOf not (has_child some MALE)		non-valid
4	***X SubClassOf not (has_child only MALE)***		valid
5	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child some MALE	non-valid
6	X SubClassOf has_child only (not MALE)		valid
7	X SubClassOf not (has_child some MALE)		valid
8	***X SubClassOf not (has_child only MALE)***		non-valid

N.B. All questions have the same putative conclusion: X DisjointTo Y

The fact that the first axioms fall into two groups of semantically equivalent axioms, combined with the two alternatives for the second axiom, means that the questions can be regarded as four pairs of semantically equivalent questions: {1, 4}, {2, 3}, {5, 8} and {6, 7}. Being semantically equivalent, the members of each pair will have the same mental model representation. This enables a test for the following hypothesis:

H7.7 The mental model representation of the questions determines their difficulty.

The four sets of semantically equivalent questions are characterised by two factors. One factor identifies the semantics of the first axiom, i.e. coincides with whether the first axiom is expressed in distinctive or normal typeface. The other factor differentiates between the two alternatives for the second axiom.

For questions 1, 2, 5 and 6, the negation immediately precedes a named class (*MALE*). For questions 3, 4, 7 and 8, the negation precedes an anonymous class, and syntactically immediately precedes an object property (*has_child*). This permits another characterization of the questions using three binary factors, differing from the previous characterization only in that the first factor (nature of restriction in the first axiom) has been replaced by the semantics of the first axiom (for which a proxy is the typeface used in Table 7-17). The division into questions with negation before an anonymous class and negation before an object property enables a test of the hypothesis:

H7.8 A difference in reasoning difficulty exists between negation before a named class and negation before an anonymous class.

Questions 2, 4, 5 and 7 have the same type of restriction in the first axiom as in the second axiom. For questions 2 and 4 this is the universal restriction. For questions 5 and 7 it is the

These are the analogues of the FOL identities: $\exists x \neg P(x) \equiv \neg \forall x\ P(x)$; $\neg \exists x\ P(x) \equiv \forall x \neg P(x)$. These, in turn, are analogues of De Morgan's laws in Propositional Logic.

existential restriction. One possibility is that participants might find it easier to reason when a question uses only one restriction type. If they are creating a mental model, then they only need to retrieve from memory the model for one type of restriction. If they are reasoning syntactically, then manipulations may be aided by using only one restriction type. This leads to the hypothesis:

> H7.9 The use of both existential and universal restrictions will cause reasoning to be more difficult than the use of one type of restriction.

Consideration of the interaction of restrictions can be taken one step further. The questions can be partitioned into the following four pairs:

- questions with universal restrictions in both axioms – questions 2 and 4;
- questions with existential restrictions in both axioms – questions 5 and 7;
- questions with an existential restriction in the first axiom and a universal restriction in the second – questions 1 and 3;
- questions with a universal restriction in the first axiom and an existential restriction in the second axiom – questions 6 and 8.

This leads to the hypothesis:

> H7.10 The pattern of first and second axiom restrictions (only … only, some … some, only … some, some … only) is an indicator of performance[130].

7.5.1. Negation and restriction – the laboratory study

As with the previous sections, the questions were presented to half the participants in one order, and to the other participants in the reverse order. Order 1 was as in Table 7-17, i.e. 1, 2, …8. Order 2 was the reverse, i.e. 8, 7, …1. Table 7-18 shows the results from the laboratory study.

Table 7-18 Negation and restriction – results from laboratory study

No.	valid / non-valid	%age corr N=28	mean time (s.d.) – secs, N = 24		
			overall	correct	incorrect
1	valid	61%	52 (39)	38 (24)	80 (50)
2	non-valid	50%	33 (18)	32 (14)	34 (22)
3	non-valid	68%*	45 (22)	43 (24)	49 (16)
4	valid	75%*	43 (25)	40 (24)	57 (24)
5	non-valid	64%	41 (30)	42 (32)	38 (27)
6	valid	50%	44 (40)	38 (25)	52 (55)
7	valid	79%*	43 (37)	34 (26)	79 (53)
8	non-valid	68%*	60 (37)	61 (42)	58 (29)

** answered significantly better than chance ($p < 0.05$)*

Table 7-19 shows the data for the questions in the two orders of presentation. The questions answered first by participants, i.e. questions 1 and 8, appear to have suffered a significant time penalty. For both questions, the ratio of the mean response time between those participants who saw the question first and those who saw the question last was approximately 2:1. This time penalty has appreciably increased the overall mean response

[130] H7.10 is a generalization of H7.9. H7.10 compares each of the combinations of restrictions: *only … only*, *some … some*, *only … some*, *some … only*. H7.9 compares the first two with the second two.

time for those two questions. In fact, questions 1 and 8 have the longest overall mean response times of the eight questions.

Table 7-19 Negation and restriction data – broken down by ordering

No.	valid / non-valid	position		%age correct		mean time (s.d.) - secs	
		order 1	order 2	order 1 N = 15	order 2 N = 13	order 1 N = 12	order 2 N = 12
1	valid	1	8	53%	69%	70 (44)	34 (25)
2	non-valid	2	7	47%	54%	28 (12)	39 (22)
3	non-valid	3	6	67%	69%	43 (17)	47 (27)
4	valid	4	5	73%	77%	48 (28)	39 (22)
5	non-valid	5	4	60%	69%	48 (36)	33 (22)
6	valid	6	3	40%	62%	43 (49)	46 (30)
7	valid	7	2	80%	77%	32 (21)	54 (46)
8	non-valid	8	1	73%	62%	41 (28)	78 (37)
MEANS				62%	67%	44 (33)	46 (32)

7.5.2. Negation and restriction – the online study

The questions for the online study were identical to those used in the laboratory study. There were 16 completed responses. Table 7-20 shows the percentage of correct responses for the online study and for the two studies combined. Considering the aggregated data from the two studies, five of the questions were answered significantly better than chance.

Table 7-20 Negation and restriction – results from online study and two studies combined

No.	valid / non-valid	percentage correct	
		online study N = 16	combined studies N = 44
1	valid	63%	61%
2	non-valid	50%	50%
3	non-valid	69%	68%*
4	valid	63%	70%*
5	non-valid	75%*	68%*
6	valid	50%	50%
7	valid	75%*	77%*
8	non-valid	81%*	73%*

** answered significantly better than chance ($p < 0.05$)*

7.5.3. Semantic equivalence and position of negation – H7.7 and H7.8

Based on the aggregated accuracy data from both the laboratory and online study, a logistic analysis of deviance was conducted using three factors: the semantics of the first axiom, as represented by the use of a normal or distinctive typeface in Table 7-17; the type of restriction used in the second axiom; and whether the negation precedes a named or anonymous class. Note that the first two of these determine the semantics of the question. The result of the analysis indicated a significant dependence of accuracy on the position of the negation ($p = 0.003$). However, neither of the other top-level effects (semantics: $p = 0.180$; second restriction: $p = 0.371$), nor the interaction effects (p values greater than at least 0.1) were significant[131]. Thus, accuracy was affected by the position of the negation but not by the semantics of the questions. This is entirely consistent with what one would expect

[131] The actual p-values vary slightly depending on the order of the factors used in the analysis. However, in all cases the effect of negation was significant ($p < 0.05$) and all the other top-level and interaction effects were not significant ($p > 0.05$).

from inspection of Table 7-20. Based on the aggregated data, the worst-performing question with negation before the anonymous class was question 3, which was answered exactly as accurately as the best-performing question with negation before a named class, i.e. question 5.

An ANOVA was conducted for time, using the same three factors. None of the top-level factors were significant (semantics: $F(1, 184) = 1.8749$, $p = 0.174$; second restriction: $F(1, 184) = 0.0005$, $p = 0.982$; position of negation: $F(1, 184) = 3.2545$, $p = 0.073$). Nor were the second-level interactions significant (p at least 0.6). However, the third-level interaction effect was significant ($F(1, 184) = 3.9376$, $p = 0.049$). One interpretation of this third-level interaction can be seen from the mean response times in Table 7-18:

- When the second axiom contains *only*, then for the questions in normal typeface, negation before a class (question 2) is answered more quickly than negation before a property (question 3), whereas for the questions in the special typeface, negation before a class (question 1) is answered more slowly than negation before a property (question 4).
- When the second axiom contains *some*, then the above inequalities are reversed. For the questions in normal typeface, negation before a class (question 6) is answered more slowly[132] than negation before a property (question 7), whereas for the questions in special typeface, negation before a class (question 5) is answered more quickly than negation before a property (question 8).

This behaviour appears to be related to the existence of syntactic clues. Considering the first bullet above, and considering questions 1 and 4, then the latter can be answered quickly by realising that the first axiom refers to the complement of the class in the second axiom (*has_child only MALE*). Question 2 is answered more quickly, but also less accurately than question 3. This could arise when *only (not MALE)* in the first axiom of question 2 is wrongly interpreted as *not (only MALE)* and hence appears to complement the class in the second axiom. In the same way, referring to the second bullet, question 7 offers a syntactic clue which makes it easier to answer than question 6; whilst question 5 can encourage a rapid erroneous response if *some (not MALE)* is mistaken for *not (some MALE)*.

In summary, questions with negation before the anonymous class were answered significantly more accurately, but took longer than questions with the negation before the named class. The semantics of the questions, and hence the corresponding mental models, had no effect on accuracy or response time. It may be that at least some of the questions were being answered by syntactic manipulation, without recourse to mental models. It may also be the case that, where mental models were being used, the questions were differentiated not by the difficulty of manipulating the models but by the difficulty of forming them, which can depend on question syntax.

7.5.4. Combining restriction types – H7.9

The hypothesis here is that participants may perform significantly better with questions which have the same type of restriction in both axioms (i.e. questions 2, 4, 5 and 7), compared to questions which use both types of restriction (i.e. questions 1, 3, 6 and 8). Based on the aggregated data from both laboratory and online studies, 66% of the responses to the

[132] In this one case, there is in fact only a small difference between the two mean times.

former set of questions were correct, compared with 63% of the responses to the latter set. This difference was not significant (p = 0.289, Fisher's Exact Test, one-sided)[133].

The mean time for the responses from the four questions using only one type of restriction is 40.1 seconds; that for the four questions using both types of restriction is 50.2 seconds. However, the latter figure is distorted because both of the questions which appear first in the two orderings, i.e. questions 1 and 8, use different restrictions in the two axioms. When the participants who saw these two questions as first questions are removed, the latter figure is reduced to 42.3 seconds[134]. On this basis, i.e. with the data relating to questions seen first removed, there was no significant difference between the two sets of questions (t(149.72) = 0.40064, p = 0.689).

In summary, there was no support for hypothesis H7.9. There is no evidence that using different types of restrictions in the two axioms makes a significant difference to accuracy, nor does it significantly increase the response time, compared with using the same restriction in both axioms.

7.5.5. Order of restrictions and position of negation – H7.10 and H7.8

H7.8 and H7.10 can be investigated together by using three factors: position of negation, nature of restriction in first axiom, and nature of restriction in second axiom. Based on the aggregated data from both laboratory and online studies, a logistic analysis of deviance using these factors confirmed that the position of negation significantly affects accuracy, as discussed above (p = 0.004). However, neither of the other top-level effects (first restriction: p = 0.114; second restriction: p = 0.364), nor any of the interaction effects (p values greater than at least 0.1)[135] were significant.

Applying an ANOVA to the time data revealed that none of the top-level effects was significant (first restriction: F(1, 184) = 0.103, p = 0.749; second restriction: F(1, 184) = 0.001, p = 0.982; position of negation: F(1, 184) = 3.2545, p = 0.073). With one exception, none of the interaction effects were significant (p > at least 0.1). The exception was the interaction between the choice of restriction for the first and second axiom (F(1, 184) = 3.9376, p = 0.049). A subsequent Tukey HSD analysis showed no significant pairwise difference. In particular, on the basis of the Tukey HSD analysis there is no significant pairwise difference in response time between any of the individual questions. Further investigation of the interaction between choice of restrictions in the first and second axioms indicated that the effect was caused by the high values recorded for questions 1 and 8 when they were answered first, as discussed above.

In summary, there was partial support for hypothesis H7.8. Questions with negation before an anonymous class were answered significantly more accurately than those with negation before a named class. However, note the discussion in the following subsection suggesting

[133] This result can be contrasted with that of Johnson-Laird et al. (1989). They were working with two-axiom problems akin to syllogisms, except using a variety of relations (e.g. "in the same place as") rather than equivalence. They found that, for problems which were represented by only one mental model, those problems using the same quantifier with the shared term in both axioms were answered significantly more correctly than those using a different quantifier. The percentage of correct responses for the former was 80%, for the latter 63%. For questions represented by more than one mental model, there were not enough correct responses for the difference to be significant.

[134] Specifically, this is the mean response time for questions 3 and 6 for all participants, for question 1 for the participants who saw order 2, and for question 8 for participants who saw order 1.

[135] As with the previous three-factor analysis of deviance, the precise p-values varied slightly with the order of factors, but not such as to materially affect the results.

that this may be an effect of the questions rather than a generic effect. Moreover, there was no significant difference in response time between the two sets of questions.

There was no support for hypothesis H7.10. The pattern of axiom restrictions made no significand difference, neither to accuracy nor response time.

7.5.6. Conclusions

Table 7-21 summarises the conclusions relating to hypotheses H7.7, H7.8, H7.9 and H7.10, based on the data available from the laboratory study for conclusions relating to response time, augmented with the data from the online study for conclusions relating to accuracy.

Table 7-21 Negation and restriction – summary of hypotheses

	accuracy	**response time**
H7.7 – mental model representation (i.e. semantics) determines difficulty	No significant difference.	No significant difference.
H7.8 – negation before a named class vs. negation before an anonymous class affects reasoning performance	Questions with negation before an anonymous class were answered significantly more accurately than those with negation before a named class.	No significant difference
H7.9 – use of different restrictions will create more difficulty than multiple uses of same restriction	No significant difference.	No significant difference.
H7.10 – pattern of axiom restrictions (only … only, some … some, only … some, some … only) affects reasoning performance	No significant difference.	No significant difference.

Table 7-21 indicates that the only factor which had any effect was the position of negation. Questions with negation before an anonymous class were answered significantly more accurately than those with negation before a named class.

It may be that there is an inherent difficulty in negation before a named class. Whilst the term 'negation' is being used here, in line with the use of *not* in MOS, the actual operation is complementing of a class, not negating a proposition. This may cause some confusion in participants' minds. The use of *not* before an object property is also performing a complement. However, it may read more naturally, being akin to the alignment of negation with a verb in English.

However, closer examination of the questions provides an additional insight into participants' thought processes. Comparison between questions with valid and non-valid putative conclusions is difficult because of possible bias. For example, Johnson-Laird et al. (1989), in experiments with multiple quantification, found a bias in favour of questions with

non-valid conclusions. Therefore, questions with valid conclusions and questions with non-valid conclusions will be considered separately.

Consider, first, the questions with valid conclusions, which are repeated for convenience in Table 7-22, along with the percentage of correct responses. Consistent with the results for hypothesis H7.8 shown in Table 7-21, the most correctly answered of these were questions 4 and 7, with negation before the anonymous class. Inspection of these two questions reveals that the same anonymous class occurs in both axioms. In question 4 this was *has_child only MALE*; in question 7 *has_child some MALE*. In each question, in the first axiom the class is complemented. So *Y* is a subclass of the anonymous class and *X* is a subclass of its complement. Thus it is immediately apparent from the syntax of the question that *X* and *Y* must be disjoint. There is no need for participants to construct a detailed mental model of the scenario. It may also be relevant that in both cases the mean times for the correct responses were appreciably less than the mean times for the incorrect responses. It may be that participants who got the question right quickly spotted the syntactic clue. The other questions with valid conclusions, questions 1 and 6, appear to offer no syntactic clues. Participants would be required either to perform some more complex syntactic manipulation or else form mental models to represent the questions. They clearly found this difficult. However, question 1 is semantically identical to question 4 and question 6 is semantically identical to question 7. Moreover, given the identities discussed earlier in this section, it is relatively easy to transform question 1 to question 4 and question 6 to question 7. For example, for question 1, the anonymous class in the first axiom can be transformed:

has_child some (not MALE) ≡ not (has_child only MALE)

Emphasizing these identities in training would help users of MOS to deal with the kinds of constructs found in questions 1 and 7.

Table 7-22 Negation and restrictions: valid conclusions

No.	First axiom – constraining X	Second axiom – constraining Y	%age correct[136]
1	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child only MALE	61%
4	***X SubClassOf not (has_child only MALE)***		70%
6	X SubClassOf has_child only (not MALE)	Y SubClassOf has_child some MALE	50%
7	X SubClassOf not (has_child some MALE)		77%

N.B. All questions have the same putative conclusion: X DisjointTo Y

Turning to the questions with non-valid conclusions, shown again in Table 7-23, the question which stands out is question 2, which was answered appreciably less well than the other three. This is the one question in the section which requires an understanding of the trivial satisfiability of the universal restriction. This feature of the universal restriction was pointed out in the handout. However, it is known to confuse, e.g. see Rector et al. (2004), and under the pressure of responding to the question, participants may overlook it.

[136] Based on the combined laboratory and online studies, N = 44.

Table 7-23 Negation and restrictions: non-valid conclusions

No.	First axiom – constraining X	Second axiom – constraining Y	%age correct[131]
2	X SubClassOf has_child only (not MALE)	Y SubClassOf has_child only MALE	50%
3	X SubClassOf not (has_child some MALE)		68%
5	*X SubClassOf has_child some (not MALE)*	Y SubClassOf has_child some MALE	68%
8	*X SubClassOf not (has_child only MALE)*		73%

N.B. All questions have the same putative conclusion: X DisjointTo Y

The universal restriction can be represented in mental model form as shown in Table *7-24*. Here *P* is an arbitrary object property, *X* an arbitrary class, and *x* an arbitrary member of *X*. The expression in brackets in the first disjunct ($\neg P\ \neg x$) would normally be assumed in a mental model representation. It is the second disjunct ($P\ \bot$) which corresponds to the trivial satisfaction of the universal restriction. In English, the word 'only' has the implicature that there is at least one individual which acts as object of the property P. The second disjunct may be omitted, either by simply overlooking it or as a result of this implicature. As a consequence, it appears to the participant that question 2 has a valid conclusion, because there is no potential for overlap between classes *X* and *Y*.

Table 7-24 Mental models for the universal restriction

OWL expression	Mental models
P only X	P x ¬(P ¬x) P ⊥

Apart from stressing the need to take account of the second mental model in the expansion of the universal restriction, understanding the identities discussed earlier would help with this question. Question 3 is semantically equivalent to question 2 and yet was answered with an appreciably greater level of accuracy. Identifying *has_child only (not MALE)* with not *(has_child some MALE)* would transform question 2 into question 3 and avoid the problems arising from the use of the universal restriction.

Another possible cause of difficulty for question 2 can arise from confusing *has_child only (not MALE)*, as in question 2, with *not (has_child only MALE)*. If a participant makes this mistake, then question 2 appears identical to question 4 which, as already discussed, can very easily be seen to have a valid conclusion. A similar problem can arise with question 5. Confusing *has_child some (not MALE)* with *not (has_child some MALE)* makes question 5, for which the putative conclusion is non-valid, appear the same as question 7, for which again the valid conclusion is easily seen. Indeed, Rector et al. (2004) have commented on confusion between "some not" and "not some". Again, stressing the two identities involving negation and restriction, as discussed at the beginning of this section, would help to clarify this confusion.

7.6. Nested restrictions

This section is concerned with investigating the effect of nested restrictions. Table 7-25 shows the questions, as used in the laboratory study. For this section, there was no data from the online study[137]. In the second column from the left there are either one or two axioms

[137] An error in framing the questions in the online study meant that the results were not strictly compatible with the laboratory study.

which constrain the class *X*. Four of the questions use one axiom; the other four use two axioms. In the former case, the second instantiation of the property (*has_child*) creates an anonymous class which is the object of the first instantiation of this property. In the latter case, a named class (*Y*) is used to connect the two axioms and plays the same role as the anonymous class.

Questions 1 to 4 between them use all four of the patterns *some … some*, *some … only*, *only … some*, *only … only*. Questions 5 to 8 then repeat the pattern *some … some*, *some … only*, *only … some*, *only … only*. Thus, each of the four patterns occurs twice, once in one axiom using an anonymous class to connect the property instantiations; and once in two axioms connected by the named class *X*. As a consequence, the first axiom of question 1 is semantically equivalent to the first two axioms of question 5. Similarly, the first axiom of questions 3, 6 and 8 is equivalent to the first two axioms of questions 5, 2 and 4.

The third column from the left then shows the remaining two axioms which are used to constrain an individual *a*. Questions 1 to 4 use the pattern *some not* whilst questions 5 to 8 use the pattern *not … some*. For all the questions, the putative conclusion was *a Type (not X)*. The structure of the questions was designed to enable the investigation of a number of hypotheses.

Table 7-25 Nested restrictions questions

No.	**First axiom(s) – constraining X**	**Remaining axioms – constraining a**	**valid / non-valid**
1	X SubClassOf (has_child some (has_child some FEMALE))	a has_child b; b Type has_child some (not FEMALE)	non-valid
2	X SubClassOf has_child some Y; Y EquivalentTo has_child only FEMALE		non-valid
3	X SubClassOf (has_child only (has_child some FEMALE))		non-valid
4	X SubClassOf has_child only Y; Y EquivalentTo has_child only FEMALE		valid
5	X SubClassOf has_child some Y; Y EquivalentTo has_child some FEMALE	a has_child b; b Type (not (has_child some FEMALE))	non-valid
6	X SubClassOf (has_child some (has_child only FEMALE)		non-valid
7	X SubClassOf has_child only Y; Y EquivalentTo has_child some FEMALE		valid
8	X SubClassOf (has_child only (has_child only FEMALE))		non-valid

N.B. All questions have the same putative conclusion: a Type (not X)

The previous section investigated whether reasoning performance differed when negation occurred before a named class or before an anonymous class (i.e. before an object property). This was formalised as hypothesis H7.8. In the context of the previous section, questions with negation before the anonymous class were answered significantly more accurately. Whilst there could be inherent differences in how participants react to the two different situations, the differences could also arise because of the structure of the questions. Subsection 7.5.6 discussed how the use of syntactic 'cues' could be responsible for the increase in accuracy, for those questions with valid putative conclusions, when negation occurs before an anonymous class. In this section, comparison of questions 1 to 4, where

negation occurs before a named class, with questions 5 to 8, where negation occurs before an anonymous class, enables this issue to be further investigated.

The previous section also investigated whether the use of the existential and universal restrictions in a question creates more difficulty than when only one type of restriction is used. This was formalised as hypothesis H7.9. The current section offers another opportunity to investigate this. Each question uses three instances of restrictions, of which the first two are nested. In questions 1 and 5 these are all of the same type (*some*), in the other questions both types of restrictions are used.

In the previous section there were pairs of question with the same mental model, which allowed investigation of the hypothesis (H7.7) that the mental model determined the difficulty of the questions. In this section there are no semantically equivalent questions, i.e. no questions which share the same mental model. However, the axiom or axioms which constrain the class *X* do occur in semantically equivalent pairs: {1, 5}, {2, 6}, {3, 7} and {4, 8}. One member of each pair uses one axiom; the other member uses two axioms joined by the class *Y*. This leads to the hypothesis:

> H7.11 Semantic equivalence between part of a question (here that part constraining class *X*) will be an indicator of performance.

The availability of four questions using one axiom to constrain *X* and four questions with two axioms allows the investigation of the following hypothesis:

> H7.12 There will be a difference in reasoning performance between the equivalent use of a named class and an anonymous class.

7.6.1. Nested restrictions – the laboratory study

As before, the questions were represented in two reverse orders, as shown in Table 7-26. Table 7-27 shows the percentage correct responses and mean response time for each question. As can be seen, only three of the questions were answered significantly better than chance.

Table 7-26 Nested restrictions - question ordering

order 1			**order 2**	
question no.	**valid / non-valid**		**question no.**	**valid / non-valid**
1	non-valid		5	non-valid
2	non-valid		6	non-valid
3	non-valid		7	valid
4	valid		8	non-valid
8	non-valid		4	valid
7	valid		3	non-valid
6	non-valid		2	non-valid
5	non-valid		1	non-valid

Table 7-27 Nested restrictions – results from laboratory study

No.	valid / non-valid	%age corr N=28	mean time (s.d.) – secs, N = 24		
			overall	correct	incorrect
1	non-valid	71%*	69 (45)	71 (46)	60 (44)
2	non-valid	57%	79 (53)	101 (55)	52 (37)
3	non-valid	71%*	63 (43)	63 (46)	65 (38)
4	valid	57%	63 (39)	56 (27)	72 (52)
5	non-valid	54%	88 (62)	94 (68)	78 (53)
6	non-valid	64%	73 (45)	66 (44)	84 (46)
7	valid	71%*	80 (36)	71 (34)	108 (29)
8	non-valid	50%	55 (30)	50 (23)	59 (36)

** answered significantly better than chance ($p < 0.05$)*

Table 7-28 shows the data for the two orders of presentation. Inspection of the data suggests that questions answered at the beginning of the section were not answered appreciably worse than questions answered later. For example, the percentage of correct responses for question 1 differs little between when the question was answered first and when it was answered last. Indeed, the percentage was slightly higher in order 1 when question 1 was the first to be answered, compared with order 2 when it was the last. The picture is different when we look at the mean response times. The mean response time for question 5 when answered in first position was 2.7 times the mean response time when the question was answered last. For question 1 the ratio was 2.3. This effect may have significantly distorted the overall response time for these two questions. A similar effect, although not quite as marked, may apply to questions 2 and 6, each of which were answered second by half the participants.

Table 7-28 Nested restrictions – results broken down by ordering

No.	valid / non-valid	position		%age correct		mean time (s.d.) - secs	
		order 1	order 2	order 1 N = 15	order 2 N = 13	order 1 N = 12	order 2 N = 12
1	non-valid	1	8	73%	69%	96 (45)	41 (23)
2	non-valid	2	7	67%	46%	96 (56)	62 (46)
3	non-valid	3	6	73%	69%	71 (41)	56 (45)
4	valid	4	5	67%	46%	55 (43)	71 (36)
5	non-valid	8	1	47%	62%	47 (28)	128 (60)
6	non-valid	7	2	73%	54%	51 (34)	95 (44)
7	valid	6	3	87%	54%	62 (36)	98 (26)
8	non-valid	5	4	53%	46%	41 (22)	68 (32)

7.6.2. Effect of pattern of restrictions and position of negation – hypotheses H7.8 and H7.11

To understand how the pattern of restrictions and the position of negation affected the proportion of correct responses, a logistic analysis of deviance was conducted using three factors: nature of first restriction; nature of second restriction; and position of negation. None of the top-level and interaction effects were significant (first restriction: $p = 1.000$; second restriction: $p = 0.267$; position of negation: $p = 0.592$; interactions: $p > 0.2$ in all cases). This result does not support hypothesis H7.8, i.e. that the position of negation affects performance. Moreover, the lack of any significant effects due to the first two restrictions, and their interaction, does not support hypothesis H7.11, i.e. that the semantics of the axiom(s) constraining the class X will be an indicator of performance.

A comparison of response times suffers from the difficulty that the first, and even the second question, may suffer a time penalty. To overcome this, the analysis was conducted on timing data taken only from those responses which participants provided during the second half of the section[138]. Thus the timing data for questions 1 to 4 were taken from order 2, whilst the timing data for questions 5 to 8 were taken from order 1. This means that all the timing data used was taken after the participants had accustomed themselves to the nature of the question, at a cost of halving the number of data points available. On this basis, an ANOVA was undertaken using the same three factors as for the logistic analysis. Again, none of the top-level factors were significant (first restriction: $F(1, 88) = 0.8855$, $p = 0.349$; second restriction: $F(1, 88) = 0.2759$, $p = 0.601$; position of negation: $F(1, 88) = 0.4488$, $p = 0.505$). This, in particular, does not support hypothesis H7.8 that the position of negation affects performance. With one exception, all the interaction effects were not significant ($p > 0.3$, in all cases). The exception was the interaction between the nature of the second restriction and the position of negation ($F(1, 88) = 4.1681$, $p = 0.044$). A subsequent Tukey HSD analysis revealed no significant comparisons. As with the accuracy data, the lack of any significant effects relating to the nature of the first two restrictions and their interaction does not support hypothesis H7.11.

In summary, there was no support for hypothesis H7.8. There was no significant difference, neither in accuracy nor response time, between questions where negation was before an anonymous class and questions where negation was before a named class. Similarly, there was no support for hypothesis H7.11. Partial semantic equivalence made no significant difference to accuracy or to response time.

Table 7-29 provides insight into the one significant effect found, the interaction between the nature of the second restriction and the position of negation. In questions 1 to 4, when negation occurred directly before the class *FEMALE*, the mean response time was less when *some* was the second restriction. This may have to do with a rapid realisation that although *not FEMALE* and *FEMALE* are complementary, there is no contradiction in having children of both genders. However, in questions 5 to 8, when the negation occurred before the object property *has_child*, the mean response time was less when *only* was the second restriction. In the case of question 6, the mean time for correct responses is appreciably less than for incorrect responses, and the low response time may have arisen in part because of the rapid realisation that *has_child only FEMALE* and *not (has_child some FEMALE)* are compatible in the case that there are no children.

[138] The motivation here is that, in addition to any gradual change in response time as the participants move through each section, there is a very considerable reduction in response time after the first one or two questions in each section. The bias from this initial effect can be avoided by using data only from questions answered in the second half of the study section, where alteration of response time with question position is much less. This does mean that there is no compensation for any slight change in response time moving through the later questions, since data is no longer being taken from two mutually reverse orderings. Moreover, the study is no longer within-participants, since the data for questions 1 to 4 is taken from different participants than the data for questions 5 to 8.

Table 7-29 Interaction between second restriction and position of negation – mean response times

		negation before class - secs	negation before property - secs	TOTALS - secs
second restriction	**only**	66.5	46.0	56.2
	some	48.5	54.7	51.6
TOTALS		57.5	50.3	53.9

7.6.3. Use of one type of restriction vs. use of different restrictions – hypothesis H7.9

These questions provide another opportunity to investigate H7.9. Table 7-30 shows the percentage of correct responses and the mean response time for each pattern of restrictions in the axioms constraining the class *X*. Each of the columns in the table represents two questions, one of which involves negation before a named class, and the other of which involves negation before an object property. Since the final axiom always involves *some*, then the case *some … some* represents questions 1 and 5 in which only one axiom type is used. As before, the mean response times are based purely on the response times for the questions met by participants in positions 5 to 8.

Table 7-30 Effect of different patterns of restrictions

	pattern of first two restrictions			
	some … some	*some … only*	*only … some*	*only … only*
%age correct	63%*	61%	71%*	54%
mean time (s.d.) - secs	44 (25)	56 (40)	59 (40)	56 (33)

** answered significantly better than chance ($p < 0.05$)*

Accuracy with the *some … some* pattern, requiring only the use of the existential restriction, is neither relatively high nor low. In fact, a logistic analysis of deviance, using one factor with four levels representing the columns of Table 7.27, confirmed that the pattern of the first two restrictions made no difference to accuracy ($p = 0.273$). On the other hand, the mean response time for the *some … some* pattern is appreciably less than for the other three patterns. However the difference is not significant; an ANOVA indicated that there was no significant difference in response times across the four patterns ($F(3, 92) = 0.6938$, $p = 0.558$).

One feature of note in Table 7-30 is the low level of accuracy achieved by the *only … only* pattern. This pattern is represented by questions 4 and 8; the former question was not answered significantly better than chance; the latter question, which requires an understanding of the trivial satisfaction of the universal restriction, was answered at chance.

7.6.4. Effect of named vs. anonymous class – hypothesis H7.12

It has already been noted that half of the questions use a named class, *Y*, whereas the other half replace this with an anonymous class. As Table 7-25 shows, each pattern of the first two restrictions (*some … some*, *some … only* etc) occurs once with class *Y* and once with an anonymous class. This provides an opportunity to investigate whether the choice of named versus anonymous class has any effect on performance.

Of the questions employing the named class, *Y*, 60% were answered correctly, whilst of the questions employing an anonymous class, 64% were answered correctly. This difference was not significant (p = 0.582, Fisher's Exact Test, two-sided). The mean response time for those questions employing the named class was 77 seconds; that for the questions with the anonymous class was 65 seconds. This difference was also not significant (t(189.62) = 1.7518, p = 0.08142). In this analysis, all the responses are being used, since the use of named and anonymous classes counterbalance, i.e for each position, when order 1 uses an anonymous class, order 2 uses a named class, and vice versa. If only those questions which appeared to a participant in the second half of each section are used, the mean response time for those questions employing the named class is 61 seconds; that for the questions with the anonymous class was 47 seconds. In this case, the difference is significant (t(93.926) = 2.0714, p = 0.041).

In summary, there was no support for hypothesis H7.12. The use of a named or anonymous class made no significant difference to accuracy. The mean reponse time for questions with the named class was appreciably longer than for questions with the anonymous class. However, the difference was not significant when data from all the responses were used. Moreover, it should be borne in mind that the two sets of questions (i.e. with named and anonymous classes) were different. Hypothesis H7.12 is reconsidered in the final study.

7.6.5. Conclusions

Table 7-31 summarises the conclusions regarding H7.8, H7.9, H7.11 and H7.12.

Table 7-31 Nested restrictions – summary of hypotheses

	accuracy	**response time**
H7.8 – negation before a named class vs. negation before an anonymous class affects reasoning performance.	No significant difference.	No significant difference. However significant interaction between position of negation and nature of second restriction.
H7.9 – use of different restrictions will create more difficulty than multiple uses of the same restriction	No significant difference.	
H7.11 – partial semantic equivalence (here of the axiom(s) constraining the class *X*) will be an indicator of performance	No significant difference.	
H7.12 use of named vs. anonymous class affects reasoning performance	No significant difference.	No significant difference; response time with named class significantly longer if we restrict analysis to those responses during second half of each section.

The findings for H7.8 contradict those from the previous section. They do, however, support the suggestion in subsection 7.5.6, that any difference in the previous section related to the

position of negation was not inherent in the particular constructs, but was caused by the overall structure of the questions, e.g. the ability to pick up a syntactic 'clue'.

The findings for H7.9 were exactly as in the previous section. There seems to be no disadvantage, either in accuracy or response time, to mixing existential and universal restrictions. In fact, questions 1 and 5, the two questions which used only the existential restriction, were answered quite differently. Question 1 was one of the most accurately answered (71%) whilst question 5 was the least well answered (54%). Turning to response time, question 1 was in the middle of the range, whilst question 5 had the longest mean response time.

Similarly, there is no evidence to support hypothesis H7.11, that the semantics of the constraint on class *X*, determined by the pattern of the first two restrictions, would affect accuracy or response time.

The lack of any evidence for these hypotheses supports the opinion that accuracy, and perhaps also response time, are determined by a subtle interplay of features in the question, rather than by the more simplistic viewpoints underlying these hypotheses.

When all the responses were taken into account, there was similarly no support for hypothesis H7.12, i.e. there was no significant difference in accuracy or response time between the case when one axiom was used and when two axioms were joined by a named class, *Y*. However, if analysis is restricted to those responses given by each participant during the second half of each section, then response time with the anonymous class is significantly longer than with the named class. Moreover, the questions using one axiom differed in semantics from those using two axioms. This topic is investigated more rigorously in the next chapter.

The remainder of this subsection considers the questions in detail, considering first the two questions with valid putative conclusions and then the six questions with non-valid conclusions.

Questions with valid putative conclusions

There were two questions with valid putative conclusions: questions 4 and 7. In considering these and the other questions, for convenience of explanation *Y* will be used to denote both the named class *Y* and the analogous anonymous class. In question 7 it is clear from the syntax that *Y* is the complement of the anonymous class defined in the final axiom, i.e. *not (has_child some FEMALE)*. This immediately implies that *b* cannot be in *Y*, and hence, given the use of the universal restriction in the first axiom, that *a* cannot be in *X*. This contrasts with question 4, which was answered less well. Here there is no such syntactic clue and participants are likely to find it difficult to construct a model of the class *X* and a model of the possibilities for *a* to identify that these models have no shared elements. Whilst for both questions the mean time to answer correctly was less than the mean time to answer incorrectly, the difference was only significant for question 7 (question 7: $t(16.932) = 3.11$, $p = 0.006$; question 4: $t(13.235) = 0.14744$, $p = 0.885$), suggesting that participants who answered this question correctly picked up the syntactic clue quickly. Note that if the final axiom of question 4 was transformed, using the rule for negation and restrictions, to *b Type has_child not (only FEMALE)*, then this question would offer a syntactic clue, analogously to question 7, and would be likely to be answered more accurately. As for certain questions in the previous section, if participants were aware of this rule, this would have helped their reasoning.

Questions with non-valid putative conclusions

The remaining six questions all had non-valid conclusions. The best answered were questions 1 and 3. These questions are relatively straightforward, in that it is clear the *b* can be in *Y*. The first axiom states that a member of *Y* has a female child and the final axiom states that *b* has a non-female child. There is no contradiction in *b* being in *Y*, although it does require an awareness of the second mental model in the expansion of the existential restriction, as shown in Table 7-32. In both questions, having deduced that *b* can be in *Y*, it is then clear, from the first axiom, that *a* can be in *X*.

Table 7-32 Mental models for the existential restriction

OWL expression	**Mental models**
P some X	P x ¬(P ¬x) P x (P ¬x)

Question 6 was also answered relatively well. The non-validity of the conclusion can deduced in two ways. It is possible for *b* to be in *Y*, since there is no contradiction in logic between having only female children and having no female children, if there are no children at all. However, this requires an understanding that the universal restriction can be trivially satisfied, i.e. it requires an awareness of the second model in Table *7-24*. Although this was pointed out in the handout, participants might easily overlook this and revert to the more natural sense of *only*, with the implicature that a female child does exist. However, even if participants conclude wrongly that *b* cannot be in *Y*, it is still possible for *a* to have another child (i.e. besides *b*) which itself does have a female child, and hence for *a* to be in *X*. This also requires an awareness of the second mental model in the expansion of the existential restriction, as shown in Table 7-32. A participant who initially takes the first model for the first (existential) restriction, and then fails to take note of the second model of the second (universal) restriction, can still answer the question correctly by backtracking to the first restriction and taking its second model.

None of the remaining three questions were answered significantly better than chance. In question 5 there is a strong syntactic clue that *b* cannot be in *Y*; *not (has_child some FEMALE)* is clearly the complement of *has_child some FEMALE*. To answer the question correctly, it is essential to backtrack to the first restriction and use its second model to understand that *a* can have another child, that does have only female children, and hence *a* can be in *X*. A similar argument applies to question 2. It is relatively easy to see that *b* cannot be in *Y*, since *b* cannot have only female children and some non-female children. Again, to understand the non-validity of the conclusion it is essential to backtrack to the second model for the first restriction.

Finally, question 8 has similarities with question 6. It is possible for *b* to be in *Y*, and hence *a* in *X*, since there is no contradiction in logic between having only female children and having no female children, if there are no children at all. However, unlike with question 6, if participants fail to understand this, backtracking to the first (universal) restriction offers no second chance. It is clear that the second model in the interpretation of the universal restriction is not applicable here, since this would require that *a* has no children, and we know that *a has_child b*. So, although question 6 can be correctly answered by making use of the trivial satisfaction of the universal restriction, this knowledge is not essential. Only question 8 requires that knowledge for a correct answer.

7.7. Participant feedback

At the end of the study, participants were asked three questions:

- What did you find difficult about the quiz?
- What did you find easy about the quiz?
- Do you have any general comments about the quiz?

There were a large number of responses, and this section discusses those which are most representative or seem to offer most insight. Rather than consider the feedback under the headings above, comments are separated here into: methodological comments about the study itself; general comments about the questions; and comments specific to the study sections.

7.7.1. Methodological comments

There was some positive feedback about the organization of the study. One participant commented that "guidelines and explanations help a lot in answering". Presumably this was a reference to the handout and the on-screen text at the beginning and before each section. On the other hand, another participant commented that there was "quite a lot of reading material beforehand". Generally, participants seemed happy to take part. One participant commented that "it was quite interesting", and one even that "it was fun".

There were some comments about the order of the questions. One participant commented that the "order of questions in terms of difficulty did not help". The participant would have liked to have seen some simple questions presented first. Another participant suggested "a few examples at the start to warm up the brain". On a related point, two participants' comments gave evidence of learning as they proceeded through the study; one noting doubts on previous questions arising as the study progressed, and another participant noting that understanding "a hard one" makes some of the others "easier and quicker to answer".

There were three responses which called into question the ecological validity of the approach, on the grounds that in practice one would not perform the more complex operations mentally. Two of these participants commented that in practice one would use pen and paper, whilst another would have used a theorem prover "much as I would use a calculator to do long division". One of the participants who wanted to write things down stated that "my working memory was being tested as much as my logic reasoning skills". Indeed, the objective was to see how well people can reason about DL constructs mentally. Given enough time and pen and paper, it is likely that the accuracy could be appreciably improved, but this would have failed to differentiate the various alternative constructs and syntaxes. Moreover, the goal of language developers should be to create syntaxes which are easily intelligible and require the minimum of reasoning with pen and paper.

Two participants were confused by the use of the word "valid"; one of them suggested that the phrase "does this axiom necessarily follow" would have been clearer. On the other hand, a third participant thought that "the answering mechanism of valid or not valid was quite easy to use".

Finally, some participants would have liked to have known the correct answers at the end. Several expressed a wish, either through the online feedback or verbally, to know their scores.

7.7.2. General comments

A number of people commented on the difficulty of at least some of the questions. One participant commented on the need to "concentrate a lot to keep a mental model of the statements". Another participant noted, perhaps not surprisingly, that "the longer the list of expressions, the harder I found the questions". One participant commented that the order of the axioms was not helpful; the participant often having "to start from the bottom". The issue of axiom ordering was not given particular attention in framing the questions. Nor is the author aware of it being discussed elsewhere in the literature. It potentially can affect comprehension and would be a useful issue to consider when automatically creating justifications.

More specifically, two participants commented on the difficulty of using abstract symbols, rather than names, as in the section on Boolean operators. Conversely, one participant commented that the property names were too long, noting that "the English meaning is not essential to answer the quiz", but that "it takes more effort to read the axioms".

One participant commented that the syntax was not "that intuitive", in particular drawing attention to the "and / or / not" rules of precedence. In contrast, another participant responded to the question "what did you find easy?", with "the basic notions of simplified OWL".

A number of participants talked about a need to draw or visualize. One went so far as to say "I am a very visual person … I draw graphs a lot when considering inferences". The syntactic approach was also represented by one participant, who commented "some questions were easy to solve when ignoring the actual class expressions, just looking at the general form of the formulas". Presumably this participant was reasoning syntactically, where possible, rather than building mental models. In a similar vein, and specific to the object property questions, another participant commented on the difficulty of "trying to juggle the relations in my head and substitute them appropriately".

7.7.3. Object properties

A number of responses indicated that the questions using transitive properties were easier than the questions using functional properties. These comments took two forms. Some participants referred explicitly to functionality or transitivity, e.g. commenting that they found functionality difficult or that "the transitivity questions were a lot easier"[139]. One even stated that "I am much more familiar with transitivity". On the other hand, some participants referred to the exemplar properties. A number of participants identified the *greater_than_or_equal* property as being relatively easy; a number identified the *has_nearest_neighbour* property as being relatively hard. One participant commented that "'greater than or equal' is a more familiar relationship than 'nearest neighbour'". This suggests that some participants were drawing on their understanding of the property, not simply on the stated characteristics. This was despite the clear instruction in the handout that the meaningful names should not affect any conclusions to be drawn. One participant appeared to be confused by the property *has_nearest_neighbour* being functional, since in some real situations it is possible for an entity to have more than one nearest neighbour.

[139] Some of these comments may, in part, reflect the fact that that the two transitivity questions were designed to be potentially on the same level of difficulty as the two easiest functionality questions, and would be expected to be easier than the other four functionality questions.

There was evidence that some participants were trying to visualize the object properties. One participant commented that it was difficult to mentally visualize the nearest neighbour relations. A participant from the online study responded that the nearest neighbour questions were easy once a graph had been drawn as a model. As with the laboratory study, participants in the online study were asked not to use pen and paper. However, there was clearly no control over this. So this participant may have drawn models on paper, or possibly the reference was to drawing mental models. In any case, these two responses suggest that some participants were using their understanding of nearest neighbour to create a visualization.

Finally, there were two comments relating to features with which the participants were not very familiar. One participant noted that the *DifferentFrom* keyword was not something often used, and hence the participant "took a while to think about it". Another participant had not had much experience with individuals, noting that "the ontologies I use are mainly about classes".

7.7.4. Boolean operators

There were conflicting views on the difficulty of this section; two participants placing it into the easy category and another into the difficult. One even claimed to have given random answers because of the difficulty of working without pen and paper. As already noted in subsection 7.7.2 above, another participant drew attention to the difficulty of the *and* / *or* / *not* rules of precedence. These conflicting views may relate to the background of the participants. Indeed, one of the participants who found this section easy commented on the fact that "*or*s, *not*s and *and*s … appear more often in non-logic programming languages". Related to this, another participant "thought that knowledge of electronic gates would have helped – but that was long ago".

There were some comments on the naming conventions. One participant liked the easy naming convention, e.g. A, A_1, A_1_X. Another two commented favourably on the use of the same hierarchical pattern throughout. On the other hand, two participants were confused by the name TOP_CLASS. One commented that "this can be easily confused with Top". In fact, TOP_CLASS was intended to be a synonym for *Top* (⊤), chosen to have a more obvious meaning for those not familiar with the term *Top*.

Although the order of precedence for *not*, *and* and *or* was explained in the handout, one participant admitted to not being clear whether *(not A or A_1)* was equivalent to (*not (A or A_1))* or *((not A) or A_1)*. Another participant would have liked a refresher "on how not (X or Y) expands".

7.7.5. Restrictions

This subsection considers together the two study sections concerned with the existential and universal restrictions. As with the section on Boolean operators, there were mixed responses. One participant regarded both sections as being "almost impossible". Another identified the "unnamed class section" as being particularly difficult; presumably this was a reference to the use of restrictions to create anonymous classes. Some participants identified *some* and *only* as being difficult. Negation was also identified as difficult, specifically in the context of its use with restrictions.

On the other hand, one participant thought the usage of existential and universal restrictions was easy, whilst another participant commented that the section on negation and restrictions, as described in Section 7.5, was "fairly easy". Referring to the same section, a different

participant commented that "when the sets are obviously disjoint it is very quick to recognize because the questions are similar". This is presumably a reference to picking up on the syntactic clues discussed in Section 7.5.

7.7.6. Conclusions

Some of the participants admitted to finding many, if not most, of the questions very difficult. There was variation in which sections were found the more easy or the more difficult, perhaps arising from participants' varying backgrounds.

Some participants were clearly using visualization. Others seemed to be reasoning syntactically, and picking up on the syntactic clues already discussed. This is consistent with Ford's (1995) observation, that people can be categorized as verbal or spatial reasoners; although one should also take into account that the same person may reason differently when confronted with different types of questions.

Transitivity was found easier than functionality, in line with the quantitative findings discussed in Section 7.3.4. People appeared to be drawing on their pre-existing knowledge of the two relations, *greater_than_or_equal_to* and *has_nearest_neighbour*, and it may be that a greater familiarity with the former helped to make the transitive questions seem easier. It also raises the question to what extent participants in such studies can ever really ignore their pre-existing models of the meanings of words. In the real world such pre-existing models are used all the time and can override purely formal reasoning, as Johnson-Laird (2002) has pointed out in his study of the various interpretations of the conditional.

The precedence of Boolean operators was a problem for some people. It was also clear that some participants would have benefited from a reminder of De Morgan's Laws. On the other hand, there is generally greater familiarity with Boolean logic than with the use of restrictions, and some participants appear to have profited from this familiarity.

Finally, whilst the use of restrictions was found very difficult by some participants, including in conjunction with negation, some of the feedback does suggest that they were being helped by the syntactic clues, i.e. where an anonymous class and its complement appeared in the same question.

7.8. Effect of participant prior knowledge and experience

At the beginning of the study participants were asked to rate their knowledge of logic and DL, and their usage of DL. The questions were exactly as in the previous study, with exactly the same response categories, see Section 6.8. Table 7-33 shows the number of participants, the proportion of correct responses and the mean response time for each category of expertise for logic and DL. Unlike in the previous study, there were some participants who claimed no knowledge of logic and some who claimed no knowledge of DL.

Table 7-34 shows similar data for the categories describing usage of DL. Some participants did not respond to the question. The assumption is that these participants do not use DL. There were also some participants who responded with the 'other' option. This category consisted chiefly of participants who have not used DLs or do not use them currently. Because of the ambiguity over the reasons given for the 'other' response, data for this response is excluded from the following analysis. Excluding 'other', increasing expertise is assumed to be represented by movement down the rows of Table 7-34 i.e. from 'no response' to 'researching'.

In this study, participants were also asked the question "please indicate which of the following most accurately represents your relationship with English". The possible responses were:

- English is my main language, i.e. I am more comfortable using English than any other language.
- English is one of my main languages, i.e. it is one of a number of languages which I am equally comfortable using.
- English is not my main language, i.e. I am more comfortable using another language.

The motivation for this question was research indicating that choices are less heuristically biased when made in a foreign language (Costa et al., 2014). In fact, the cited research suggests that this phenomenon is limited to emotional choices. Nevertheless, it was felt that under conditions of cognitive stress, when working in a foreign language, participants might be less prone to introduce bias and more likely to rely on pure reasoning. On the other hand, lack of familiarity with English might reduce performance. Table 7-35 shows the breakdown of accuracy and response time data by the three categories.

The same questions were asked both to participants in the laboratory study and participants in the online study. Because the results presented here are based on the full study, rather than the individual sections, the data is taken entirely from the laboratory study. Interpretation of these data needs to take into account that self-assessment of expertise is often influenced by personality and may not accurately reflect actual competence (Scheuermann et al., 2013).

Table 7-33 Participant performance by knowledge of logic, and by knowledge of OWL or another DL formalism

	Knowledge of logic				**Knowledge of OWL or other DL formalism**			
	accuracy		**time**		**accuracy**		**time**	
	N	%age correct	N	mean time (s.d.) - secs	N	%age correct	N	mean time (s.d.) - secs
No knowledge	2	51%	1	2484 (NA)	3	72%	2	2545 (429)
A little knowledge	3	80%	3	2087 (217)	8	58%	7	2573 (871)
some knowledge	16	65%	14	2522 (816)	9	67%	7	2401 (776)
expert knowledge	7	77%	6	1929 (768)	8	80%	8	1965 (700)

Table 7-34 Participant performance by usage of OWL or another DL formalism

	accuracy		**time**	
	N	%age correct	N	mean time (s.d.) - secs
no response (no use)	4	63%	4	2313 (491)
learning	1	47%	1	1556 (NA)
use in work	12	69%	11	2623 (732)
researching	5	84%	5	1703 (611)
other	6	63%	3	2485 (1057)

Table 7-35 Participant performance by relationship with English

	accuracy			time	
English is:	N	%age correct		N	mean time (s.d.) - secs
my main language	13	67%		11	2173 (800)
one of my main langs	8	71%		7	2378 (711)
not my main language	7	70%		6	2513 (832)

Inspection of the accuracy data suggests that relationship with English had no effect. This is confirmed by a chi-square test (χ^2 (2) = 1.7375, p = 0.420). Inspection of the response time data suggests that participants tended to take longer, the less dominant their English language skills were. However, an ANOVA revealed that this effect was not significant (F(2, 21) = 0.524, p = 0.600).

Table 7-36 shows the Spearman's rank correlation between the other three categories and both accuracy and response time. In all cases the tests were one-sided, since there was an expectation that increasing knowledge and experience would improve accuracy and reduce response time.

Table 7-36 Impact of participant profile on accuracy and response time – Spearman's rank correlation

	accuracy	**response time**
knowledge of logic	$\rho = 0.19$; p = 0.163	$\rho = -0.13$; p = 0.266
knowledge of OWL or other DL formalisms	$\rho = 0.41$; p = 0.015	$\rho = -0.29$; p = 0.087
usage of OWL or other DL formalisms	$\rho = 0.49$; p = 0.010	$\rho = -0.17$; p = 0.217

N.B. p-values not computed exactly because of ties; p-values represent one-sided test

The first point to note is that, whilst two of the three factors correlate significantly with accuracy, none of the factors correlate significantly with response time. The two factors which correlated significantly with accuracy, i.e. knowledge and usage of DL, were themselves significantly correlated (Spearman's $\rho = 0.74$, $p < 0.0001$, one-sided test)[140], so the effects cannot be regarded as completely independent. The lack of any significant performance correlation with knowledge of logic may be because of the non-uniform distribution between categories; well over half the participants were in the 'some knowledge' category. To better investigate the effect of prior knowledge, a broader range of participants would be needed. In fact, study participants were biased towards computer scientists and may not well represent domain experts who might be occasional users of ontologies.

7.9. Effect of question position

Previous sections have described how the questions in each section were presented in two reverse orders so as to mitigate any effect of position. It was also noted that, from inspection of the data in each section, there was no apparent effect of question order on accuracy. On the other hand, inspection of the response time data suggested that the initial question in each section often suffered a time penalty, and this was sometimes also true for the second

[140] This was based on 22 participants, comprising the 28 laboratory study participants for which accuracy data was available after removing the 6 participants in the 'other' category. When all 28 participants were used, the correlation was not significant ($\rho = 0.26$, p = 0.087).

question. This section investigates the effect of question position, first considering the effect on accuracy and then on response time.

7.9.1. Question position and accuracy

The analysis in this subsection does not make use of the full 28 data points relating to accuracy available from the laboratory study, but rather of the data from the 24 participants who also provided response time data. This is because this dataset was balanced to include each of the possible 24 permutations of the four section orders, and also each section was seen by half the participants in one order, and by the other half in the reverse order. This reduces the risk of confounding the difficulty of individual questions with the effect of position in the study.

On this basis, there is no evidence that the section position made any difference to the overall accuracy of the section (χ^2 (3) = 3.1119, p = 0.375). Moreover, taking all the questions together, there is no evidence that the position of a question within a section affected accuracy ($\chi^2(9)$ = 11.8586, p = 0.221).

On a section-by-section basis, there is no evidence of position in section significantly affecting accuracy (section on properties: χ^2 (7) = 5.832, p = 0.560; on Boolean operators: $\chi 2$ (9) = 16.2784, p = 0.061; on negation and restrictions: χ^2 (7) = 2.4334, p = 0.932; and on nested restrictions: χ^2 (7) = 6.7447, p = 0.456).

7.9.2. Question position and response time

Table 7-37 shows the mean time for each question by section position. From inspection, it does appear that there is a tendency for reduced response time as the participant proceeds through the study. An ANOVA of time versus section position revealed that there was a significant effect (F(3, 812) = 4.82, p = 0.002). A subsequent Tukey's HSD analysis revealed a significant difference between the first and third section positions to be encountered (p = 0.005) and the first and fourth section positions (p = 0.018). By ensuring that each of the 24 permutations of section order is seen by one participant, the design of the study is intended to compensate for this effect.

Table 7-37 Mean response time by section position

Section position	1	2	3	4
mean time (s.d.) - secs	73 (51)	68 (50)	60 (44)	62 (47)

A regression analysis of log response time versus question position showed a significant decline in response time with question position (F(1, 814) = 77.76, p < 0.001). When the regression analysis was repeated on a section-by-section basis, there was a significant effect for the sections on: Boolean operators (F(1, 238) = 64.37, p < 0.001); negation and restrictions (F(1, 190) = 11.32, p < 0.001); and multiple restrictions (F(1, 190) = 47.3, p < 0.001), but not for the section on object properties (F(1, 190) = 3.606, p = 0.059). In every case, the effect was a decline in mean response time with increasing position in the section.

Figure 7-1 to Figure 7.4 show the effect for each section. For every position there are two points, representing the mean time for the two questions asked at each position, i.e. in order 1 and order 2. Note also that, because the regression was performed on log time, the line is not strictly a straight line. However, in each case it is close to being straight. Finally, as can be seen from the degrees of freedom quoted above for the F-values, the regression analysis

was performed on the individual data from each participant, not the means displayed in the figures.

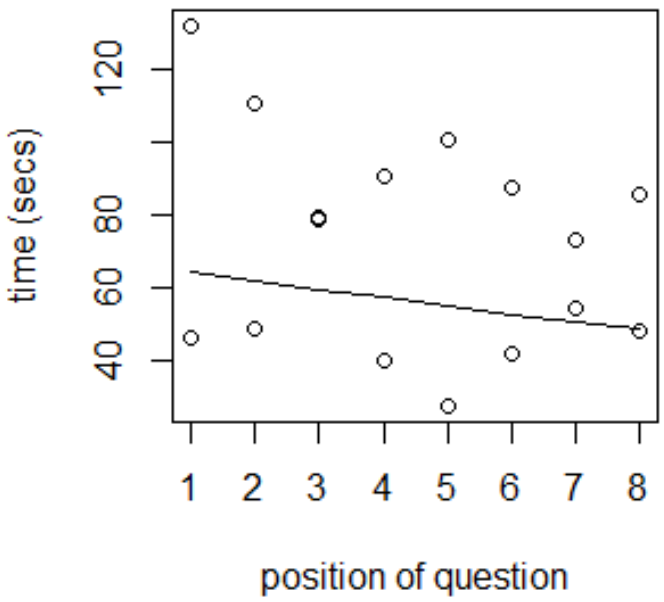

Figure 7-1 Decline in mean response time with position – object properties section

N.B. the appearance of only one point at position 3 is because of overprinting; the two points are very close together.

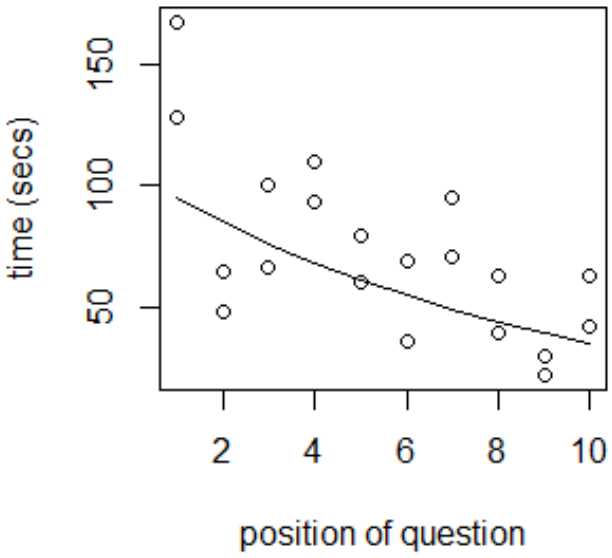

Figure 7-2 Decline in mean response time with position – Boolean operators section

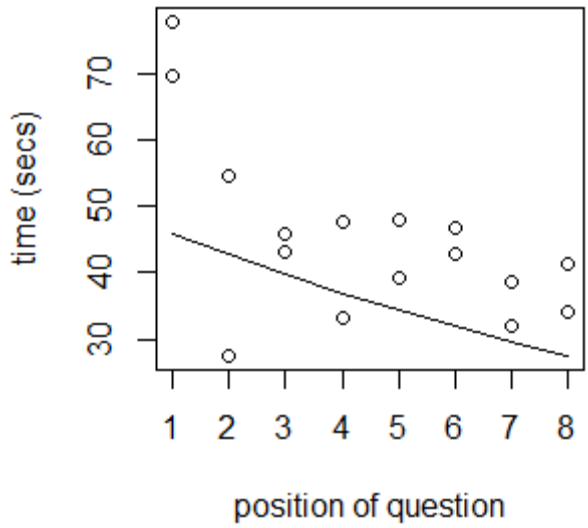

Figure 7-3 Decline in mean response time with position – negation and restrictions section

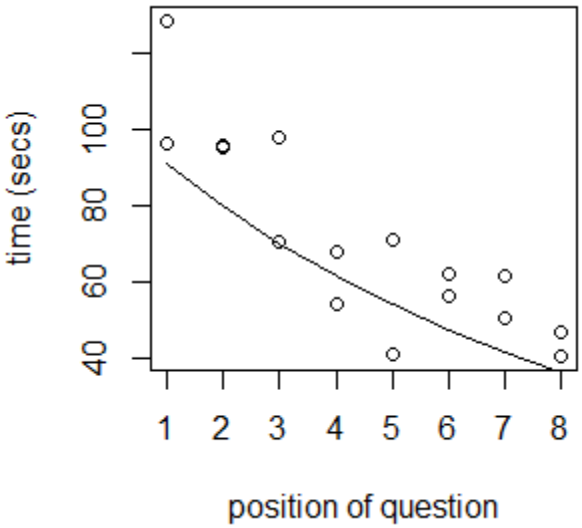

Figure 7-4 Decline in mean response time with position – nested restrictions section

All four figures suggest, to a greater or lesser extent, that there is an initial question penalty; in each case at least one of the initial questions appears to be appreciably above the regression line. To an extent this is compensated for by the reversal of the question order. A better strategy would be to randomize the question order, as was done in the third study.

7.10. Varying effects of valid and non-valid putative conclusions

It is possible that participants have a bias towards a particular answer. For example, it may be that participants answer 'not valid' whenever they are unable to prove validity, even when they cannot be sure that the answer is not valid. Subsection 7.10.1 investigates this possibility.

In the previous study, described in Chapter 6, there was some evidence that the questions with non-valid putative conclusions take longer to answer than the questions with valid

putative conclusions. This study offers the opportunity to investigate this more thoroughly, and this is done in subsection 7.10.2.

7.10.1. Validity and accuracy

Considering all the responses from the 28 participants who provided accuracy data in the laboratory study, the questions with valid conclusions were answered correctly 67.6% of the time, whilst the questions with non-valid conclusions were answered correctly 69.4% of the time. This difference is not significant ($p = 0.576$, Fisher's Exact Test, two-sided).

There were 16 questions with valid conclusions and 18 questions with non-valid conclusions. It might be that, on average, the two sets of questions were of differing difficulty. However, there were eight pairs of questions where the two questions shared the same axioms but had different conclusions; one valid and one non-valid. These questions occurred in Sections 7.3 and 7.4. In Table 7-5 all the questions occurred in such pairs, in Table 7-10 all but questions 1 and 2 occurred in such pairs.

Taking just these 16 questions, the percentage of correct responses to the 8 questions with valid conclusions is 69.3%, and to the 8 questions with non-valid conclusions is 77.6%. This difference is not significant ($p = 0.083$, Fisher's Exact Test, two-sided).

Two caveats need to be made. Firstly, this analysis is limited to two of the four sections, in particular excluding the two sections using restrictions, where each question had different axioms. Secondly, the difficulty of the questions with non-valid conclusions depends upon the plausibility of the putative conclusions. The conclusions were designed to be highly plausible, i.e. *prima facie* as plausible as the valid conclusions. It would have been possibly to have created much less plausible non-valid conclusions, which would have made these questions much easier and hence increased the apparent bias. On the other hand, with greater skill on the part of the author, it might have been possible to have created more plausible non-valid conclusions.

7.10.2. Validity and response time

As with the previous study, the mean response time for the valid questions (63 seconds) was less than for the non-valid questions (68 seconds). This was also the case when the comparison was limited to the correctly answered questions, although not when the comparison was limited to the incorrectly answered questions[141]. A two-factor analysis indicated a significant dependence on both validity ($F(1, 812) = 4.2682$, $p = 0.039$) and accuracy ($F(1, 812) = 11.9038$, $p < 0.001$) with a significant interaction effect ($F(1, 812) = 10.2729$, $p = 0.001$). A subsequent Tukey HSD indicated a significant difference between the times for the valid correct responses and non-valid correct responses ($p = 0.004$) and between the valid correct responses and both categories of incorrect responses ($p < 0.001$ valid incorrect; $p = 0.013$ non-valid incorrect). The other three pairwise comparisons were not significant ($p >$ at least 0.1 in all cases). In particular, this means that whilst there is a significant difference between the valid and non-valid response times for the correct responses, there is no significant difference between the valid and non-valid responses for the incorrect responses[142].

[141] The mean times are: valid / correct: 55.7 secs; valid / incorrect: 80.7 secs; non-valid / correct: 67.8 secs; non-valid / incorrect: 68.0 secs.

[142] In interpreting these results it should be remembered that there were fewer incorrect than correct responses. This may influence significance.

However, in both these studies the questions were not balanced to enable a proper comparison. The intrinsic differences in the questions might have been the reason for the differences observed. As with accuracy, the eight pairs of questions identified in the previous subsection offer a more controlled way to investigate the relationship between response time and validity of question. These 16 questions provide 384 response times from the 24 participants. These represent 192 pairs of responses, each pair consisting of responses from the same participant and from questions with the same axioms, one with a valid putative conclusion, one with a non-valid putative conclusion.

Of these 192 pairs, there were 103 pairs in which both questions had been answered correctly. This enables a comparison of accurately answered questions with valid conclusions and accurately answered questions with non-valid conclusions. Moreover, a paired t-test can be used in which the pairing is both by the question axioms and participant, balancing any variation in difficulty of the axioms and any variation in facility of the participants. Whilst the mean response time for the 103 questions with valid conclusions (67.0 seconds) was less than for the 103 questions with non-valid conclusions (70.6 seconds), this difference was not significant ($t(102) = 0.71797$, $p = 0.474$, two-sided paired test)

7.10.3. Conclusions

The first study suggested that correct responses to valid questions were significantly faster than correct responses to non-valid questions, suggesting that a frequent strategy was to begin by attempting to prove validity.

However, a controlled comparison in this study revealed no significant difference between the valid and non-valid questions, neither in accuracy nor in response time. This controlled comparison was necessarily limited to the questions in the two sections concerned with object properties and Boolean operators, which were the two sections with the highest proportions of correct responses. Given the nature of the questions, it was not possible to conduct a similar analysis for the two sections with restrictions, which participants found harder.

More extensive controlled comparisons are required to investigate further, ideally augmented by interviews with participants to understand how they arrive at conclusions in particular cases.

7.11. Discussion

This study set out to investigate further the difficulties detected in the previous study, and at the same time to extend the investigation to the interaction of negation with existential and universal restrictions, and also the effect of nested restrictions. As with the previous study, the intention was to use the theoretical framework discussed in Chapter 3 to understand participant behaviour. Subsection 7.11.1 lists the main findings, whilst subsection 7.11.2 discusses the threats to validity.

7.11.1. Main findings of the study

Taking the sections of the study in turn, the main findings were as follows.

For the section on object properties:

- After controlling for relational complexity, reasoning about functionality takes longer and is less accurate than reasoning about transitivity; although the effects are limited to the valid questions. Participant feedback supported this; several participants commented

that the transitivity questions were the easiest. Reasoning about functionality in situations of RC 4 is less accurate and takes longer than reasoning about functionality in situations of RC 3.

For the section on Boolean operators:

- Different algebraic formulations of semantically equivalent axioms made no difference to the accuracy or response time for the questions with valid conclusions. For the questions with non-valid conclusions, there was an increase in accuracy as the algebraic formulation approached DNF, whilst the response time was significantly greater for DNF than for the form intermediate between DNF and the Rector (2003) form.

For the two sections on restrictions, taken together:

- Where possible, participants appear to be using syntactic clues, perhaps to avoid creating detailed models of the situation. This led to some questions being answered relatively accurately and rapidly. One participant commented "some questions were easy to solve … just looking at the general form of the formulas". On the other hand, in other cases participants appear to be misled by apparent, but false, clues, e.g. by confusing *some ... not* with *not ... some*. Where participants were using mental models to represent restrictions, there appeared to be a propensity to start with the most obvious of the two models. This sometimes led to the second model being ignored, in particular in the case of the universal restriction, where the possibility of its trivial satisfaction may sometimes not have been taken into account. On other occasions, backtracking led to use of the second model and arrival at the correct conclusion.

In some cases, the complexity of a question's mental model representation was an indicator of question difficulty. However, in other cases questions with the same mental model but different syntactic structures were answered quite differently.

Apart from specific comments like those quoted above, participant feedback shed light generally on how they were reasoning about the questions. There was evidence of both visual and syntactic techniques being used. There was also evidence that some participants were using prior understanding, e.g. associated with a property name, rather than reasoning purely from the axioms. This is likely to be how many people do reason when working with real ontologies.

Whilst the majority of participants had a good knowledge of logic, there was a greater variation in terms of knowledge and previous usage of DL. Those participants with more knowledge or previous usage of DL tended to be more accurate in their responses. Since the primary concern of the study was to compare questions, and the constructs used in the questions, the varying competences of the participants might be regarded as irrelevant. On the other hand, it is likely that experienced ontologists with a good understanding of logic will have different problems and need different support than, e.g. domain experts.

Question position did not appear to have any effect on accuracy. However, there was a significant effect on response time, sometimes with an apparent time penalty for the first or second questions in a section. The strategy of having half the questions in one order, and the other half in reverse order may not have been adequate to compensate for this. As a consequence, in certain cases only part of the response time data was used. This emphasizes the importance of randomizing the question order.

Finally, there was no significant difference in accuracy between the valid and non-valid questions; neither when all questions were considered nor when balanced pairs of valid and non-valid questions were analysed. When all the questions were considered, there was a significant difference in response time, with the non-valid questions taking longer. However, a controlled comparison revealed no significant difference in response time. The controlled comparison did not include questions employing restrictions; further investigation is needed, as part of research to understand people's reasoning strategies when faced with DL and other logic languages.

7.11.2. Threats to validity

Threats to construct validity

The same comments apply here as were made in subsection 6.12.3. Responses were recorded automatically. However, response time was recorded manually, leading to possible errors of ±2 seconds; this is relatively small compared with the response times being measured.

Threats to internal validity

In study 1 the section order was varied between participants to compensate for any inter-section effect. However, the question order within sections was not varied. Study 1 gave no indication that accuracy was dependent on question position, and this was confirmed for study 2 by the analysis in Section 7.9. However, study 1 did suggest that reponse time depended on question position within section.

For study 2, in the laboratory study, the questions were again permuted to compensate for inter-section effects. Moreover, participants were divided into two groups and each group saw mutually reverse question orders within sections. In this way, it was hoped to compensate for intra-section effects.

Figures 7-1 to 7-4 indicate an approximately linear decline in response time with increasing question order. However, in some cases there is also evidence that there was an appreciable time premium for the first question, which was not fully compensated for by the reversal of question order. In order to avoid any bias due to this time premium, in some of the response time analyses, only part of the data was used.

In the online study question order was less critical, since this study was only used to collect accuracy data. For the oneline study, each section could be undertaken in a separate session, and thus the experimenter had no control over inter-section effects. However, the order of questions in each section was randomized, to compensate for any possible intra-section effects.

In study 3, to be described in the next chapter, both section order and question order within section was randomized to better compensate for any inter or intra-section effects.

Threats to external validity

Subsection 6.12.3 noted that the generalizability of the study 1 results might be limited by the nature of the questions and the profile of the participants. As with study 1, study 2 employed constructs which are known to be commonly used. In study 1, those constructs were set within the context of commonly used patterns. In order to make comparisons between different combinations of constructs, study 2 did not use such patterns, and in some cases this might be taken to detract from ecological validity. However, these combinations

of constructs identified certain general principles, e.g. the use of syntactic clues and the failure to make use of the less salient mental model for the universal and existential restrictions.

Subsection 6.12.3 also noted that the sample of participants in study 1 was biased towards computer scientists and not representative of domain experts. In study 2, there was a slightly wider range of participants. Comparison of Tables 6-12 and 7-33 shows that in study 2, unlike in study 2, there were some participants who claimed no knowledge of logic or DLs. However, the majority of participants had at least some knowledge of logic and DLs.

8. Study 3: Modifications to Manchester OWL Syntax

> *It is a commonplace of philosophical logic that there are, or appear to be, divergences in meaning between … the FORMAL devices … and … what are taken to be their analogs or counterparts in natural language.*
>
> *H.P. Grice, 'Logic and Conversation', 1975*

The studies described in the previous two chapters, have attempted to understand the difficulties experienced with DLs and interpret those difficulties in terms of the theories of reasoning developed by cognitive psychologists. The study described in this chapter builds on that understanding to develop and trial additions to MOS which might mitigate these difficulties. It also draws upon the concept of *implicature*, mentioned in Chapter 3 and first developed by Grice in the paper from which the above quotation is taken (Grice, 1975).

Section 8.1 provides an overview of the study and how it was conducted. As with the previous two studies, the response time data was transformed using the logarithmic transformation, and a justification for this is provided in section 8.2. The study was organised into four parts, analogous to study 2, and sections 8.3 to 8.6 describe each of these parts and the main findings. Section 8.7 reports on participant feedback and section 8.8 analyses the effect of participant experience and prior knowledge. Section 8.9 analyses the effect of question position. Section 8.10 investigates differences in accuracy (i.e. proportion of correct responses) and response time between the valid and non-valid questions. Finally, section 8.11 reviews the main findings. The study's main findings are also described in Warren et al. (2017).

8.1. Organisation of the study

The four parts of the study were concerned with: functional and inverse functional object properties; Boolean concept constructors, in particular negated conjunction; negation and restriction; and nested restrictions. With a few exceptions explained in later sections, the questions were analogous to questions in study 2, to enable the comparison with the results of that study. Note that all the comparisons of accuracy with study 2 made in this chapter are based on the study 2 data collected under laboratory conditions, i.e. not including the data from the online part of the study. This was to maintain consistency with study 3.

Each part contained eight questions with the same format as in the previous studies. Participants were presented with a set of axioms and a putative conclusion and asked if the conclusion was valid or non-valid. There were two variants of the study, with some of the questions differing between the variants. There were 30 participants, 15 for each variant, drawn from the Open University, another U.K. university and an industrial research laboratory. Participants were allocated alternately between the two variants. The study used the same simplified form of MOS as in the other two studies, except that in certain cases different keywords were used. In one case, that of *solely* described in section 8.3, the keyword was additional to existing keywords. In the other cases the keywords were alternatives. As with study 2, but not study 1, all keywords in the questions were identified in blue, as opposed to the black type used for the identifiers. As with the previous two studies, participants were provided with a handout which they could read beforehand and was available to them during the study; there were different handouts for each version of the study. Also as with the previous studies, there was a preliminary section where participants

were invited to provide information about themselves, and a final section where they could provide feedback on the study.

The questions were presented to the participants on a lap-top computer using the MediaLab application from Empirisoft[143], which recorded both the response and the response time. Statistical analysis was conducted using the R statistical package (R Core Team, 2014).

The analysis of the previous study in section 7.9 indicated that response time, although not accuracy, depended significantly on the position of the section and the position of the question in the section. To avoid any bias due to this effect, the order of sections and order of questions within each section were randomized, using a facility available in MediaLab.

8.2. Response time data

As with the two previous studies, examination of the distribution of the response times showed a positive skew. Figure 8-1 shows the distribution of response times over all the 960 measurements (i.e. 30 participants each answering 32 questions).

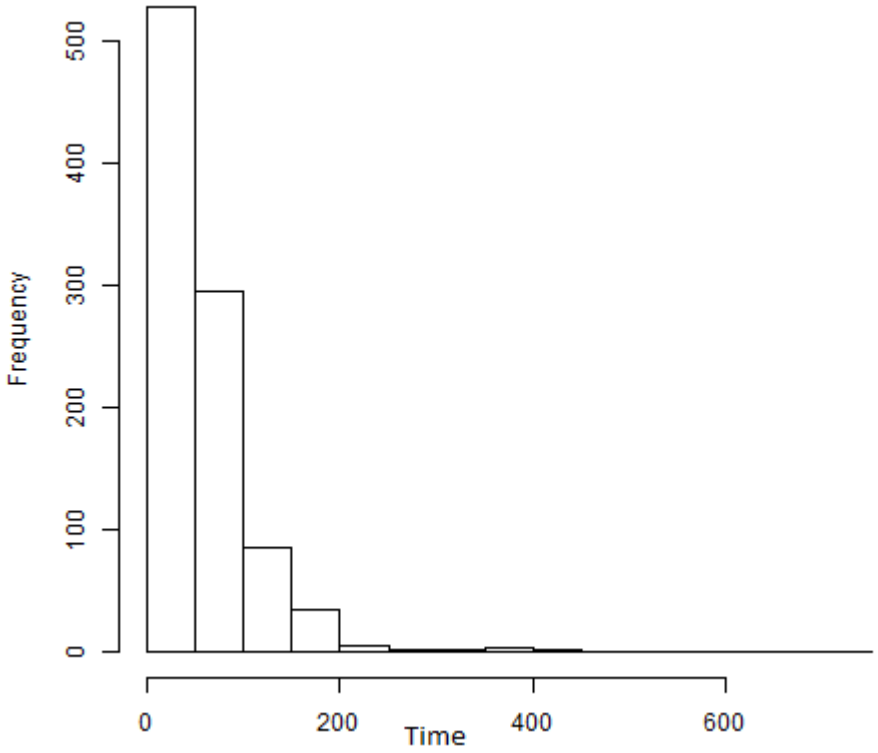

Figure 8-1 Frequency of response time; time shown in seconds

A Shapiro-Wilk normality test gave a p-value of less than 2.2×10^{-16}. Moving along Tukey's ladder of powers (Scott, 2012), the best fit to normality was achieved with a log transformation, which resulted in a Shapiro-Wilk p-value of 0.00726. Figure 8-2 shows the distribution of the response times after a $\log_{10}$ transformation.

[143] www.empirisoft.com

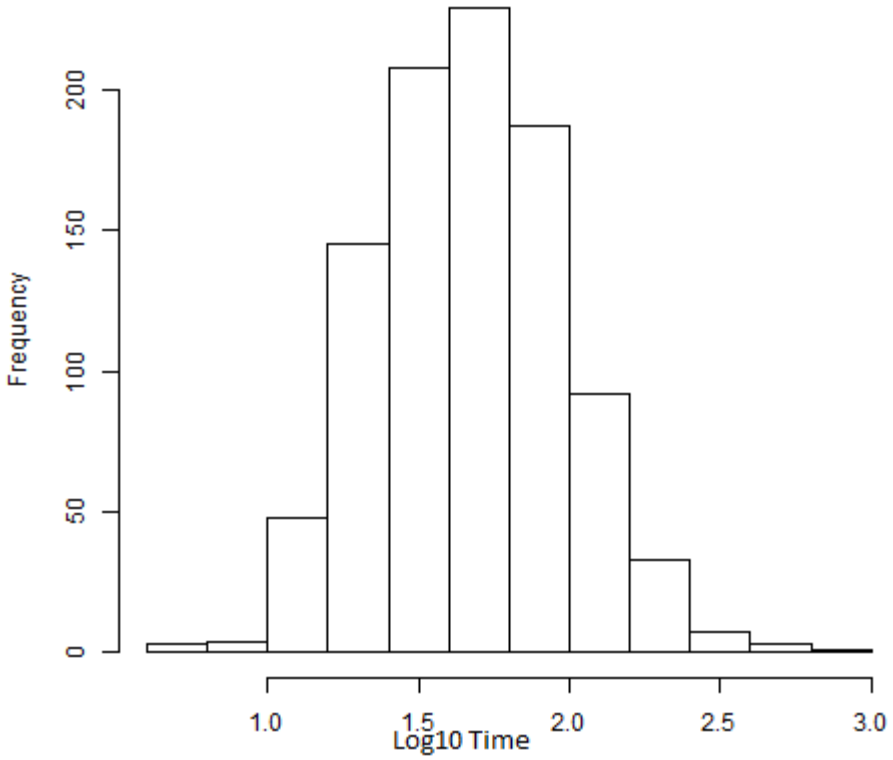

Figure 8-2 Frequency of response time; shown after $\log_{10}$ transformation

As in the previous studies, response time data was tested for each question using the Shapiro-Wilk test, resulting in a minimum p-value of 5.21×10^{-7} for the untransformed data. The two variants of each question were treated separately, making a total of 64 questions. The response time data was then transformed, using points from the Tukey ladder of powers, and in each case the Shapiro-Wilk test was repeated for each of the 64 questions. The minimum of the 64 p-values was noted. The objective was to obtain the transformation for which worst-case deviation from the normal (i.e. the worst case of the 64 questions) was minimized. This occurs when the minimum p-value from the 64 questions is greatest. Specifically, this happened when the logarithmic transformation was used, giving a minimum p-value of 0.00679. Although this is a relatively low p-value, for the logarithmic transformation there were only three other p-values below 0.05, and the great majority were above 0.1. This can be seen in Figure 8-3 and Figure 8-4 which show the Shapiro Wilk p-value for each question in each of the two variants before and after logarithmic transformation. Consequently, time was transformed using the logarithmic transformation prior to all ANOVA, t-tests and regression analyses reported here.

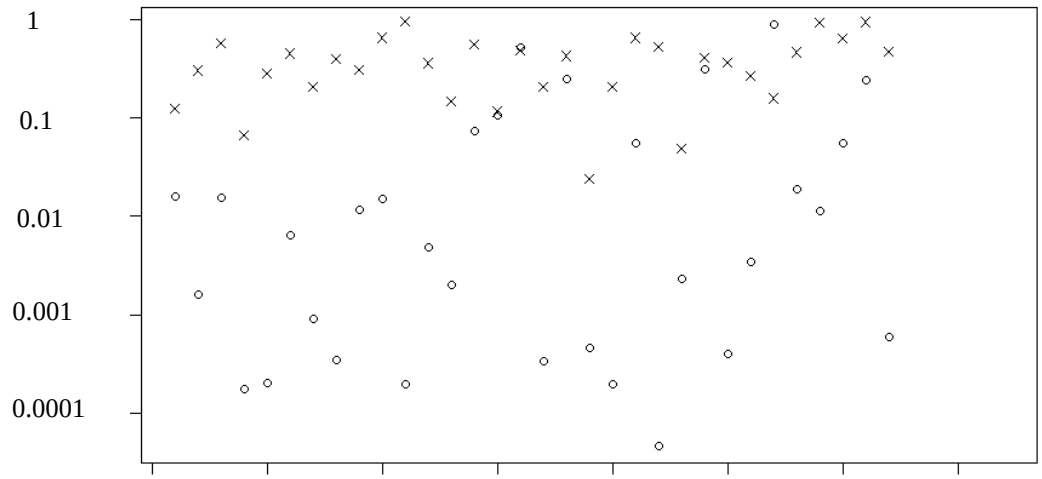

Figure 8-3 Shapiro-Wilk p-values for variant 1 questions, on a log scale. Circles indicate before logarithmic transformation, crosses after transformation.

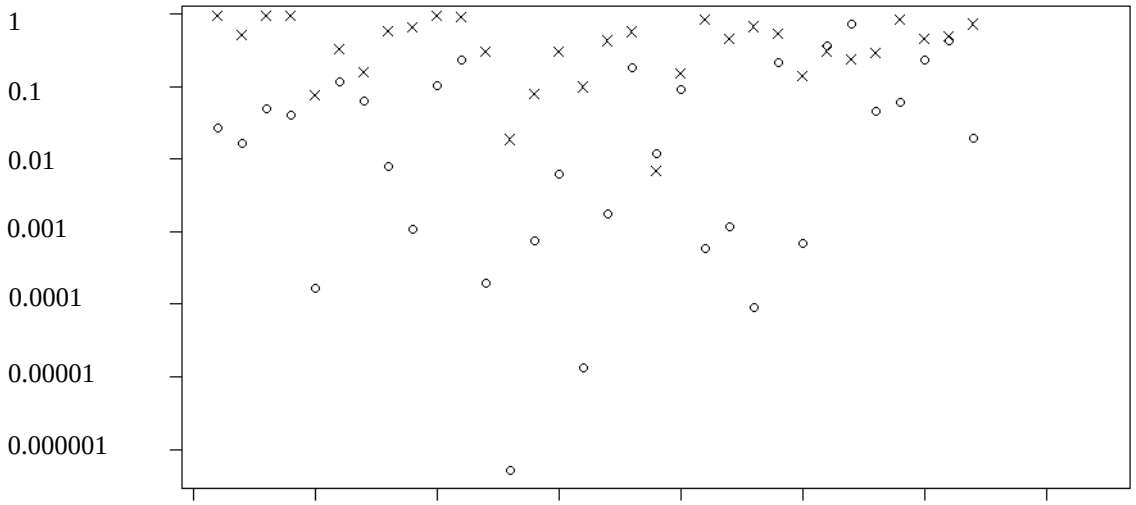

Figure 8-4 Shapiro-Wilk p-values for variant 2 questions, on a log scale. Circles indicate before logarithmic transformation, crosses after transformation.

8.3. Functional and inverse functional object properties

Study 2 concluded that, under conditions of equal relational complexity, reasoning about functionality took significantly longer than reasoning about transitivity, i.e. functionality is inherently harder than transitivity. It was hypothesized that one reason for this was the existence of confusion in participants' minds between functionality and inverse functionality, i.e. confusion about whether it is the subject or the object of the relation which is unique. In this study, six functional questions were created isomorphic to the six functional questions in study 2, i.e. questions 3 to 8 in Table 7-3. The resultant questions are shown in Table 8-1. A keyword, *solely*, was inserted before the object of the property to make clear that it is the object of the property which is unique. *solely* was chosen because of its anticipated power to convey uniqueness. A possibly more natural choice, *only*, was rejected because of its existing use in MOS. Unlike the other additions to MOS proposed in this study, *solely* is not an alternative to other keywords, but is additional. This leads to the hypothesis:

> H8.1 The introduction of the additional keyword *solely* between a functional property and its object will improve participant performance.

This hypothesis is investigated in subsection 8.3.1

The six functional questions were of varying complexity, and this enabled a further investigation of hypothesis H7.2, already discussed in section 7.3:

> H7.2 Reasoning about functionality becomes increasingly difficult with increasing relational complexity.

Another goal was to investigate whether performance with inverse functional properties could be improved by the use of a keyword to make clear that it is the subject of such properties which is unique. As neither of the previous studies included inverse functional properties, comparison of any syntactic changes could not be made between studies. Consequently, variant 1 of this study included two questions using an inverse functional proprty in our simplified version of MOS, whilst variant 2 included the same questions, except that the keyword *solely* was included before the subject to indicate its uniqueness. These questions are showin in Table 8-4. This leads to the hypothesis:

> H8.2 The introduction of the additional keyword *solely* before the subject of an inverse functional property will improve participant performance.

This hypothesis is investigated in subsection 8.3.3.

Finally, the availability of analogous questions using functional and inverse functional object properties enables a comparison of the relative difficulty of the two property characteristics. This is formalised as the hypothesis:

> H8.3 There will be a difference in performance between reasoning with functional and inverse functional properties.

This hypothesis is investigated in subsection 8.3.4.

8.3.1. Functional object properties – hypothesis H8.1

Table 8-1 shows the six study 3 questions using a functional object property. For these questions, there was no difference between variants 1 and 2, i.e. all 30 participants were presented with identical questions. As in the tables in the previous chapter, for brevity the tables in this chapter omit the declarative statements included in the actual questions. They also use semi-colons to separate some of the statements, rather than new lines as in the actual questions. The questions are grouped into three pairs, in order of increasing complexity, the questions in each pair having the same axioms but different putative conclusions, and arranged such that the odd-numbered questions are valid and the even-numbered are non-valid. Note that the numbering in Table 8-1 does not correspond to that in Table 7-3. Questions 1 to 6 in Table 8-1 are isomorphic to questions 3 to 8 in Table 7-3. The two sets of questions use different names. Specifically, the individual names *a, b, c, d, e, f* in the previous study have been mapped into *r, s, t, v, w, x* in this study. This was done to reduce the probability of any recollection from the previous study, which had taken place approximately twelve months earlier, since many of the participants in the two studies were the same. The property name used, *has_nearest_neighbour*, was the same as previously. For brevity, the property name is represented by F in the table. The table also shows the relational complexity of the valid questions, which is equal to the relational complexity of the analogous questions in the previous study.

Table 8-1 Functional object property questions

	axioms (F = *has_nearest_neighbour*)	**putative conclusion**	**validity**	**relational complexity**
1	r F solely s; r F solely t;	v sameAs w	valid	3,2,3
2	s F solely v; t F solely w	r sameAs t	not valid	n/a
3	r F solely s; r F solely t; v F solely s;	v DifferentFrom w	valid	3,2,4
4	w F solely x; t DifferentFrom x	r DifferentFrom v	not valid	n/a
5	r F solely s; t F solely v;	w DifferentFrom x	valid	4,2,4
6	s DifferentFrom v; w F solely r; x F solely z; t SameAs z	r DifferentFrom x	not valid	n/a

Table 8.2 shows the proportion of correct responses for each question. For comparison, the analogous data for the previous study is repeated from Chapter 7. For each of the valid questions, the proportion of correct responses was greater for the current study than for study 2. However, Fisher's Exact Test revealed no significant difference between the proportions of correct responses for the two studies at the level of individual questions. For each question, the mean time overall and the mean time for the correct responses were less for this study than for study 2. A comparison of the response times for each question using the t-test revealed a significant difference for question 1 ($t(51.251) = 2.0161$, $p = 0.049$) and question 5 ($t(44.063) = 2.2331$, $p = 0.031$), but no other significant differences.

Table 8-2 Functional object property questions – accuracy and response times

	study 2 – without *solely*				**study 3** – with *solely*			
	%age correct N = 28	mean time (SD) – secs N = 24			%age correct N = 30	mean time (SD) – secs N = 30		
		overall	correct	incorrect		overall	correct	incorrect
1	75%	52 (36)	56 (39)	36 (9)	83%	39 (31)	39 (32)	38 (26)
2	96%	61 (46)	62 (47)	35 (NA)	83%	50 (29)	52 (30)	43 (24)
3	61%	84 (67)	90 (66)	73 (70)	70%	58 (27)	52 (26)	71 (27)
4	79%	92 (66)	86 (62)	111 (80)	83%	78 (49)	77 (48)	81 (63)
5	43%	109 (79)	96 (55)	119 (96)	63%	73 (37)	76 (44)	67 (23)
6	71%	96 (47)	93 (50)	101 (43)	70%	90 (46)	92 (46)	86 (48)

The small sample sizes reduce the likelihood of significant differences for individual questions, particularly when comparing accuracy. A more realistic analysis is based on aggregated questions. Table 8-3 shows the percentage of correct responses and mean time to respond for all six functional questions aggregated and for the three valid and three non-valid questions aggregated. Using Fisher's Exact Test there was no significant difference in accuracy between the two studies when all questions were taken together ($p = 0.334$), nor for the aggregated valid questions ($p = 0.081$), nor the non-valid questions ($p = 0.703$).

A two-factor ANOVA indicated that response time varied significantly between the studies ($F(1, 320) = 7.559$, $p = 0.006$) and between the valid and non-valid questions ($F(1, 320) = 4.928$, $p = 0.027$), with no significant interaction ($F(1, 320) = 1.761$, $p = 0.185$). A subsequent Tukey HSD analysis revealed a significant difference in response time between the two studies for the valid questions ($p = 0.022$) but not for the non-valid ($p = 0.746$). This conclusion is supported when t-tests are applied to determine the effect of *solely* on the valid and non-valid questions separately. In both cases the mean response time is less when *solely* is used. In the case of the valid questions, the difference is significant ($t(144.7) = 2.8373$, $p = 0.005$). In the case of the non-valid questions, the difference is not significant ($t(148.61) = 1.0043$, $p = 0.317$).

In summary, there was partial support for hypothesis H8.1. The introduction of *solely* had no significant effect on accuracy. It did lead to a significant reduction in response time for the valid questions but had no significant effect on response time for the non-valid questions.

Table 8-3 Functional object property questions

	study 2 – without *solely*				**study 3 –** with *solely*			
	% corr	mean time (SD) - secs			% corr	mean time (SD) - secs		
		overall	correct	incorr		overall	correct	incorr
All questions	71%	83 (61)	78 (54)	93 (75)	76%	65 (41)	64 (42)	67 (38)
valid questions	60%	81 (67)	77 (55)	89 (83)	72%	57 (35)	54 (37)	63 (27)
non-valid questions	82%	83 (55)	78 (54)	101 (60)	79%	73 (45)	73 (45)	73 (49)

8.3.2. Effect of complexity – hypothesis H7.2

The data from questions 1 to 6 of this study enable a further investigation into the effect of complexity. There are three levels, represented in increasing complexity by questions {1, 2}, {3, 4} and {5, 6}. A logistic analysis showed no dependence of accuracy on complexity ($p = 0.102$). A one-way ANOVA showed that time to respond was significantly dependent on complexity ($F(2, 177) = 20.24$, $p < 0.001$). A subsequent Tukey HSD analysis showed that the difference in time was significant when comparing questions {1, 2} with {3, 4}, and when comparing questions {1, 2} and {5, 6} ($p < 0.001$ in both cases). The difference between questions {3, 4} and {5, 6} was not significant ($p = 0.080$).

In summary, there was partial support for hypothesis H7.2. Complexity did not significantly affect accuracy but did significantly affect response time. This contrasts with the results from the analogous questions in study 2, where complexity significantly affected both accuracy and response time.

8.3.3. Inverse functional object properties – hypothesis H8.2

Table 8-4 shows the format of the questions in the two variants. For brevity, I is used to represent the property *is_nearest_neighbour_of* used in this study. Questions 7 and 8 were generated from questions 1 and 5 of Table 8-1 by restating using *is_nearest_neighbour_of* rather than its inverse *has_nearest_neighbour*. Furthermore, to create question 7, *r* and *s* were swapped; and to create question 8, *r* was swapped with *s*, *t* with *v*, and *x* with *z*.

Table 8-4 Inverse functional object property questions

	axioms (I = *is_nearest_neighbour_of)*	**putative conclusion**	**validity**	**relational complexity**
variant 1				
7	r I s; t I s; v I r; w I t	v SameAs w	valid	3,2,3
8	r I s; t I v; r DifferentFrom t; s I w; x I z; v SameAs x	w DifferentFrom z	valid	4,2,4
variant 2				
7	solely r I s; solely t I s; solely v I r; solely w I t	v SameAs w	valid	3,2,3
8	solely r I s; solely t I v; r DifferentFrom t; solely s I w; solely x I z; v SameAs x	w DifferentFrom z	valid	4,2,4

Table 8-5 shows the percentage of correct responses and the mean response time for these questions. For question 8 the percentage of correct responses was the same in both variants. Using Fisher's Exact Test, there was no significant difference in accuracy of response

between the two variants, neither when the questions were aggregated ($p = 0.760$) nor for question 7 alone ($p = 0.651$). A two-way ANOVA showed that there was a significant difference in response time between the two questions ($F(1, 56) = 26.836$, $p < 0.001$), reflecting their difference in complexity, but not between the two variants ($F(1, 56) = 0.640$, $p = 0.427$); nor was there an interaction effect ($F(1,56) = 0.382$, $p = 0.539$).

Table 8-5 Inverse functional object property questions

	variant 1 – without solely; N = 15				**variant 2 –** with solely; N = 15			
	% corr	mean time (SD) - secs			% corr	mean time (SD) - secs		
		overall	correct	incorr		overall	correct	incorr
question 7	73%	38 (18)	33 (19)	51 (9)	87%	48 (23)	47 (25)	53 (9)
question 8	73%	105 (92)	115(107)	78 (16)	73%	90 (43)	89 (49)	94 (27)
both questions	73%	72 (74)	74 (86)	65 (19)	80%	69 (40)	66 (43)	80 (30)

In summary, there was no support for hypothesis H8.2. In the case of inverse functional properties, the use of the keyword solely has no significant effect on accuracy or response time. It may be that *solely* is taken to refer to the whole <subject, predicate, object> triple without making clear the uniqueness of the subject rather than object.

8.3.4. Difficulty of functional and inverse functional questions – hypothesis H8.3

The relationship between questions 1 and 7, and questions 5 and 8 enable investigation of hypothesis H8.3. For the functional questions in this study variants 1 and 2 were identical, both using *solely*. Hence a reasonable comparison is between:

A. questions 1 and 5 from study 2 and questions 7 and 8 from variant 1 of this study, i.e. the questions which do not use *solely*;
B. questions 1 and 5 from both variants of this study and questions 7 and 8 from variant 2, i.e. the questions which use *solely*.

Table 8-6 shows the relevant data for the two cases. In both cases a Fisher's Exact Test showed no significant difference in accuracy of response between the aggregated functional and inverse functional questions (case (A): $p = 0.241$; case (B): $p = 0.606$). Similarly, t-tests showed no significant difference in response time (case (A): $t(56.167) = 1.0853$, $p = 0.282$; case (B): $t(64.429) = 1.9113$, $p = 0.060$)

Table 8-6 Functional / inverse functional vs. with / without solely

	Case (A): without *solely*		**Case (B): with *solely***	
	%age corr	mean time (secs)	%age corr	mean time (secs)
Functional	59%	80	73%	56
Inverse functional	73%	72	80%	69

N.B. Case (A) compares questions 1 and 5 aggregated from study 2 with questions 7 and 8 aggregated from variant 1 of this study. Case (B) compares questions 1 and 5 aggregated from both variants of this study with questions 7 and 8 aggregated from variant 2 of this study.

In summary, there was no support for hypothesis H8.3. There was no significant difference, neither in accuracy nor response time, between the questions with functional and the questions with inverse functional properties.

8.3.5. Discussion

Table 8-7 summarises the conclusions for hypotheses H7.2, H8.1, H8.2 and H8.3. The key finding is that introduction of *solely* did significantly reduce response time for the functional object properties, but only in the case of the valid questions. However, *solely* had no effect for the inverse functional object properties. Other keywords and other choices of position might improve performance. One could, for example, experiment with the use of *alone* after the subject, e.g. *s alone has_nearest_neighbour r*. As was found in section 7.3, increasing complexity led to increased response time; although unlike in section 7.3, increasing complexity did not reduce accuracy significantly[144]. Finally, there was no apparent difference in performance between functional and inverse functional object properties.

Table 8-7 Object properties – summary of hypotheses

	accuracy	**response time**
H7.2 – effect of relational complexity on reasoning abut functionality	no significant difference	Response time increases significantly when a step with RC 3 is replaced by a step with RC 4.
H8.1 – use of *solely* with functional object properties	no significant difference	significantly reduces response time – valid questions only
H8.2 – use of *solely* with inverse functional object properties	no significant difference	
H8.3 – comparison of functional and inverse functional properties	no significant difference	

8.4. Boolean concept constructors

The eight questions in this part of the study were designed to test out whether amendments to MOS could reduce the difficulties experienced with Boolean concept constructors, in particular relating to negated conjunction.

Study 1 noted that negated conjunction was significantly harder than negated disjunction, as can be seen by comparing the data in Table 6-12 for questions 2 and 4. This effect has also been observed by Khemlani et al. (2012a). At the same time, it is known that *and* and *or* in everyday language are used ambiguously, e.g. see Mendonça et al. (1998). It was hypothesized that performance might be improved by the use of less ambiguous terminology for conjunction and disjunction. This leads to the following hypothesis:

> H8.4 The use of the keyword *intersection* in place of *and* will improve performance for negated conjunction.

[144] Note that the analysis in section 7.3 was based on 65 data points, i.e. over twice as many as were available in this study. As discussed in section 9.5, the reduced number of data points in study 3 means that a larger effect will be required to achieve significance. If the percentage of correct responses achieved at the three levels in study 3, aggregated over the two questions at each level, were achieved with a sample of 65 data points, then a logistic analysis of deviance would give a significant result ($p = 0.009$).

In study 2 the consecutive keywords *and not* were used to create exceptions, following a format used by Rector (2003). For this study, it was thought that *except* might be more intuitively understandable, including within nested exceptions, i.e. *and not (... and not ...)*, which give rise to negated conjunction. This is formalised as the hypothesis:

> H8.5 The use of *except* in place of *and not* will improve performance.

In some syntaxes, OWL already use the terms *intersection* and *union* as part of prefix operators. A relevant question is how such a prefix notation compares with the infix notation used in MOS. To investigate this, the following hypothesis was proposed:

> H8.6 There will be a difference in performance between the prefix notation *IntersectionOf()* and *UnionOf()*, and the infix notation *intersection* and *union*.

Table 8-8 shows the original questions from previous studies which were used to generate questions for this study and to investigate these three hypotheses. Note that the numbering differs from that used for studies 1 and 2 in Chapters 6 and 7, but is that used in this chapter for the analogous questions for study 3.

Questions 1 and 2 are analogous to questions 2 and 4 from study 1 in Table 6.11. As shown here they differ from the questions in study 1 in that the latter used an ontology pattern, whilst the questions here show the essential features abstracted from the original questions. This makes clearer the relationship with the questions in this study.

Questions 3 to 8 are taken from the study 2 questions shown in Table 7-10. To create these questions, six questions had to be chosen out of the ten original Boolean concept constructor questions. The rationale for the choice is explained in subsection 8.4.2 below.

Table 8-8 Boolean concept constructor questions as used in studies 1 and 2

	axioms	**putative conclusion**	**validity**
study 1			
1	Entity DisjointUnionOf Event, Abstract, Quality, Object; A Type Entity; A Type not (Event and Quality);	A Type (Abstract or Object)	not valid
2	Entity DisjointUnionOf Event, Abstract, Quality, Object; A Type Entity; A Type not (Event or Quality);	A Type (Abstract or Object)	valid
study 2			
3	Z EquivalentTo (TOP_CLASS and not A and not B); TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	valid
4	Z EquivalentTo (TOP_CLASS and not (A or B)); TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	valid
5	Z EquivalentTo (TOP_CLASS and not (A and not A_1)); TOP_CLASS DisjointUnionOf A, B; A DisjointUnionOf A_1, A_2	Z EquivalentTo (B or A_1)	valid
6	As for question 5	Z EquivalentTo B	not valid
7	Z EquivalentTo (TOP_CLASS and not (A and not (A_1 and not A_1_X))); TOP_CLASS DisjointUnionOf A, B; A DisjointUnionOf A_1, A_2; A_1 DisjointUnionOf A_1_X, A_1_Y	Z EquivalentTo (B or A_1_Y)	valid
8	As for question 7	Z EquivalentTo A_1_Y	not valid

To generate the questions for this study, variant 1 was formed by replacing *and* with *intersection*. This enables the investigation of H8.4. At the same time, for consistency, *or* was replaced with *union*. In questions 3 to 8 of variant 1, the consecutive keywords *and not* have been replaced with *except*. This enables the investigation of hypothesis H8.5.

To investigate hypothesis H8.6, variant 2 was formed by using *IntersectionOf()* for conjunction and *UnionOf()* for disjunction. In question 3, *IntersectionOf()* was used with three arguments; in other cases two arguments were used. In addition, to lessen the probability of any recollection of the previous study, the variable names were changed. In the first two questions: *Entity*, *Event*, *Abstract*, *Quality*, *Object* were changed to *UNIVERSE*, *W*, *X*, *Y*, *Z*; and *A* was represented lowercase. In the other six questions: *TOP_CLASS* was changed to *UNIVERSE*; *A*, *B* and *C* were changed to *X*, *Y*, *Z*; *Z* was changed to *W*; and *X* and *Y* were changed to *A* and *B*, thus *A_1_X* became *X_1_A*. Table 8-9 and Table 8-10 show the questions for variants 1 and 2 respectively.

Table *8-11* shows the percentage of correct responses for variants 1 and 2 of this study and the analogous questions from previous studies. The table also shows the percentage of correct responses for questions 3 to 8 aggregated. Table 8-12 similarly shows the mean response times.

Table 8-9 Boolean concept constructor questions – study 3: variant 1

	V / NV	axioms	putative conclusion
Derived from study 1			
1	NV	UNIVERSE DisjointUnionOf W, X, Y, Z; a Type UNIVERSE; a Type not (W intersection Y);	a Type (X union Z)
2	V	UNIVERSE DisjointUnionOf W, X, Y, Z; a Type UNIVERSE; a Type not (W union Y);	a Type (X union Z)
Derived from study 2			
3	V	W EquivalentTo ((UNIVERSE except X) except Y); UNIVERSE DisjointUnionOf X, Y, Z	W EquivalentTo Z
4	V	W EquivalentTo (UNIVERSE except (X union Y)); UNIVERSE DisjointUnionOf X, Y, Z	W EquivalentTo Z
5	V	W EquivalentTo (UNIVERSE except (X except X_1)); UNIVERSE DisjointUnionOf X, Y; X DisjointUnionOf X_1, X_2	W EquivalentTo (Y union X_1)
6	NV	As for question 5	W EquivalentTo Y
7	V	W EquivalentTo (UNIVERSE except (X except (X_1 except X_1_A))); UNIVERSE DisjointUnionOf X, Y; X DisjointUnionOf X_1, X_2; X_1 DisjointUnionOf X_1_A, X_1_B	W EquivalentTo (Y union X_1_B)
8	NV	As for question 7	W EquivalentTo X_1_B

Table 8-10 Boolean concept constructor questions – study3: variant 2

	V / NV	axioms	putative conclusion
Derived from study 1			
1	NV	UNIVERSE DisjointUnionOf (W, X, Y, Z)[145]; a Type UNIVERSE; a Type not IntersectionOf (W, Y);	a Type UnionOf (X, Z)
2	V	UNIVERSE DisjointUnionOf (W, X, Y, Z); a Type UNIVERSE; a Type not UnionOf (W, Y);	a Type UnionOf (X, Z)
Derived from study 2			
3	V	W EquivalentTo IntersectionOf (UNIVERSE, not X, not Y)); UNIVERSE DisjointUnionOf (X, Y, Z)	W EquivalentTo Z
4	V	W EquivalentTo IntersectionOf (UNIVERSE, not UnionOf(X, Y)); UNIVERSE DisjointUnionOf (X, Y, Z)	W EquivalentTo Z
5	V	W EquivalentTo IntersectionOf (UNIVERSE, not IntersectionOf (X, not X_1)); UNIVERSE DisjointUnionOf (X, Y); X DisjointUnionOf (X_1, X_2)	W EquivalentTo UnionOf (Y, X_1)
6	NV	As for question 5	W EquivalentTo Y
7	V	W EquivalentTo IntersectionOf (UNIVERSE, not IntersectionOf (X, not (IntersectionOf (X_1, not X_1_A)))); UNIVERSE DisjointUnionOf X, Y; X DisjointUnionOf (X_1, X_2); X_1 DisjointUnionOf (X_1_A, X_1_B)	W EquivalentTo UnionOf (Y, X_1_B)
8	NV	As for question 7	W EquivalentTo X_1_B

[145] For variant 2, *DisjointUnionOf* was followed by parentheses. This was done for consistency with *UnionOf()*. Variant 1 used the standard MOS format for *DisjointUnionOf*, without parentheses, see https://www.w3.org/TR/owl2-manchester-syntax/.

Table 8-11 Boolean concept constructor questions – percentage correct

	study 1; N = 12	**study 3: variant 1; N = 15**	**study 3: variant 2; N = 15**
1	25%	80%	67%
2	92%	87%	100%
	study 2; N = 28		
3	82%	100%	100%
4	86%	100%	93%
5	61%	53%	53%
6	64%	100%	73%
7	54%	60%	40%
8	68%	80%	60%
Q3 to 8	69%	82%	70%

Table 8-12 Boolean concept constructor questions – mean time (standard deviation)

	study 1; N = 12			**study 3: variant 1; N = 15**			**study 3: variant 2; N = 15**		
	overall	correct	incorr	overall	correct	incorr	overall	correct	incorr
1	75 (48)	65 (30)	79 (54)	53 (41)	48 (25)	73 (88)	47 (16)	44 (15)	53 (17)
2	44 (19)	42 (19)	62 (NA)	42 (28)	40 (30)	53 (7)	39 (25)	39 (25)	NA (NA)
	study 2; N = 24								
3	39 (26)	36 (23)	56 (37)	39 (25)	39 (25)	NA (NA)	47 (30)	47 (30)	NA (NA)
4	43 (29)	36 (24)	78 (27)	35 (26)	35 (26)	NA (NA)	46 (35)	44 (35)	69 (NA)
5	96 (56)	99 (64)	89 (37)	61 (37)	63 (49)	59 (19)	65 (38)	60 (34)	71 (45)
6	105 (78)	112 (78)	89 (79)	44 (25)	44 (25)	NA (NA)	82 (57)	81 (57)	86 (66)
7	90 (48)	91 (49)	89 (51)	97 (75)	99 (92)	94 (47)	156 (126)	128 (39)	174 (161)
8	94 (47)	94 (50)	95 (45)	88 (60)	97 (63)	52 (16)	93 (51)	89 (34)	98 (74)
Q3 to Q8	78 (56)	74 (58)	86 (50)	61 (50)	55 (48)	69 (42)	82 (74)	60 (42)	104 (103)

8.4.1. Negated conjunction and disjunction – hypotheses H8.4 and H8.6

Question 1 was designed to test H8.4 and also H8.6. In variant 1, *and* is replaced with *intersection* whilst in variant 2 the prefix form is used, *IntersectionOf()*. Question 2 was designed to determine what effect similar changes would have on negated disjunction, where performance in study 1 had been good. For this question, variant 1 uses *union* whilst variant 2 uses the prefix form, *UnionOf()*. Using a logistic analysis, the percentage of correct responses for question 1 varied significantly between the three cases, i.e. study 1 and the two variants of this study ($p = 0.011$). A subsequent Tukey HSD showed that the difference between study 1 and variant 1 was significant ($p = 0.020$), whilst there was no significant difference between study 1 and variant 2 ($p = 0.094$), nor between the two variants ($p = 0.691$). A one-way ANOVA indicated that there was no significant difference in response

time between the three cases, i.e. study 1 and variants 1 and 2 of this study ($F(2, 39) = 2.052$, $= 0.142$).

For question 2, a logistic analysis showed no significant variation in percentage of correct responses between the three cases ($p = 0.229$). A one-way ANOVA also showed no significant difference in response time for the three situations ($F(2, 39) = 0.4233$, $p = 0.658$). Thus the use of *union*, in infix and prefix form, made no difference to the previously high level of performance with negated disjunction.

In summary, there was partial support for hypothesis H8.4. For negated conjunction the use of *intersection*, in infix form, significantly improved the accuracy but not the response time. The use of *IntersectionOf()* made no significant difference to accuracy or response time. The use of *union*, in infix and prefix forms, made no significant difference to accuracy or response time with negated disjunction.

There was no support for hypothesis H8.6. Neither with negated conjunction nor negated disjunction was there a significant difference between infix and prefix forms.

8.4.2. Except – hypothesis H8.5

Questions 3 to 8 in variant 1 were intended to test the effect of replacing *and not* with *except*. As with questions 1 and 2, variant 2 was constructed to investigate whether participant performance would be different if the infix notation used in MOS for conjunction and disjunction was replaced with prefix notation. Study 2 included ten questions relating to Boolean concept constructors; the questions in this study are analogous to six of them. These questions were chosen to represent the three levels of complexity present in the propositional logic questions in study 2. For the questions of medium complexity, this involved selecting two questions from the six originally used in study 2. These are questions 5 and 6 in Table 8-8, and correspond to questions 3 and 4 in Table 7-11. They were chosen because the first axiom was similar to an expression used in Rector (2003), containing two occurrences of *and not* which each translate into *except*, i.e. they are the most visibly obvious statements of exceptions. Questions 5 and 6 in Table 7-11 make no use of *and not*, whilst questions 7 and 8 make only one use of *and not*.

Compared with study 2, the percentage of correct responses was greater in variant 1 for five of the six questions, whilst the percentage of correct responses was greater in variant 2 only for three questions. A logistic analysis using all six questions aggregated showed no significant variation between the three cases ($p = 0.052$). Thus, neither the use of *except* nor the use of prefix notation made a significant difference to accuracy.

A one-way ANOVA indicated that time varied significantly between the cases, ($F(2, 321) = 3.65$, $p = 0.027$). A subsequent Tukey HSD analysis indicated that there was a significant difference between study 2 and variant 1 of this study ($p = 0.044$), with variant 1 responses faster. There was no significant difference between the two variants ($p = 0.05002$), nor between study 2 and variant 2 ($p = 0.978$).

The Tukey HSD analysis showed no significant difference between the two variants; as already noted the p value was a little above 0.05 (0.05002). However, a t-test did reveal a significant difference between the variants ($t(176.22) = 2.3962$, $p = 0.018$). This difference between the Tukey HSD comparison and the t-test is to be expected. The Tukey HSD adjusts the reported p value to compensate for the multiple pairwise comparisons. Note also that inspection of Table 8-12 shows that the mean time for the six questions for variant 1 (61 seconds) was appreciably less than both variant 2 (82 seconds) and study 2 (78 seconds).

A strong caveat is required when interpreting the comparison of study 2 and variant 1 with regard to time. The rationale for choosing the two questions, questions 3 and 4 in Table 7.11, from the six questions of medium complexity has already been explained. However, these were the two questions which occupied position 1 in the two orderings of the section questions. Inspection of Table 7-14, in particular the difference between the mean times in the two orderings, shows that this may have biased the response times, making the comparison with this study more favourable. When, for the purposes of the statistical analysis, questions 3 and 4 are replaced by the semantically equivalent but syntactically different questions 5 and 6 (albeit confounding the use of *except* with the issue of differing axiom structure), then there is no significant difference in response times between study 2 and variant 1 ($t(192.67) = 0.96752$, $p = 0.335$).

In summary, there is no evidence to support hypothesis H8.5. The use of *except* in place of *and not* did not significantly affect accuracy. Moreover, on the basis of the evidence, it is not safe to conclude that *except* significantly reduces response time. However, it does present a strong case for further investigation. A final point to make here is that all the studies in this dissertation are concerned with comprehension. It may also be the case that the use of *except* has an advantage over *and not* when modelling with DL, since it corresponds more closely to the language generally used to discuss exceptions. This also needs to be investigated.

8.4.3. Discussion

Table 8-13 summarises the hypotheses discussed in this section. The major finding is that the use of *intersection*, in place of *and*, significantly increases accuracy. In addition, *except* appears to offer promise in reducing reasoning time; this requires further investigation.

Table 8-13 Boolean concept constructors – summary of hypotheses

	accuracy	**response time**
H8.4 – *intersection* in place of *and*	infix form significantly improves accuracy	no significant difference
H8.5 – *except* in place of *and not*	no significant difference	requires further investigation
H8.6 – infix versus prefix notation	no significant difference	

8.5. Negation and restrictions

Chapter 7 described the difficulties which study 2 participants experienced with the universal and existential restrictions, and related these difficulties in part to a failure to form both the required mental models. This and the following section of study 3 investigates whether the replacement of *only* with *noneOrOnly* and *some* with *including* would improve performance with these restrictions. *noneOrOnly* was intended to draw attention to the fact that, e.g. the class *has_child noneOrOnly MALE* includes those individuals who have no children at all. *including* was intended to draw attention to the fact that, e.g. the class *has_child including MALE* may contain individuals who have a female child in addition to a male one. Thus, the two proposed keywords were intended to draw attention to the second mental model in Table 7-24 and Table 7-32. The hypothesis to be investigated is:

> H8.7 The use of *noneOrOnly* in place of *only* and *including* in place of *some* will lead to improved participant performance.

Moreover, in study 2, two questions include the axiom *X SubClassOf not (has_child some MALE*). *not ... some* is not a natural English construct; a more natural construct is *not ... any*. It was thought that the use of the latter construct might aid comprehension and reasoning. This leads to another hypothesis:

> H8.8 The use of *any* to indicate the existential restriction, when the corresponding object property is preceded by a negation, will improve performance.

The original questions, as in study 2, were shown in Table 7-17. In this section the same numbering scheme is used as in Chapter 7. All questions have the same putative conclusion: *X DisjointWith Y.* As in Table 7-17, two different typefaces are used for the first axioms to indicate semantic equivalence of axioms and draw attention to the fact that the questions can be grouped into four semantically equivalent pairs: {1, 4}; {2, 3}; {5, 8}; {6, 7}. For this study these questions were modified by the replacements described above. For six of the questions there was no difference between the two variants. However, in their original form questions 3 and 7 contain the class description *not (has_child some MALE)*. In variant 1 some was replaced with *including* as in the other questions. In variant 2, any was used, i.e. *not (has_child any MALE).* As with the analogous questions in Chapter 7, all questions have the same putative conclusion: *X DisjointWith Y.* Table 8-14 shows the questions used in this study, including the two variants of questions 3 and 7. Table 8-15 shows the proportion of correct responses for the questions in study 2 and the current study. For the latter, separate data are shown for the two variants when the questions differ. The table also shows the data aggregated for all questions excluding questions 3 and 7, and aggregated for questions 3 and 7. Similarly, Table 8-16 shows the data for response times.

Table 8-14 Negation and restriction questions as used in study 3.

	first axiom	**second axiom**	**validity**
1	***A SubClassOf has_child including (not FEMALE)***	B SubClassOf has_child noneOrOnly FEMALE	valid
2	A SubClassOf has_child noneOrOnly (not FEMALE)		not valid
3	**variant 1**		not valid
	A SubClassOf not (has_child including FEMALE)		
	variant 2		
	A SubClassOf not (has_child any FEMALE)		
4	***A SubClassOf not (has_child only FEMALE)***		valid
5	***A SubClassOf has_child some (not FEMALE)***	B SubClassOf has_child including FEMALE	not valid
6	A SubClassOf has_child only (not FEMALE)		valid
7	**variant 1**		valid
	A SubClassOf not (has_child including FEMALE)		
	variant 2		
	A SubClassOf not (has_child any FEMALE)		
8	***A SubClassOf not (has_child only FEMALE)***		not valid

N.B. the putative conclusion in each case was A DisjointWith B.

Table 8-15 Negation and restriction questions – percentage correct

	study 2 N = 28	**study 3: both variants;** N = 30	**study 3: variant 1;** N = 15	**study 3: variant 2;** N = 15
	only, some	*noneOrOnly, including*		*not … any*
1	61%	80%		
2	50%	73%		
3	68%	70%	67%	73%
4	75%	90%		
5	64%	70%		
6	50%	70%		
7	79%	80%	73%	87%
8	68%	67%		
Exc Q3 and Q7	61%	75%		
Q3 and Q7	73%	75%	70%	80%

Table 8-16 Negation and restriction questions – mean time (standard deviation)

	study 2 N = 24			**study 3: both variants** N = 30			**study 3: variant 1** N = 15			**study 3: variant 2** N = 15		
	only, some			*noneOrOnly, including*						*not … any*		
	overall	correct	incorr	overall	correct	incorr	overall	correct	incorr	overall	correct	incorr
1	52 (39)	38 (24)	80 (50)	42 (33)	40 (35)	46 (30)						
2	33 (18)	32 (14)	34 (22)	29 (20)	30 (22)	26 (11)						
3	45 (22)	43 (24)	49 (16)	69 (127)	67 (149)	76 (57)	55 (49)	30 (22)	104 (53)	84 (175)	100 (203)	40 (43)
4	43 (25)	40 (24)	57 (24)	41 (32)	36 (19)	87 (83)						
5	41 (30)	42 (32)	38 (27)	30 (21)	27 (21)	36 (21)						
6	44 (40)	38 (25)	52 (55)	33 (33)	33 (38)	35 (18)						
7	43 (37)	34 (26)	79 (53)	29 (16)	30 (17)	26 (13)	26 (14)	25 (14)	26 (16)	33 (18)	33 (19)	27 (4)
8	60 (37)	61 (42)	58 (29)	38 (24)	39 (26)	35 (20)						
Exc Q3 and 7	45 (33)	42 (29)	52 (39)	35 (28)	34 (28)	39 (29)						
Q3 and 7	44 (30)	38 (25)	62 (37)	49 (92)	47 (103)	56 (50)	40 (38)	28 (17)	70 (57)	58 (125)	64 (139)	36 (34)

8.5.1. noneOrOnly and including – hypothesis H8.7

To avoid the confounding effect of the introduction of *not … any* in variant 2 for questions 3 and 7, an analysis was conducted based on the six questions excluding questions 3 and 7. Tables 8-15 and 8-16 show the mean results for these six questions. The use of *noneOrOnly* and *including* led to a significant increase in accuracy (Fisher's Exact Test, $p = 0.008$) and reduction in response time ($t(284.57) = 2.7897$, $p = 0.006$)[146]. There is an appreciable increase in accuracy for the two questions which were answered worst in study 2, i.e. questions 2 and 6. The former requires the second model for the universal restriction, i.e. the realisation that the universal restriction can be trivially satisfied.

It is possible to separate out the effects of *noneOrOnly* and *including*, although at the expense of reduced sample sizes. Questions 2 and 4 use only the universal restriction. For these questions, this study has a higher proportion of correct responses and a reduced mean

[146] In fact, the conclusion is the same if we retain questions 3 and 7. A Fisher's Exact Test gives $p = 0.015$ and the result for the t-test is $t(394.79) = 2.7744$, $p = 0.006$.

response time, compared to study 2. However, neither of these differences was significant ($p = 0.192$, Fisher's Exact Test, two-sided; $t(104.75) = 1.2697$, $p = 0.207$).

Questions 5 and 7 use only the existential restriction. However, the situation is complicated here by the fact that, in variant 2, question 7 uses the construction *not … any*. To avoid this confounding effect, a comparison can be made for these two questions between study 2 and variant 1 of this study. In this case the percentage of correct responses is the same in both cases (71%), but there is a significantly lower response time in study 3, variant 1 ($t(88.697) = 2.0266$, $p = 0.046$).

In summary, there was support for hypothesis H8.7. The use of *noneOrOnly* and *including* in place of *only* and *some* significantly increased accuracy and significantly reduced response time. Attempting to separate out the effects of the two changes appreciably reduces sample size. The only significant result when the two changes are separated, is that the replacement of some by *including* significantly reduced response time.

8.5.2. *not … any* – hypothesis H8.8

Questions 3 and 7 provide an opportunity to investigate the effect of using *not ... any* in variant 2. Inspection of the percentage of correct responses indicates little difference in accuracy between study 2 and the two variants of this study. A logistic analysis showed no significant difference in percentage of correct responses between the three cases ($p = 0.653$). A one-way ANOVA for the aggregated response time data from questions 3 and 7 also showed no variation across the three cases ($F(2, 105) = 0.602$, $p = 0.550$). Thus there is no evidence that the use of *not ... any* makes any significant difference to performance.

In summary, there was no support for hypothesis H8.8. The use of *not ... any* in place of *not ... some* had no significant effect on accuracy or response time.

8.5.3. Discussion

Table 8-17 summarises the hypotheses discussed in this section. Taken together, the use of *noneOrOnly* and *including* both significantly increase accuracy and reduce response time. More investigation is required to investigate the individual effects of *noneOrOnly* and *including*, but the evidence presented here indicates that *including* leads to a significant reduction in response time.

Table 8-17 Negation and restrictions – summary of hypotheses

	accuracy	**response time**
H8.7 - *noneOrOnly* and *including* in place of *only* and *some*	significant increase in accuracy	significant reduction in response time
H8.8 – *not ... any* in place of *not ... some*	no significant difference	

8.6. Nested restrictions

Study 2 included eight questions making use of nested restrictions, as shown in Table 7-25. In variant 1 of this study, analogous questions to the study 2 questions were created by replacing *only* with *noneOrOnly* and *some* with *including*. This provided another opportunity to investigate hypotheses H8.7. Table 8-18 shows the questions used in variant 1.

In study 2, four of the questions used a named class Y, whilst in the other four questions the corresponding class is anonymous. This enabled an investigation of hypothesis H7.12:

> H7.12 There will be a difference in reasoning performance between the equivalent use of a named class and an anonymous class.

In study 2 there was no significant difference in accuracy between the questions with two axioms joined with a named class and the questions with one axiom; for response time the analysis was more ambiguous. However, the two sets of questions were different, thereby potentially biaising the results of the analysis. In this study, variant 1 used named and anonymous classes as in the original study. This enabled the controlled comparison between study 2 and variant 1 described in subsection 8.6.1. However, for variant 2, questions 1 to 4 reversed the usage, i.e. questions 1 and 3 used a named class whilst questions 2 and 4 used an anonymous class. This enabled further investigation of hypothesis H7.12 with a controlled comparison between the use of named and anonymous classes.

In study 2, the final axiom of questions 5 to 8 included the construct *not ... some*. For this study, in variant 2 this was replaced by *not ... any*, whilst *only* and *some* were replaced with *noneOrOnly* and *including*, as in variant 1. Thus, for these questions, comparison of the two variants allows further investigation of hypothesis H8.8. Table 8-19 shows the questions used in variant 2.

Table 8-18 Questions employing nested restrictions; study 3: variant 1.

	first axiom(s)	**final axiom**	**valid**
1	A SubClassOf (has_child including (has_child including MALE))	x has_child y; y Type has_child including (not MALE)	not valid
2	A SubClassOf has_child including B; B EquivalentTo has_child noneOrOnly MALE		not valid
3	A SubClassOf (has_child noneOrOnly (has_child including MALE))		not valid
4	A SubClassOf has_child noneOrOnly B; B EquivalentTo has_child noneOrOnly MALE		valid
5	A SubClassOf has_child including B; B EquivalentTo has_child including MALE	x has_child y; y Type (not (has_child including MALE))	not valid
6	A SubClassOf (has_child including (noneOrOnly MALE))		not valid
7	A SubClassOf has_child noneOrOnly B; B EquivalentTo has_child including MALE		valid
8	A SubClassOf (has_child noneOrOnly (has_child noneOrOnly MALE))		not valid

N.B. the putative conclusion in each case was x Type (not A).

Table 8-19 Questions employing nested restrictions; study 3: variant 2.

	first axiom(s)	**final axiom**	**valid**
1	A SubClassOf has_child including B B EquivalentTo has_child including MALE	x has_child y; y Type has_child including (not MALE)	not valid
2	A SubClassOf (has_child including (has_child noneOrOnly MALE))		not valid
3	A SubClassOf has_child noneOrOnly B B EquivalentTo has_child including MALE		not valid
4	A SubClassOf (has_child noneOrOnly (has_child noneOrOnly MALE))		valid
5	A SubClassOf has_child including B; B EquivalentTo has_child including MALE	x has_child y; y Type (not (has_child any MALE))	not valid
6	A SubClassOf (has_child including (noneOrOnly MALE))		not valid
7	A SubClassOf has_child noneOrOnly B; B EquivalentTo has_child including MALE		valid
8	A SubClassOf (has_child noneOrOnly (has_child noneOrOnly MALE))		not valid

N.B. the putative conclusion in each case was x Type (not A).

Table 8-20 and Table 8-21 show the percentage of correct responses and the mean response times for each question for study 2 and each of the variants of study 3. The table also shows the mean aggregated data for all the questions and for questions 5 to 8. The aggregated data are not shown for variant 2, as the differing processes for generating questions 1 to 4 and questions 5 to 8 in variant 2 would not make this data meaningful.

Table 8-20 Nested restrictions questions – percentage correct

	study 2 N = 28	**current study: var 1** N = 15	**current study: var 2** N = 15
	only, some	*noneOrOnly, including*	
1	71%	80%	60%
2	57%	40%	67%
3	71%	60%	53%
4	57%	53%	47%
			not ... any
5	54%	40%	67%
6	64%	73%	73%
7	71%	53%	47%
8	50%	60%	53%
Mean for Q5 to 8	60%	57%	60%
Mean for all questions	62%	58%	

Table 8-21 Nested restrictions questions – mean time (standard deviation)

	study 2 N = 24			**current study: var 1** N = 15			**current study: var 2** N = 15		
	only, some			*noneOrOnly, including*					
	overall	correct	incorr	overall	correct	incorr	overall	correct	incorr
1	69 (45)	71 (46)	60 (44)	47 (30)	47 (34)	44 (6)	73 (59)	71 (66)	75 (53)
2	79 (53)	101(55)	52 (37)	65 (20)	67 (18)	64 (23)	68 (41)	74 (43)	57 (36)
3	63 (43)	63 (46)	65 (38)	54 (31)	45 (13)	68 (45)	79 (46)	92 (44)	64 (46)
4	63 (39)	56 (27)	72 (52)	100(87)	107 (112)	92 (53)	66 (49)	69 (41)	63 (58)
5	88 (62)	94 (68)	78 (53)	64 (30)	72 (42)	58 (19)	74 (67)	89 (75)	42 (30)
6	73 (45)	66 (44)	84 (46)	85 (82)	99 (93)	47 (18)	99 (90)	97 (88)	105 (108)
7	80 (36)	71 (34)	108(29)	64 (39)	65 (39)	63 (41)	97(102)	131 (140)	67 (43)
8	55 (30)	50 (23)	59 (36)	83 (35)	87 (36)	77 (36)	63 (37)	70 (42)	54 (30)
Mean for Q5 to Q8	74 (46)	71 (47)	78 (44)	74 (51)	83 (61)	62 (30)	83 (77)	96 (88)	64 (54)
Mean for all quests	71 (45)	71 (46)	71 (44)	70 (51)	73 (61)	66 (35)			

8.6.1. noneOrOnly and including – hypothesis H8.7

A Fisher's Exact Test comparing study 2 with variant 1 of study 3, aggregated over all questions, showed no significant difference in accuracy ($p = 0.420$), i.e. the use of *noneOrOnly* and *including* made no difference to accuracy. A t-test comparison of study 2 and variant 1 response times, aggregated over all questions, also showed no significant difference ($t(288.79) = 0.40724$, $p = 0.684$). This contrasts with the effect reported in subsection 8.4.1. In particular, the use of *including* did not appreciably improve questions 2 and 5, which require the understanding that the first *including* stipulates a relationship (*has_child*) with a certain individual, but does not preclude the possibility of the same relationship with other individuals. Similarly, the use of the second *noneOrOnly* in question 8 does not appear to have created the awareness that there might be no child at all.

In summary, in contrast to the results presented in subsection 8.5.1, there was no support for hypothesis H8.7. The replacement of *only* and *some* with *noneOrOnly* and *including* had no significant effect on accuracy or response time.

8.6.2. Anonymous and named classes – hypothesis H7.12

Questions 1 to 4 of study 3 enabled a comparison between the use of a named and an anonymous class in nested restrictions. Each of questions 1 to 4 occurred in one variant with a named class and in the other variant with an anonymous class. This enables a controlled comparison between the use of a named class (Q1 variant 2, Q2 variant 1, Q3 variant 2, Q4 variant 1) and an anonymous class (Q1 variant 1, Q2 variant 2, Q3 variant 1, Q4 variant 2). Note that each participant answered two of the four questions with a named class and the other two with an anonymous class. The aggregated data for these two cases is shown in Table 8-22. A Fisher's Exact Test revealed no difference in accuracy ($p = 0.268$) whilst a t-test did show a significant difference in response time ($t(117.99) = 2.5387$, $p = 0.0124$).

In summary, there was partial support for hypothesis H7.12. There was no significant difference in accuracy between the use of a named and an anonymous class. However, response time was significantly longer in the case of the named class than in the case of the anonymous class.

Table 8-22 Named and anonymous classes – percentage correct and response times

named class; N = 15				**anonymous class**; N = 15			
% corr	mean time (standard deviation)			% corr	mean time (standard deviation)		
	overall	correct	incorrect		overall	correct	incorrect
52%	79 (58)	85 (70)	73 (43)	63%	59 (39)	58 (36)	61 (44)

8.6.3. *not … any* – hypothesis H8.8

For questions 5 to 8, there was no significant difference in accuracy between the two variants (Fisher's Exact Test, $p = 0.853$). Nor was there any significant difference in response time between the variants ($t(106.86) = 0.19408$, $p = 0.847$).

In summary, as with the results presented in subsection 8.5.2, there is no support for hypothesis H8.8. The use of *not … any* in place of *not … some* had no significant effect on accuracy or response time.

8.6.4. Discussion

Table 8-23 summarizes the hypotheses discussed in this section. In contrast to the previous section, the use of *noneOrOnly* and *including* made no difference to performance, thus offering no support for hypothesis H8.7. Possibly the use of nested restrictions creates extra cognitive load which makes it difficult to take account of the implication of these keywords, i.e. that there are two associated mental models. On the other hand, the results for hypothesis H8.8 are consistent with the previous section, i.e. *not ... any* makes no difference to performance. Finally, the analysis in this section found that the use of a named class took significantly longer than the use of an anonymous class.

Table 8-23 Nested restrictions – summary of hypotheses

	accuracy	**response time**
H7.12 – anonymous versus named classes	no significant difference	responses to questions using named class significantly longer
H8.7 - *noneOrOnly* and *including* in place of *only* and *some*	no significant difference	
H8.8 – *not ... any* in place of *not ... some*	no significant difference	

8.7. Participant feedback

As in the previous studies, at the end of this study participants were given the opportunity to respond to three questions:

- What did you find difficult about the quiz?
- What did you find easy about the quiz?
- Do you have any general comments about the quiz?

In addition, some participants made verbal comments both during and after completion of the questions. This section reports on the most commonly occurring and relevant of these written and verbal comments. Very many of the comments were about the general difficulty of the questions; others were commenting on the format of the study. When interpreting the more specific comments, it should be remembered that in usability studies it is not necessarily the case that the option which participants favour is the one with which they perform best. For example, Cockburn and McKenzie (2001), working with document management systems found that there was a significant preference for 3D systems over 2D systems, but no significant difference in task performance. An additional caveat is that the comments are not necessarily representative of the overall participant experience.

The most commonly reported problem was the difficulty of holding all the necessary information in memory, represented by one comment: "Most difficult is having to keep many different sets in mind at the same time...". As with the previous study, a number of participants expressed a wish to use pen and paper. Also as with the previous study, there were diverging views on the relative difficulty of the sections. For example, one participant regarded the section on functional properties as being the hardest, another participant thought it probably the easiest. Four participants found the section with intersection and union relatively easy; one found it difficult.

A number of other groups of questions were singled out as being difficult. Two participants explicitly identified the nested restrictions section as being difficult, whilst three participants identified nested expressions generally as difficult. Two participants identified unnamed (i.e. anonymous) classes with property restrictions as difficult. Two participants reported confusion over the use of two names for the same individual (i.e. in the object property section). In addition, one participant commented on the difficulty of double negation.

There were some comments on the effects of the MOS modifications. One participant commented favourably on the notation for the existential and universal restrictions (*noneOrOnly* and *including*), but explicitly excluded from this *not ... any*. In response to the question "what did you find easy about the quiz?", another participant noted the "noneOrOnly and including questions which did not involve individuals", presumably a reference to the negation and restriction section. Another participant specifically referred to *noneOrOnly* as being "very clear". However, one of the participants who found the nested restrictions section particularly difficult, commented on finding the *noneOrOnly* and *including* keywords confusing. One participant thought *not ... any* difficult despite being intuitive, because it differed from what the participant was used to. There was one specific reference to the use of *solely*: "the solely keyword was of help on getting the difference btw functional/inv functional".

There were three comments about *except*. One was unfavourable; the participant thought "not is more clear than except". Another merely identified "the disjoint union with except" questions as being difficult. The third identified the "except/not/including" questions as being easier than the questions on inverse functional properties.

As in the previous study, the topic of realistic versus abstract names was raised. One respondent made the comment: "It would be even harder in an abstract domain, e.g. the questions about children would be much harder without a concept of grandchild". Thus the use of realistic names enabled the participant to make use of a concept not explicit in the question, in a way which would be more difficult and less intuitive if abstract names were used.

8.8. Effect of participant prior knowledge and experience

Participants were asked the same questions about themselves as in the previous study: two questions about their knowledge of formal logic and DLs; a third about their usage of DLs; and a question about their relationship to English. Tables 8-24 to 8-27 show the four questions and the percentage of responses to each answer.

Table 8-24 Knowledge of Logic – percentage breakdown of participants; N = 30

Please rate your knowledge of formal logic on the following scale	
No knowledge at all	3%
A little knowledge, e.g. from an introductory course in formal logic	23%
Some knowledge, e.g. from a university course in formal logic	47%
Expert knowledge	27%

Table 8-25 Knowledge of DLs – percentage breakdown of participants; N = 30

Please rate your knowledge specifically of OWL or other Description Logic formalisms	
No knowledge at all	13%
A little knowledge, e.g. from an introductory course	47%
Some knowledge, e.g. working knowledge to create or edit ontologies	27%
Expert knowledge	13%

Table 8-26 Usage of DLs – percentage breakdown of participants; N = 30

Please indicate whether and for what purpose you use OWL or another Description Logic formalism	
Do not use	40%
I am learning about the language (either as part of a formal course or informally)	10%
I use the language in my work, e.g. to create ontologies	27%
I am researching the language and its applications	23%

Table 8-27 Relationship with English – percentage breakdown of participants; N = 30

Please indicate which of the following most accurately describes your relationship with English	
English is my main language	40%
English is one of my main languages	33%
English is not my main language	27%

Spearman's rank correlation coefficients were computed for each pairwise correlation between the four factors. As in the previous studies, the responses to the question about usage of DLs were assumed to be on an ordinal scale: 'Do not use': 'I am learning…'; 'I use…'; 'I am researching'. There was no significant correlation between the respondents' relationship with English and the other three factors. Significance here was calculated on a two-sided basis since there was no a priori expectation of a particular polarity of correlation. Table 8-28 shows the correlation coefficients and significance levels between the other three factors. These were calculated on a one-sided basis, since one would expect a positive correlation. There was a significant correlation between all three pairs of factors. In fact, 15 of the 30 participants rated their knowledge of logic and OWL / DL the same, whilst 14 rated their knowledge of logic greater than that of OWL / DL.

Table 8-28 Spearman's rank correlation coefficients and significance.

	knowledge of OWL or other DL formalisms	usage of OWL or other DL formalisms
knowledge of logic	$\rho = 0.56$; $p < 0.001$	$\rho = 0.56$; $p < 0.001$
knowledge of OWL or other DL formalisms		$\rho = 0.67$; $p < 0.001$

N.B. the p-values are not computed exactly because of the presence of ties.

There was no significant Spearman's rank correlation between participants' relationship with English and performance on the questions, neither in terms of accuracy nor response time (accuracy: $\rho = -0.09$, $p = 0.638$; time: $\rho = -0.04$, $p = 0.823$). These significance levels

were calculated on a two-sided basis, since, as discussed in section 7.8.3, there are opposing theories about how performance would be affected. Table 8-29 shows the Spearman's rank correlation between each of the three other factors and accuracy and time. The significance levels in this case were calculated on a one-sided basis, since the *a priori* expectation is that knowledge and usage would improve performance. The first point to note is that, for each of the three factors, the magnitudes of the two correlation coefficients are broadly equivalent, differing at most by 0.09. Knowledge of logic had the greatest effect on performance, followed by usage of DLs; for both these factors there was a significant effect on accuracy and response time. Knowledge of DLs, on the other hand, did not have a significant effect on either accuracy or response time.

Table 8-29 Impact of participant profile on accuracy and response time – Spearman's rank correlations.

	accuracy	**response time**
knowledge of logic	$\rho = 0.48$; $p = 0.004$	$\rho = -0.43$; $p = 0.009$
knowledge of OWL or other DL formalisms	$\rho = 0.17$; $p = 0.191$	$\rho = -0.24$; $p = 0.097$
usage of OWL or other DL formalisms	$\rho = 0.31$; $p = 0.047$	$\rho = -0.40$; $p = 0.015$

N.B. p-values not computed exactly because of ties

These results contrast, in places, with those of previous studies. For example, study 2 found that knowledge of DL had a significant effect on accuracy. These differences are likely to reflect the different distribution of participants' backgrounds in the studies. Significant correlations are more likely to be achieved where the distribution of participants by background category is more uniform.

8.9. Effect of question position

The analysis of the previous study in section 7.9 indicated that the response time, although not the accuracy, depends significantly on the position of the section and the position of the question in the section. The randomization of question order in this study enabled a controlled investigation of the effect of question order on performance. The first subsection below looks at the effect of overall position in the study, and also the relationship between accuracy and the time taken to respond. The second subsection separates the effect of the position of a question in its section and the effect of the position of the section in the study.

8.9.1. Position in study

Figure 8-5 shows the results of a logistic regression of accuracy against the position of a question in the study. The horizontal axis represents each of the 32 positions in the study. The circles represent the proportion of correct responses for each position; note that the 30 questions at each position are randomly selected from the 32 questions in the study. The curve represents the logistic regression, which gave results very close to a straight line over the relevant region. However, this regression was not significant ($p = 0.054$).

Regression analysis did show a significant reduction in time with position ($F(1, 958) = 62.85$, $p < 0.001$). Figure 8-6 shows the average response time for each position and the regression line. The regression analysis was performed using all the time data, i.e. not just the averages shown here. Note also that time was first transformed using the logarithmic transformation, as explained in section 8.2. For this reason, the regression line shown here is not strictly linear. Figure 8-6 shows an appreciable increase in response time at the beginning of each section; the four points at the very top of the figure represent the mean times for the first question in each section.

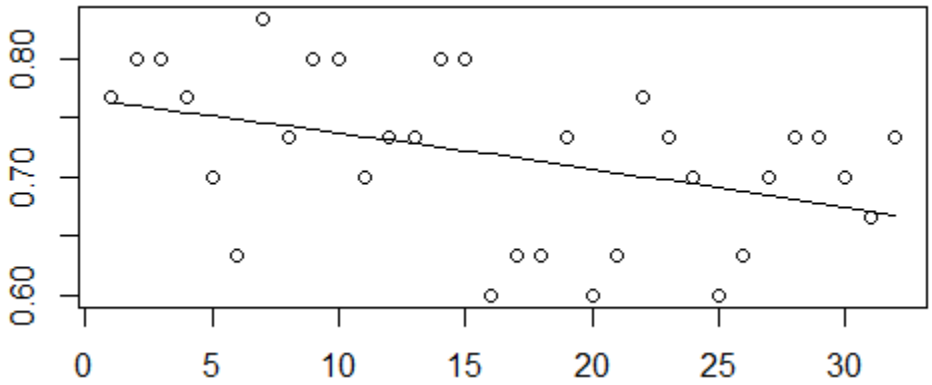

Figure 8-5 Proportion of questions correct versus position in study

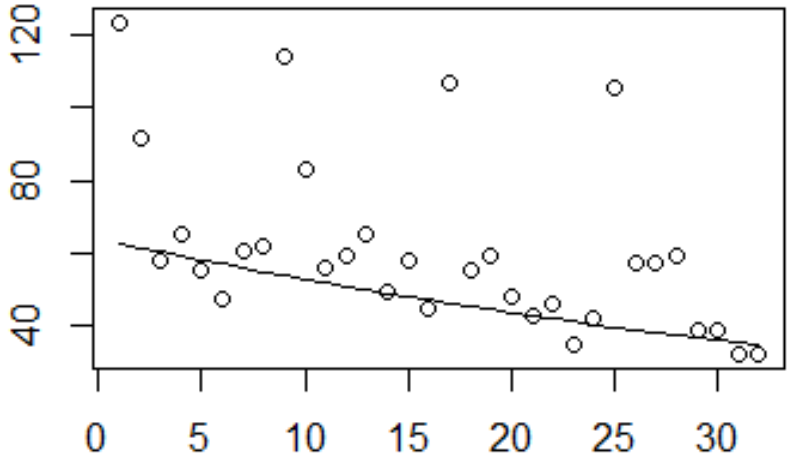

Figure 8-6 Response time (secs) versus position in study

It was thought that the time spent on a question might be influencing the accuracy with which it is answered. To investigate this, a logistic regression of proportion of correct responses against time was conducted. This result was not significant ($p = 0.078$). However, the mean time for the correct responses is 59 seconds, and for the incorrect is 66 seconds, and this difference is significant ($t(525.6) = 3.2288$, $p = 0.001$). This may be a consequence of the harder questions taking longer and being more prone to incorrect answers.

To separate out the effects on accuracy of position and response time, a logistic regression was performed of accuracy versus position and response time. This gave no significant top-level effects (position: $p = 0.996$; time: $p = 0.626$). There was, however, a significant interaction between the two factors ($p = 0.024$). Letting *c* represent the proportion of correct responses, the resultant model was[147]:

$$logit\ (c) = 1.166 + 6.145e\text{-}05 * position + 1.115e\text{-}03 * time - 3.291e\text{-}04 * position * time$$

The first thing to note is that the coefficients of position and time are positive, whereas in the single factor analysis these coefficients are both negative. This is explained by the negative coefficient for the interaction effect.

[147] logit represents the logistic function, so $logit\ (c) = ln\ (c/(1-c))$. In the equations represented here, *logit (c)* is used on the left-hand side, rather than solving the equation to provide a value for *c*. Since our interest here is in the direction of each effect, and since logit is a monotonically increasing function, the mathematics can be simplified by avoiding solving for *c*.

In order to study how one factor influences the effect of the other, we can rewrite this equation in two ways. Firstly, to understand how response time influences the effect of position, the equation can be rearranged to:

*logit (c) = 1.166 + 1.115e-03 * time + position * (6.145e-05 - 3.291e-04 * time)*

Since the minimum response time recorded is 4.5 seconds, the expression in brackets will always be negative. Thus, accuracy decreases with position for every value of response time, and moreover decreases more rapidly the greater the response time. The relationship between accuracy and position may well be a fatigue effect; the fact that this effect is greater when the response time is greater may be because the fatigue effect is greater for questions which take longer, i.e. the more difficult questions.

In order to study how position influences the effect of response time, the equation can be rearranged to:

*logit (c) = 1.166 + 6.145e-05 * position + time * (1.115e-03 - 3.291e-04 * position)*

Here, the expression in brackets will be positive for low values of position and then negative for higher values. Equating the expression in brackets to zero gives position $\approx$ 3.4; i.e. accuracy improves with time for the first three questions and then decreases with time for the fourth and later questions. There seem to be two opposing effects here. Initially, the longer a participant spends on a question, the greater the likelihood of getting it correct. Later, the time spent on a question is perhaps more an indicator of its difficulty, and hence of the chance of getting it wrong.

8.9.2. Question position in section and section position

When we perform a logistic regression based on the position of a question in its section, rather than overall, the effect of position on accuracy is not significant ($p = 0.827$). However, a logistic regression against the position of the section shows a significant negative effect ($p = 0.041$). This suggests that there may be an overall effect of fatigue which is reducing accuracy.

To understand this better, a logistic regression was undertaken of proportion of correct responses against both the position of the section in the study and the position of the question in the section. In this case, there is a significant reduction in accuracy with position of the section ($p = 0.022$), but no significant variation with position of question in section ($p = 0.184$) and no interaction effect ($p = 0.123$). It is possible that there is some kind of fatigue effect, which is too small to be significant within a section but becomes significant on a section-by-section basis. It could also be the case that within individual sections, a learning effect compensates for the fatigue effect.

The relationship between response time and question position is somewhat different. A regression analysis shows that the decline in response time with position in section is significant ($F(1, 958) = 130.3$, $p < 0.001$) as is the decline with overall position (as noted in section 8.9.1: $F(1, 958) = 62.85$, $p < 0.001$). A regression analysis of time against section order also shows a significant decline in time ($F(1, 958) = 27.65$, $p < 0.001$)[148].

[148] A comparison of the coefficients of $\log_{10}$ time against the regression variate is also interesting; for position in section: -0.046752, for overall position: -0.008320, for position of section: -0.04638. The appreciably greater magnitude of the first of these, compared with the second, is likely to be a consequence of the effect on response time of the first one or two positions. As is suggested by Figure 8-6, on a section-by-section basis this would create a greater decline in response time than the overall regression curve shown in the figure.

A regression analysis of time against position of section and position in section also gives a significant result ($F(3, 956) = 57.12$, $p < 0.001$). However, this analysis shows no significance dependence on position of section ($p = 0.546$) but a significant negative dependence on position within section ($p = 0.006$) and also a significant negative interaction effect ($p = 0.029$). The within-section effect indicates that participants accelerate as they proceed through a section; this could be a combination of a learning effect plus an effect of fatigue. The lack of a significant intersection effect suggests there is no significant inter-section learning. The significant interaction effect indicates that rate of acceleration increases for the later sections.

8.9.3. Discussion

In summary, there appears to be a weak effect possibly due to fatigue, by which accuracy declines during the study. The effect is so weak as not to be significant within individual sections, but is significant on a section-by-section basis. This has implications for the design of such experiments. It suggests that, whilst prudent, it is not absolutely necessary to randomize the position of questions within a section if only accuracy data is being used and only comparisons within sections are being made. If accuracy comparisons are being made between sections, rather than just between questions in the same section, then it is necessary to randomize the order of the sections, or to use a design which presents different orderings to different participants. These comments are sensitive to the overall structure of the study. If a study consisted of 32 questions in one section and it was desired to compare these questions, then randomization could be needed.

For response time, there is a significant variation within a section, so questions do need to be randomized, or some appropriate design strategy is required. On a one-variate approach, there is also a significant variation between sections. On a two-variate approach this is not the case, but there is a strong interaction effect between section position and question position within the section. On either analysis, section position needs to be randomized, as in this study, or some appropriate design needs to be adopted, as in the two previous studies.

Finally, when considering accuracy, there is also a significant interaction effect between position of a question in the study and the time taken to respond. On average, for the first three questions in the study, accuracy improves with time to respond; possibly because greater response time means more thought is being given. Thereafter, accuracy declines with response time; possibly because greater response time is an indicator of greater difficulty. From an alternative viewpoint, accuracy decreases with position more rapidly the greater the response time; possibly because the harder questions intensify fatigue.

8.10. Varying effect of valid and non-valid putative conclusions

Analysis of the previous study found no evidence that the questions with valid putative conclusions were answered significantly more or less accurately than the questions with non-valid putative conclusions. There was a significant difference in response time between correctly answered valid questions and correctly answered non-valid questions, with the latter taking longer. However, the difference was not significant when the analysis was restricted to a subset of the responses, so as to achieve a balanced comparison. This section re-examines these questions using data from study 3.

8.10.1. Validity and accuracy

There were 16 questions with valid conclusions and 16 questions with non-valid questions. The valid questions were answered rather more accurately than the non-valid (74% compared with 69%). This contrasts with study 2, where the non-valid questions were

answered more accurately. However, as in study 2, the difference was not significant ($p = 0.133$, Fisher's Exact Test, two-sided).

As with study 2, a more controlled comparison can be made by limiting the questions to those pairs of valid and non-valid questions sharing the same axioms. This is at the expense of eliminating a number of questions, including all the questions from the two sections with universal and existential restrictions. There are five pairs of valid and non-valid questions with identical axioms: three pairs relating to object properties, i.e. questions 1 to 6 in Table 8-1; and two pairs of questions relating to Boolean concept constructors, i.e. questions 5, 6 and 7, 8 in Table 8-9 and Table 8-10. For these questions, the proportion of correct responses to the non-valid questions (79%) is greater than to the valid questions (64%), and this difference is significant ($p = 0.007$, Fisher's Exact Test, two-sided). This compares with study 2 where the difference had the same polarity but was not significant.

8.10.2. Validity and response time

The five pairs of valid and non-valid questions investigated in the previous subsection also provide an opportunity to further investigate the effect of question validity on response time. The ten questions provide a total of 300 responses, 150 from the valid questions and 150 from the non-valid questions. Of these 150 pairs, with both members of each pair from the same participant, there are 77 pairs in which both responses are correct. This enables a balanced comparison between valid and non-valid questions. Of these 77 pairs, the mean response time for the valid questions is 64 seconds and for the non-valid is 70 seconds. This difference is not significant ($t(76) = 1.5694$, $p = 0.121$, paired two-sided test).

It might be argued that what determines the participant's experience is not whether the question is valid or non-valid, but rather whether the participant's response is valid or non-valid. This approach enables the analysis to be repeated with a larger sample. In addition to the 77 pairs where a participant has answered both questions correctly, there are also 13 pairs where a participant answered both questions incorrectly. This makes a total of 90 pairs of responses to questions with identical axioms, with one response being 'valid' and one response being 'not-valid' from each pair. Of these 90 pairs, the mean response time for the valid responses is 65 seconds and for the non-valid responses is 71 seconds. Again, this difference is not significant ($t(89) = 1.6387$, $p = 0.105$, paired two-sided test).

8.11. Discussion

This section makes some final comments on the study. Subsection 8.11.1 draws together the conclusions, whilst subsection 8.11.2 reviews the threats to validity

8.11.1. Conclusions

Study 3 investigated a number of additions to MOS, intended to mitigate the problems identified in Chapters 6 and 7, and designed in part with insights from studies in psychology and language. A number of these additions met with some success:

1. The use of *solely* significantly reduced response times for the valid questions. For these questions there was a reduction of around 30%, from 81 seconds to 57 seconds.
2. The replacement of *and* with *intersection* significantly improved accuracy, with an increase from 25% to 80%.
3. The replacement of *only* with *noneOrOnly* and *some* with *including* significantly improved accuracy and reduced response time in situations with only negation and restriction. When questions 3 and 7 were excluded, to avoid the confounding effect of the introduction of *any*, the increase in accuracy was from 61% to 75%, whilst the

reduction in time was around 22%, from 45 seconds to 35 seconds. However, the change was not effective in situations of nested restrictions.

A decision to implement any of these changes depends on a cost-benefit analysis. However, there is a *prima facie* argument for introducing changes (1) and (2). In the case of (1), a reduction in time of 30% could be regarded as worthwhile, particularly as it was associated with an appreciable but non-significant increase in accuracy from 60% to 72% for the valid questions. In the case of (2), this is a very appreciable increase in accuracy indeed. The case of (3) requires more research, as is discussed below.

The effect of replacing *and* with *intersection* illustrates the need to avoid ambiguous language. This topic is returned to in Chapter 9.

The effect of replacing *and not* with *except* may have the effect of reducing response time. However, the results are not conclusive and further investigation is required. Such an investigation should also look at the extent to which the use of *except* aids modelling. This suggests a more general research question: to what extent can comprehension be aided by using language which more naturally fits our thought patterns, rather than fitting those thought patterns into the limited vocabulary of a language composed of the operators of formal logic.

The use of *noneOrOnly* and *including* appeared successful in situations where the participant needed only to manipulate one restriction at a time, but not so when nested restrictions were involved. The complexity created by the use of nested restrictions remains a significant challenge. It may be that better results could be achieved with alternative keywords. Krötzsch et al. (2012b), discussing the universal restriction, equate the expression $\forall parentOf.Female$ to the English phrases "no children other than female ones" and "no children that are not female". This suggests *noneOtherThan* or *noneNot* as alternatives to *only*. Discussing the existential restriction, they equate $\exists parentOf.Female$ to "individuals that are parents of at least one female individual". This suggests *atLeastOne* as an alternative to *some*.

A number of strategies were not successful. The use of *solely* with the inverse functional property did not improve performance; an alternative keyword or word order might help here. Moreover, the use of *any* for the existential restriction when preceded by a negation made no difference to performance. Similarly, the use of prefix notation for conjunction and disjunction made no difference to performance.

Although not directly related to its objectives, the study also suggested that nested restrictions with two statements linked by a named class took longer to comprehend than with one statement and an anonymous class. This may relate to the overhead of storing and retrieving the extra class name.

These findings suggest further areas for research. In particular, how to improve performance with inverse functional properties and with nested restrictions.

As a final point, and returning to the discussion of subsection 2.2.2, note that the replacement of *and* with *intersection*, and *and not* with *except*, is purely concerned with syntax, i.e. with changes to MOS to avoid ambiguity. The other changes discussed are concerned with using syntax to mitigate problems inherent in DLs.

8.11.2. Threats to validity

Threats to construct validity

As in the previous studies, the responses were recorded automatically; this time using the *MediaLab* application. Moreover, *MediaLab* also automatically recorded the timepoints. This is done with a precision of 1 millisecond. The supplier, *Empirisoft*, do caution against the use of *MediaLab* for certain "cognitive / perception trials" for which millisecond accuracy is required. Nevertheless, the accuracy of *MediaLab* is far greater than that achieved by the use of *Camtasia* in the previous two studies.

Threats to internal validity

The previous two studies indicated that question position within section was having a material effect on the response time measurements. Study 3 has overcome this effect by randomizing both the order of sections and of questions within sections. Whilst this was really motivated by considerations of response time, the analysis in subsection 8.9.2 also indicates that question position does have an effect on accuracy which is not significant within sections, but does make itself felt over the duration of a whole study. Thus, randomization of section and question position is also prudent for the collection of accuracy data.

Threats to external validity

The questions used in this study were mostly isomorphic to those used in study 2. Like the study 2 questions they were representative of commonly used DL constructs but were removed from the commonly used contexts employed in study 1. However, study 2 had used these questions to identify generic difficulties, and study 3 similarly addressed solutions to those generic difficulties.

Table 8-24 shows that, as in the previous two studies the majority of participants had at least some knowledge of logic, i.e. greater than 'a little knowledge'. However, unlike in the previous studies, only a minority of the participants had at least some knowledge of DLs. In that sense, the participants were more representative of less specialist users.

Finally, all three studies were concerned with the understanding of DL inferences. In particular, study 3 was concerned with syntactic changes to improve that understanding. The studies did not address the complementary issue of modelling, and whether syntactic changes might improve ease of modelling. When modelling, one needs to understand the consequences of the axioms being written. Thus, the ability to reason with DL is necessary for modelling with DL. However, there may be other factors which determine success in modelling. These would need to be investigated in experiments which involved participants in modelling tasks.

9. Discussion – theory and methodology

There is nothing inherent in mathematical logic that makes it the study of reason.

George Lakoff, 'Women, Fire and Dangerous Things', (1987)
Chapter 14, page 223

Certain aspects of classical logic do seem to be used in reason.

ibid., Afterword, page 586

This chapter reviews the chief themes which have run through the dissertation. Section 9.1 begins by discussing the use of theory, from cognitive psychology and the philosophy of language. Section 9.2 then looks at ambiguity in natural language, taking as examples the confusion surrounding the words *and* and *or*. Confusion in interpreting *and* contributes to some of the difficulties with MOS, and is in part related to the difference between logic and DL. Section 9.3 then discusses how participants approach answering the questions, and compares reasoning in the laboratory with reasoning about real applications. Section 9.4 discusses the relationship between the formal concepts of decidability and computational tractability, and the difficulties experienced by human reasoners.

The remainder of the chapter is concerned with methodology. Section 9.5 makes some observations about the relationship between sample size, effect size and significance level. Section 9.6 then looks at the methodological lessons learned from the three studies described in the preceding chapters. Finally, Section 9.7 builds on this to make recommendations about methodologies for future studies.

9.1. The uses of theory

Chapter 3 introduced three theories of reasoning: relating to the use of rules, mental models, and the concept of relational complexity. In the psychological literature, all three theories are used to understand the difficulties people experience with reasoning. Mental model theory has also been used to predict the kinds of mistakes which people make in reasoning. In the context of this work, these theories have been used to explain performance, as discussed in subsection 9.1.1, and as a yardstick for measuring difficulty, as discussed in subsection 9.1.2. The theory of the *implicature*, developed by philosophers of language, in particular Grice (1975), offers an additional insight into the difficulties experienced with reasoning using natural language. These theories also offer the potential to guide language design, which is discussed in subsection 9.1.3. Recently, alternatives have been proposed to the theories of reasoning described in this dissertation; some of these alternatives are briefly discussed in subsection 9.1.4.

9.1.1. Explaining performance

In Chapters 6 and 7, mental model theory was used to explain the difficulties participants experienced with negated conjunction and, in certain cases, with restrictions. No explicit use was made of rule-based theory, in the sense of using the rules of logic such as *modus ponens*, as in Rips (1983). The primary reason for describing rules-based theory was to provide a background for mental model theory, which came after it and which can be seen as a reaction to it. However, it did appear that in some situations participants were reasoning syntactically rather than building models. Although not necessarily using the same set of rules as proposed by Rips (1983), they were using syntactic reasoning to avoid developing detailed models. For example, they appeared to be recognising complementary classes on

the basis of syntax. At times this helped reasoning. At other times confusion between *not… only* and *only not* and between *not … some* and *some not* may have led to errors, as discussed at the end of subsection 7.5.6. The use of syntactic and model-based reasoning may in part be a response to the particular nature of the question. It may also be influenced by the preferred mode of thinking of the participant. As already noted, Ford (1995) has observed that some people exhibit a preference either for syntactic or spatial reasoning. The latter can be seen as akin to the construction of mental models[149].

Finally, the implicature offers another insight into the difficulties associated with using natural language. In the context of MOS, the most obvious example is the use of *only*, with the implicature that, e.g. *has_child only MALE*, means that a male child does exist. Difficulties may also occur with *some*, e.g. *has_child some MALE* may have an implicature that there are no female children.

9.1.2. Measuring difficulty

Both the rules-based and model-based theories provide a measure of the difficulty of inferences. Rips (1983) calculated probabilities representing the accessibility of particular rules, and from this the probability of successfully arriving at conclusions to specific inferences. The mental model theory generally represents difficulty in terms of the number of mental models required for an inference. Additionally, it was suggested in Chapter 7 that the complexity of the individual models was another differentiator. Relational Complexity theory (Halford et al., 1998) also provides a measure of complexity. Whilst in principle applicable to any type of reasoning, in the context of DLs, RC theory is most readily applicable to reasoning about object properties, since these describe relations. Work reported in Chapter 7 demonstrated that RC is a useful measure of difficulty. RC theory also enabled a controlled comparison between different types of reasoning, i.e. based on transitive and functional properties, which illustrated that the characteristics of the relation is also a factor influencing difficulty[150].

9.1.3. Language design

Some of the new syntactic constructs described in Chapter 8 were influenced by the theories of reasoning and language. *noneOrOnly* and *including* were intended to guide the participant in formulating both the mental models associated with the universal and existential restrictions, and to avoid the implicatures associated with *only* and *some*. The use of *solely* with the functional object property was not directly inspired by any theory. However, RC theory was used to identify that functional properties posed a particular difficulty. That this difficulty was, at least in part, to do with the direction of uniqueness, was suggested by the comment of at least one participant. On the other hand, the use of *not … any* and *except* was not influenced by theory, but by a desire to match the syntax of DL to natural language[151].

---

[149] Ford (1995) claims that spatial reasoning is different from the construction of mental models, in that in the former "the class itself and not the finite members of the class is represented". Khemlani and Johnson-Laird (2012) dispute this, based on the experiment protocols quoted by Ford (1995). Any difference may have to do with the availability of pen and paper in the latter's experiment. In any case, spatial and mental model reasoning both create models and are clearly differentiated from syntactic reasoning.

[150] In both the transitive and functional properties compared in study 2, the ordering of the entities in the relation is significant. In symmetric relations, the order is irrelevant and this would presumably facilitate reasoning. RC theory does not appear to take this into account. Relations in which the ordering is irrelevant are sometimes termed 'relationships' (Codd, 1970).

[151] An alternative inspiration for language design might come from logical patterns or macros, e.g. exclusive or, as discussed in Horridge et al. (2006).

It is possible to use intuition to guide the development of a formal language and then use experiment to determine the value of that intuition, as was done for *not … any* and *except*. However, such experiments are time and labour-intensive. Theory, when tested and accepted can be used to guide language design, at least reducing the need for experiment. Indeed, the theories outlined in this dissertation could be used generally to guide the design of any formal language.

9.1.4. Alternative theories of reasoning

The theories of reasoning used here were chosen in part because of their fundamental position in reasoning research. They were the first to offer insight into reasoning and are still of interest in the psychological research community. Later, probabilistic models were developed. For an early discussion see Oaksford and Chater (2001), who apply the probabilistic approach to conditional inference, Wason's selection task and syllogistic reasoning. The use of probabilistic reasoning makes it theoretically possible to unify deductive with inductive reasoning, based on plausibility. Deductive reasoning can be seen as reasoning about statements with an associated probability equal to one. Inductive reasoning can be seen as reasoning about statements with associated probability potentially less than 1. In fact, prior to Oaksford and Chater's work (2001), Johnson-Laird (1994) had put forward the view that mental models can explain both types of reasoning. In this view, the proportion of mental models for which a putative conclusion is consistent is a measure of the likelihood of its correctness in the real world, and is the basis of inductive reasoning; whilst consistency across all mental models, indicating a probability of one, is the basis of deductive reasoning.

Rips (2001a) compares deductive and inductive reasoning. He shows that participants will state that a deductively incorrect argument is correct if it is plausible. The converse, stating that a deductively correct argument is incorrect if it is implausible, occurred much less frequently. However, Rips (2001b) argues against the idea that one theory can apply to both inductive and deductive reasoning. In particular, he rejects the view that mental models can explain both types of reasoning (Johnson-Laird, 1994). More recently, Lassiter and Goodman (2015) claim to have refuted objections to the probabilistic approach, and in particular the view that two forms of reasoning are required, one for induction and one for deduction.

Subsection 9.1.1 noted that, in some cases participants seemed to use syntactic reasoning, akin to the rule-based approach of Rips (1983), whilst in other cases they seemed to develop models. Perhaps, in some cases, the syntactic approach offers a simple, quick route to a solution. In other cases, this approach is difficult and participants fall back on building models. Supporting this standpoint, and arguing against the view that all human reasoning is inductive, Rips (2001b) points out that people can appreciate some reasoning arguments where they must rely on the pattern of the argument. Lakoff (1987, Chapter 17) offers an explanation for the schemas used in reasoning. He argues that such schemas arise from bodily experience. One of the most basic of these is what he calls the container schema, i.e. something is contained inside of another object, whilst a third object may be outside. He suggests that this is the basis of modus ponens and of the logic of Boolean classes. It would also appear to be the basis of the most simple of syllogisms: all A are B; all B are C; therefore all A are C. Perhaps this is why modus ponens and this particular syllogism are easy – they rest on a very simple and obvious schema.

An altogether different direction in reasoning research is exemplified by Ragni et al. (2016), who were concerned with formally modelling non-monotonicity. This is observed in human reasoning when additional information suppresses inferences which would otherwise have been made. An example they quote comes from Byrne (1989):

1. If she has an essay to write, then she will study late in the library.
2. If the library stays open, she will study late in the library.
3. She has an essay to write.

Here, statements (1) and (3) should lead to the conclusion that "she will study late in the library". However, the presence of statement (2) resulted in 38% of study participants making this conclusion and 62% concluding that "she may or may not study late in the library". Ragni et al. (2016) observe that "some non-monotonic logics seem to be adequate to describe human commonsense reasoning". The work represents an interesting attempt to relate observed facts of human reasoning to non-classical formal logic.

All these various theories of reasoning may offer insight to guide the design of languages for human-computer interaction, particularly those languages, like DLs, which are based on logic. From a practical perspective, it is not important that no universal model of reasoning has won acceptance. Each of the alternative theories may offer some insight in particular circumstances.

9.2. Ambiguity in natural language – the examples of *and* and *or*

Natural language offers the possibility of providing a knowledge representation format which conforms with human thought patterns. However, this can be at the price of ambiguity. This is exemplified by the word *and*. Evidence of the ambiguity which surrounds the word is provided by Mendonça et al. (1998), whose analysis of medical terminology, already cited in Chapter 4, found that 51% of the occurrences represented conjunctions, 46% inclusive disjunctions, and 3% exclusive disjunctions. Partee and Rooth (1983) consider this ambiguity from the standpoint of formal semantics. The details of their analysis are beyond the scope of this dissertation. However, some of their examples are instructive. In the next subsection, 9.2.1, they are accompanied by this author's own analysis, aimed at illustrating the difficulties which arise with the use of the word in MOS. *or* also presents difficulties, and this is discussed in subsection 9.2.2. Subsection 9.2.3 then discusses the difference between the use of *and* and *or* in logic and MOS, and makes a general observation about the difference between logic and DL. Subsection 9.2.4 makes a final comment.

9.2.1. *and* as conjunction and union

Amongst other examples, Partee and Rooth (1983) give:

1. Susan will retire and buy a farm.
2. John and Mary are in Chicago.
3. She was wearing a new and expensive dress.

(1) is clearly a conjunction. It is shorthand for the conjunction of two propositions:

1'. (Susan will retire) and (Susan will buy a farm).

(2) can also be interpreted as a conjunction:

2'. (John is in Chicago) and (Mary is in Chicago).

However, *and* here can also be regarded as representing union:

2". {John} ∪ {Mary} are in Chicago.

(3) can again be regarded as a conjunction:

3'. (She was wearing a new dress) and (She was wearing an expensive dress).

However, here *and* can be regarded as representing intersection:

3". She was wearing a {new dress} ∩ {expensive dress}.

In each of these three examples, the underlying semantics of *and* is the conjunction of two propositions. However, in (2) and (3), an interpretation in terms of class operations may be more natural. Moreover, it could be argued that union, rather than intersection, is a more natural interpretation of *and*, since intersection is often achieved simply by juxtaposing adjectives, e.g.:

3'". She was wearing a new, expensive dress.

Thus, the use of *and* in MOS, to represent intersection, can be ambiguous for those who rely on intuition from natural langauge.

9.2.2. *or* as union and uncertainty

or also presents difficulties. Menonça et al.'s (1998) analysis identified the ambiguity between inclusive and exclusive disjunction. However, a more fundamental confusion is between *or* representing union and *or* representing uncertainty.

Example (2) above used *and* to create a union. The use of *or* to represent union, as in MOS, is clearly wrong in this context:

2A. John or Mary is[152] in Chicago.

Here, *or* represents uncertainly, since 2A is a contraction of the disjunction[153]:

2A'. (John is in Chicago) or (Mary is in Chicago).

However, Partee and Rooth (1983) give another example:

4. The department is looking for a phonologist or a phonetician.

The underlying semantics of *or* is again disjunction:

4' (The department is looking for a phonologist) or (the department is looking for a phonetician)

However, this statement is ambiguous. It could imply uncertainty on the part of the originator of the sentence. It could also imply that the department is looking for someone from either discipline, i.e. from the class which is the union of phonologists and phoneticians:

4" The department is looking for {phonologist} ∪ {phonetician}.

[152] *are* has been replaced here by *is*, in accordance with normal usage. Retaining the plural form would not create the desired semantics of union, merely render the sentence ungrammatical.

[153] Whether we regard *or* here as inclusive or exclusive disjunction, the effect is still to represent uncertainty. For inclusive disjunction this is between two possibilities (John is in Chicago or Mary is in Chicago). For exclusive disjunction it is between three possibilities (John is in Chicago, or Mary is in Chicago, or they are both in Chicago).

In this example, *or* has the same meaning as in MOS.

9.2.3. Logic and DL

The preceding two sections have demonstrated that whilst *and* and *or* in natural language can have the meanings ascribed to them in MOS, they can also have alternative meanings: union in the case of *and*; uncertainty in the case of *or*. Language, when used to express facts rather than, e.g. emotions, is comprised of propositions. In this respect, natural language is similar to formal logic. When we write natural language as a set of propositions, then *and* and *or* have the meaning of conjunction and disjunction respectively. The ambiguity with *and* arises because the contracted forms (2) and (3) above lead to different interpretations in terms of sets. The ambiguity with *or* arises because both (4) and its expanded form (4') have two potential interpretations: that either discipline will suffice or that there is uncertainty as to which.

Baader and Nutt (2003) describe DLs as "a family of knowledge representation formalisms … equipped with a formal, logic-based semantics". This quote identifies that DLs are logic-based; they are not, strictly speaking, logics. Logics are concerned with propositions. For example, Boolean operators can be used to combine propositions, and quantifiers are used to define propositions. DLs are a calculus of classes. Boolean operators, or Boolean concept constructors, are used to combine classes, and restrictions are used to define classes[154].

In the context of logic, *and* and *or* are unambiguous. In the context of classes, they are ambiguous and do not necessarily equate to their use in MOS. Indeed, it could be argued that *and*, when used between individuals and classes normally represents union, whilst *or* normally conveys uncertainty.

The adoption of the keywords *union* and *intersection* in DL avoids this ambiguity. A point to note is that *union* is used in everyday English, e.g. 'the United Kingdom is the union of England, Scotland, Wales and Northern Ireland'. *Intersection*, in everyday usage, is confined to lines, and by extension streets and roads. It is not used more generally. As noted above, in everyday language the intersection of two sets is often represented by juxtaposing two adjectives ('new, expensive dress').

The lack of a commonplace word for set intersection in English was presumably the motivation for the search for an alternative in MOS. The choice of *and* presumably arises because, if we take P(x) and Q(x) to be predicates representing membership of classes C_P and C_Q respectively, then (P and Q)(x) represents the intersection of these two classes, $C_P \cap C_Q$. Similarly, the choice of *or* arises because (P or Q)(x) represents the union of the two classes, $C_P \cup C_Q$. This is also presumably the motivation for using conjunction, disjunction, and (by a similar argument) negation as synonyms for intersection, union and complement, as in Krötsch et al. (2012), and as used in this dissertation. However, this is a logician's usage, not a commonplace one.

9.2.4. Comment

The preceding analysis has demonstrated that *and* and *or* are ambiguous in everyday English; and that they are very commonly used with meanings quite different from those in MOS.

[154] The distinction between Propositional Logic and Boolean concept constructors in DL is analogous to that between two common interpretations of Boolean algebra. In one, Boolean algebra is interpreted as an algebra of propositions; in the other it is interpreted as an algebra of classes. The two cases are frequently distinguished by the use of the operator symbols $\wedge$ and $\vee$ for the algebra of propositions and $\cap$ and $\cup$ for the algebra of classes.

The aim was not to denigrate the use of natural language. Indeed, in Chapter 7 use was made of *except* (with some evidence of success) and *not … any* (with no evidence of success) in an attempt to improve reasoning by coming closer to natural language usage. However, when natural language is adopted, great care needs to be taken to avoid any ambiguous usage.

9.3. Reasoning in the wild and in the laboratory

The axioms and conclusions used in the studies were chosen to represent the kind of reasoning which would take place in real life situations. Nevertheless, an experimental study is necessarily an artificial situation. However close the questions are to real life reasoning, there remains scope for debate as to how close is the experience. Two aspects of the experimental experience are considered here: how participants answer the questions, in particular whether they attempt to confirm or refute validity; and the use of meaningful or non-meaningful names.

9.3.1. Confirming or refuting validity

An important motivation for this work, as for that of Horridge et al. (2011) and Nguyen et al. (2012, 2013), is to improve the understanding of how an entailment follows from a particular justification. In a real situation, people will start with the assumption of validity. It is not *a priori* clear whether this would generally also be the case in the context of an experimental study, where participants are expecting some questions with non-valid conclusions. The quantitative data from the three studies could offer some indication as to how these questions are answered. In particular, if participants start by seeking to prove validity, a non-valid response would arise after a failure to prove validity. As a consequence, we would expect the valid responses to be given more quickly than non-valid responses. Note that valid responses are a combination of the correct responses to valid questions and incorrect responses to non-valid questions, and *vice versa* for non-valid responses.

Table 9-1 Mean and standard deviation of valid and non-valid response times

study		**Mean (s.d.) time - secs**		**t-test**
		Valid responses	**Non-valid responses**	
all data	1	49 (35)	59 (40)	t(206.73) = 2.8986, p = 0.004
	2	60 (44)	71 (52)	t(813.98) = 3.4977, p < 0.001
	3	56 (50)	66 (61)	t(952.46) = 3.987, p < 0.001
balanced data	2	69 (50)	75 (57)	t(115) = 1.1616, p = 0.248
	3	65 (54)	71 (44)	t(89) = 1.6387, p = 0.105

Table 9.1 shows that in all three studies, when all the data was used, the valid responses took at least ten seconds less than the non-valid responses, and in each case this difference was significant. However, this analysis is open to the objection that the effect of the nature of the responses may be confounded with other factors. The table also shows, for studies 2 and 3, a balanced comparison. Here, only the responses are used from those questions which occur in pairs, each pair with the same axioms and a valid and non-valid putative conclusion. Moreover, only pairs from each participant were used where the participant had answered both questions correctly or both questions incorrectly. This has the effect of providing a matching set of valid and non-valid responses to the same questions, and enables a paired t-test to be conducted, where the pairing is both on the participants and the question-pairs. Again, for both studies 2 and 3, the valid responses were received more quickly than the

non-valid. However, with the balanced data the difference in mean response times reduced to 6.2 seconds for study 2 and 5.3 seconds for study 3, and in each case the difference was not significant.

Thus, on the basis of the balanced data, there is no evidence of a significant difference between the valid and non-valid responses. Hence there is no evidence that participants start by trying to prove validity, but equally, there is no evidence that participants try to disprove validity. The reason for the reduction in effect size when only the balanced data is used requires further investigation[155].

9.3.2. Meaningful and non-meaningful names

Studies such as those described in Chapters 6, 7 and 8 ought to correspond as closely as possible to the experience of a real application. On the other hand, they need to ensure that no prior domain knowledge will bias the results. Horridge et al. (2011) satisfy the latter requirement by using abstract names, e.g. C1, C2, prop1, prop2. However, this deviates from the former requirement, in that in general ontologists use meaningful names. The use of abstract names may impose difficulties not present in real applications. In particular, it may make recalling entities more difficult and thereby impose an irrelevant burden on the participant. As an alternative, Nguyen et al. (2012) use a combination of meaningless names ('kalamanthis', 'tendriculos'), meaningful names ('plant', 'animal'), and also names which are not real English words but have a semblance of being real words ('merfolk', 'lizardfolk').

The studies described here adopted a mixed approach. All three studies used a combination of meaningful and abstract names. In particular, object properties were represented by meaningful names, as were some classes. In each study the handout noted that the names used should not affect any conclusions to be drawn. However, it is doubtful whether participants really achieve this detachment. The difficulty, noted in subsection 7.7.3, that one participant had in regarding *has_nearest_neighbour* as functional is illustrative of this. The participant, mentioned section 8.7, who used the concept of 'grandchild', not mentioned in the question, offers another illustration.

It is also doubtful whether this degree of detachment is ideal, since it is removed from actual experience. A particular example of this is associated with the characteristics of object properties. It is likely that the user of an ontology has a reasonable idea of whether a particular object property is, e.g. transitive or functional. The *Characteristics* statement may serve to confirm any expectations, but it is primarily for the benefit of the computer, not the human. As a result, the human reasoner may often be using an inherent understanding of the characteristics of object properties, rather than falling back on their defined characteristics. Johnson-Laird et al. (1989, page 668) make the same point when they argue that the logical properties of a term, e.g. the transitivity and symmetry of "is in the same place as" emerges "without any need to use explicit statements of these properties in the form of meaning postulates". This suggests that the best environment for investigating difficulties and testing out new syntax is the real application which the ontologist uses. Such an approach will be limited in sample size. An alternative is to use a real application, but

[155] It might be thought that the differing results when using all the responses and when using the balanced responses was due to the absence of restrictions in the balanced questions. However, this is not the case. In studies 2 and 3, when all the responses from the object property and Boolean operator sections were used, the difference between the mean times for the valid and non-valid responses was at least 10 seconds, and this difference was significant:

study 2 – valid: 67 seconds; non-valid: 77 seconds; $t(425.85) = 2.7442$, $p = 0.006$

study 3 – valid: 59 seconds; non-valid: 73 seconds; $t(475.37) = 4.8675$, $p < 0.001$

one with which the participants are not likely to be familiar. As an example of this, in a different kind of study, Vigo et al. (2014b), used a potato ontology.

9.4. Decidability and tractability

The theoreticians of machine reasoning have played a great deal of attention to the issues of decidability and computational tractability. OWL DL and OWL 2 DL were designed to achieve decidability. OWL 2 EL, OWL 2 RL and OWL 2 QL were designed not just to achieve decidability, but also computational tractability in particular modelling use cases (Horridge et al., 2012). At the same time, a great deal of development effort has been expended developing efficient reasoners for particular language profiles, e.g. ELK[156] for OWL 2 EL. Students of human reasoning, on the other hand, have paid little attention to these issues.

Consider first decidability. Decidability is assured in DLs by observing certain rules, e.g. see the rules for OWL DL (Bechhofer et al., 2004). Failure to observe these rules does not guarantee undecidability, but rather means that decidability cannot be guaranteed. In any case, all the questions in the three studies in this dissertation are decidable, i.e. the validity of the putative conclusion can be proved or a counter-example can be constructed. This is an inevitable consequence of their being within OWL DL.

Lack of computational tractability is a phenomenon which occurs at scale. To take an example from another domain, it is perfectly possible for a human reasoner, with pen and paper, to solve the travelling salesman problem for a limited number of cities by comparing every possible route. With a computer it is possible to solve the problem for many more cities than with unaided human reasoning. The problem is the behaviour of any solution algorithm as the number of cities increases. All the problems in this dissertation can be solved in a reasonable time. What might happen if, for example, the number of quantifiers present in a question was increased very considerably is of no relevance. In effect, working memory limitations are exceeded even for algorithms which might be regarded as computationally tractable, and long before any unreasonable time requirements become apparent.

Levesque (1988) is an author who has considered the tractability of human reasoning, asking how it is possible for humans to perform reasoning tasks which, on the surface, make physically unrealisable demands on resources, e.g. time. He makes the point that difficulties arise in computation when there are a number of possibilities, each of which needs to be considered. Since, each time this happens there is a multiplicative effect, this can give rise to an exponentially increasing number of possible reasoning chains. He suggests there are a number of ways this is avoided in everyday reasoning. One example is the subsumable disjunction. He suggests that a typical disjunction in everyday life is "Joe is 71 or 72 years old". Frequently we do not need to consider each case, but can subsume them into a more generic case, e.g. Joe is over 70. From that one statement we can go on to make all the relevant deductions, e.g. that Joe is entitled to a state pension and a free bus pass. Levesque (1988) contrasts this with a disjunction such as "Joe is 71 years old or Mary has the flu", making the point that the latter type of statement does not occur in everyday life, although it might in puzzles. Such unlikely disjunctions also occur in some of the questions human reasoning researchers pose to study participants, e.g. "June is in Wales, or Charles is in Scotland, but not both" (Johnson-Laird et al., 1992). This is not to discredit such questions

[156] https://www.cs.ox.ac.uk/isg/tools/ELK/

as instruments for investigating human reasoning, but merely to point out that they may stray some way from normal everyday reasoning and that this may explain their relative difficulty.

9.5. Sample size, effect size and significance

The studies described in Chapters 6, 7 and 8 required access to participants with some knowledge of ontologies, or alternatively some background in computer science[157]. Obtaining those participants poses a greater problem than that met by psychologists who are interested in the general human population. The consequent limitation in number of participants places a great emphasis on the need to understand the relationship between sample size, effect size and significance. For any given sample size, effects below a certain size are not likely to be detected as significant. Subsection 9.5.1 discusses the relationship for accuracy data between effect size and required sample size. Subsection 9.5.2 presents a similar discussion for response time data. Response time data exhibits the problem that, particularly at low sample sizes, the mean and standard deviation can be very much influenced by just one extreme data point. This is discussed in subsection 9.5.3. Subsection 9.5.4 then draws on the preceding subsections to make some final comments.

9.5.1. Accuracy and sample size

For study 3, the number of participants was 30. For the laboratory alternative of study 2, with which the study 3 results are compared, there were 28 participants. To illustrate the limitations which such sample sizes place on the determination of a significant effect, consider as an example the comparison between the responses to two questions given by a group of 30 participants. Assume that the accuracy for question 1 was 50%, i.e. exactly chance[158]. This is an extreme case, but one which did occur in practice. Table 9-2 illustrates how the p-value from a Fisher's Exact Test decreases as the accuracy for question 2 increases. Note that even an accuracy of 70%, i.e. an increase in 20 percentage points over the performance for question 1, will not be significant. In fact, there will be no significant effect until the accuracy of the question 2 reaches 80%, i.e. 24 correct responses out of 30; 23 out of 30 (77%) will not suffice ($p = 0.060$).

Table 9-2 Comparison between two questions, with 30 participants, assuming 50% accuracy for question 1

accuracy for question 2	60%	70%	80%	90%	100%
p-value	0.604	0.187	0.029	0.002	0.000006

N.B. using Fisher's Exact Test

Table 9-3 illustrates the other extreme. Here it is assumed that the participants achieve 100% accuracy with question 1. In this case, question 2 would need to be at 80% accuracy or less in order for there to be a significant difference.

[157] The requirement is for study participants who are similar to ontology users, i.e. domain experts with a training in the use of ontologies or computer scientists working in the field of ontologies.

[158] For the sake of simple illustrations, the assumption here is that the samples accurately represent the underlying populations, so that in this case a 50% accuracy would be achieved with the total population. A more rigorous analysis would consider the ensemble of sample accuracies, each with their probability of occurrence, for any given underlying population distribution.

Table 9-3 Comparison between two questions, with 30 participants, assuming 100% accuracy for question 1

accuracy for question 2	50%	60%	70%	80%	90%
p-value	0.000006	0.0001	0.002	0.024	0.237

N.B. using Fisher's Exact Test

Table 9-2 and Table 9-3 represent infrequent extremes. A more likely example, representing a 20 percentage point difference, would be between 60% and 80%. With 30 particpants, this difference would not be significant ($p = 0.158$). In fact, 50 participants would be needed for the difference between 60% and 80% to be significant ($p = 0.049$).

In practice, two other factors need to be taken into consideration. Firstly, usually more than one question was involved in the comparison, e.g. between study 2 and study 3, and this increases sample size, for a given number of participants. Secondly, in study 3 some questions existed as two variants, and this reduces sample size. Each case needs to be analysed separately. The key conclusions are that the sample size needs to be appropriate to the minimum effect size to be identified; and that the kinds of sample sizes used in these studies will only detect relatively large increases in accuracy.

9.5.2. Response time and sample size

The discussion is inevitably more complicated for response time data because the time distribution has two parameters. The analysis, the detail for which is provided in Appendix C, can be simplified by making certain assumptions. Firstly, as throughout the preceding three chapters, it is assumed that response time has a log-normal distribution, i.e. that the log of the response time has a normal distribution; and that the t-statistic is calculated after the log transformation. Secondly, the response time data in the three preceding chapters indicate that the standard deviation increases with increasing mean response time, and that the ratio of standard deviation to mean response time is somewhat less than 1. In fact, the ratio of standard deviation to mean response time for studies 1, 2 and 3 is: 0.71; 0.74; 0.91. The analysis assumes that the ratio is the same for both groups being compared; in the analysis the value of 0.75 is taken as representative. Thirdly, the assumption is made that the sample means and standard deviations are equal to those of the underlying distributions. Under these assumptions, the t-statistic is purely a function of the size of each of the two samples and the percentage difference in mean times between the samples. For simplicity, it is also assumed that the two samples being compared are the same size. Finally, the analysis looks at relative differences in the mean response times of the two samples. Specifically, the analysis has been conducted for differences of 10%, 20%, … 100%[159]. Table 9-4 shows the minimum size for each sample necessary to achieve significance at the 0.05 level. This is based on a two-sided t-test; a one-sided test would require smaller samples.

Table 9-4 Minimum sample size to achieve significance

%age difference	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
min. sample size	379	105	52	32	23	17	14	12	10	9

N.B. size shown is for each of two samples

The table shows that, under these assumptions, even with a 40% difference in mean times, 32 data points would be required in each sample. If the comparison is between two versions

[159] Because of the assumptions made, and the use of the logarithmic transformation, the sample size required for significance will depend only the percentage difference between the mean times, not their absolute values.

of one question in a within-participants design, this would require 32 particpants; a between-participants design would require 64. Where a particular language feature is being investigated, the use of more than one question will also reduce the number of participants needed, at the risk of having a less representative sample. Note that this analysis is necessarily approximate; in particular, in its assumption of a log-normal distribution for time, in the assumption that the sample statistics correspond exactly to the underlying distribution statistics, and in the assumption that the standard deviation to mean ratio is 0.75. It is intended to give only an estimate of the required sample sizes. Van Schaik and Weston (2016) discuss the topic of sample size in detail and provide an extensive set of references.

9.5.3. Response time and outliers

Response time data is subject to another difficulty of interpretation caused by sensitivity to extreme data points, or outliers. An example can be drawn from section 8.5.2. Based on two questions, an ANOVA was used to compare study 2, using *not ... some*; variant 1 of study 3, using *not ... including*; and variant 2, using *not ... any*. The analysis revealed no significant variation across the three cases ($F(2, 105) = 0.602$, $p = 0.550$). However, the mean response times for variant 1 and variant 2 were 40 and 58 seconds, which *prima facie* appears an appreciable difference.

For simplicity, the discussion here will use a t-test to compare the two variants of study 3. There were 15 participants in each variant, i.e. a total of 60 data points, 30 with *not ... including* and 30 with *not ... any*. Despite the appreciable difference in mean response time, a t-test is consistent with the ANOVA in revealing that this difference is not significant ($t(57.832) = 0.79496$, $p = 0.423$). However, the standard deviations of the response times in the samples in variant 1 and variant 2 were 38 and 125 seconds respectively. The latter value is very high, in fact more than twice the sample mean. Examination of the response time data in the tables in Chapters 6, 7 and 8 shows that it was very rare for the sample standard deviation to exceed the mean. Such a high standard deviation will appreciably reduce the value of the t-statistic and hence appreciably increase the p value[160]. Further examination of the variant 1 data reveals one response time of 710.7 seconds[161]. Removal of this one data point reduces the response time mean for variant 2 to 36 seconds, i.e. now slightly less than for variant 1. The standard deviation is now 20 seconds, and the differences between the two groups remains not significant ($t(52.556) = 0.2977$, $p = 0.767$), as would now be intuitively expected given the small difference in means.

This example illustrates that an apparently appreciable difference in mean response times can be caused by one extreme data point. The presence of such a point can often be detected by a disproportionately high value for the standard deviation. This high standard deviation will also have the effect of reducing the t-statistic, indicating that the difference in group means is not significant.

9.5.4. Comments

The point of the discussion in subsections 9.5.1 and 9.5.2 is not to suggest that increasing the sample sizes in the three studies would have necessarily resulted in more significant results. Increasing the sample size is likely to increase the accuracy of the sample statistics,

[160] To an extent, the effect of this high standard deviation is mitigated by the logarithmic transformation of time.

[161] At the beginning of each experiment, participants were encouraged not to take too long over any question. On occasions this advice was repeated when a participant appeared to be considering one question for an excessive time. However, in a few cases very long response times were recorded.

i.e. how closely they represent the statistics of the overall population. In a particular case this might lead to a reduction in the difference between two sample means, so that even with increased sample sizes, the difference remains non-significant. The point being made is that quite large data sets can be needed for even appreciable differences to be regarded as significant.

An ideal approach would be to start by considering what size of difference, i.e. for accuracy and mean response times, would be material and decide on that basis what sample size is required. In their discussion of magnitude-based inference, van Schaik and Weston (2016) stress the need for researchers "to choose a smallest important effect size". For example, when comparing the syntax modifications in study 3 with the original MOS in study 2, a starting point would be to decide what increase in accuracy or reduction in mean response time would justify a change to the language. To an extent this is a subjective decision, although it can be informed by consideration of the practicalities involved in implementing any policy decision arising from the analysis.

Care needs particularly to be taken when comparing different changes, e.g. two different syntactic changes to MOS. If the comparisons are based on different sample sizes, then it might be concluded that one change is significant, the other not, when the difference in significance could be due to sample size. Something analogous to this was noted in section 8.3. The observation was made there that, when considering the object property questions concerned with functionality, there was a significant effect of complexity on accuracy in study 2 but not in study 3; and that this difference might be the result of the difference in sample sizes between the two studies.

The discussion in subsection 9.5.3 raises the question whether it is appropriate to remove extreme data points. As it stands, the mean response time for variant 2, given in Table 8-17 is deceptively large. In all the studies reported in this dissertation there has been no attempt to remove outliers. An alternative would be to provide the statistics, and undertake analysis, before and after the removal of any outliers. For this, an objective definition of an outlier would be required. In the context of the work reported here, a rule of thumb might be any point which causes the sample standard deviation to exceed the sample mean.

An alternative approach to this problem would be to limit the amount of time a participant could spend on any one question. However, this further complicates the statistical analysis. Moreover, it is particularly difficult for the kinds of studies reported here because there is considerable variation in response times between participants and between questions.

9.6. Experimental methodology – lessons learned and recommendations

This section considers the major methodological lessons from the experience of the three user studies, and makes some recommendations for alternative approaches. Subsection 9.6.1 considers methods for achieving adequate sample size and subsection 9.6.2 considers the problem of bias due to the order of the questions. Finally, subsection 9.6.3 discusses the interpretation of response time data and suggests an alternative approach to that used here.

9.6.1. Achieving adequate sample size

The previous section identified the need for adequate sample sizes, bearing in mind what effect size is regarded as important. The requirement for participants with some knowledge of ontologies or computer science makes it hard to achieve satisfactory sample sizes. At the same time, the work has had the pragmatic goal of identifying problems with the commonly used MOS features, and recommending how to overcome those problems. Whilst

psychologists may devote a single study to one aspect of reasoning, e.g. the syllogism, such a focussed study did not seem reasonable in this context. The need was to consider a range of common MOS features and a range of possible changes to syntax. This had the effect of confounding factors. It also had the effect of reducing sample size, in particular in study 3 where two variants were used. In an exploratory study, this is to an extent unavoidable. However, more refined studies should avoid confounding factors in this way.

One approach to achieving greater sample size is exemplified by Nguyen et al. (2012), who used a crowd-sourcing technique. This means that the respondents are chiefly lay-people, i.e. not familiar with ontologies[162], and that the axioms and conclusions need to be represented in an accessible way, i.e. not in a formal language. Nguyen et al. (2012) used 51 questions, each with a set of axioms and a valid conclusion[163]. There were four questions which were answered significantly worse than chance. The least accurately answered question had only 2 correct responses out of 49 (compared with chance: $p = 2 \times 10^{-12}$). It is noteworthy that this question required an understanding of the trivial satisfaction of the universal restriction; it is not clear to what extent this was explained to participants. Another two of the four questions used rather technical language, talking about lack of convertibility between double values and integer values and between string values and integer values. The fourth question, which was answered appreciably better than the others, although still significantly worse than chance, required thinking about inverse properties ("toves from" and "toves into"); the use of different prepositions after the verb to create different property names may have caused difficulties here. It may be that participants were not sufficiently prepared to cope with these questions. In any case, the participants were not representative of the user community.

An alternative is to use participants who may not be familiar with ontologies, but are not randomly selected through crowd sourcing. One possibility would be to use the kind of people who might be regarded as domain experts, i.e. the kind of people who might conceivably be involved in building ontologies, even if they are not currently doing so. They would need to be well prepared for the study. However, for a study which was not exploratory but rather looking at a focussed issue, the degree of preparation could be relatively limited. They would only need to be familiar with that small subset of the language necessary for the investigation, rather than the relatively wide spectrum of language dealt with in the three studies in this dissertation.

As in studies 2 and 3, it is possible to increase the number of data points, without increasing the number of participants, by increasing the number of questions. For example, in the work described in section 8.3.3, one could have more questions employing an inverse functional object property. The drawback to this approach is that the sample may not be as representative of the total population as would be achieved by increasing the number of participants.

Studies 1 and 2 were within-participant studies. The comparisons were between different questions answered by the same participants. However, Chapter 8, in sections 8.3 to 8.6 describes mostly between-participant comparisons, i.e. between study 3 and both studies 1 and 2, and between the two variants of study 3. Between-participant studies introduce variability due to the participants, unlike within-participant studies where paired

[162] Whilst many users of ontologies are not computer scientists but domain experts, the latter will have received some training in ontology development and will differ considerably from a random sample of lay people.
[163] Questions with non-valid conclusions were used as foils but not included in the analysis.

comparisons can be made. Within-participant studies also have the advantage that, within the limit set by the realistic length of a study, twice as many questions can be posed, since each question is put to each participant. For these reasons, a more focussed study to investigate the issues explored in study 3, would benefit from a within participants approach.

A final point on sample size relates to analysis of response time. Section 4.4.1 noted that both correct and incorrect responses were used for the response time analyses, to increase sample size. Analyses based purely on the correct responses would have been a useful complement. Whilst such analyses might have produced fewer statistically significant results, the use and comparison of the two approaches might have provided additional insight.

9.6.2. Avoiding order bias

Analysis from studies 1 (section 6.9) and 2 (section 7.9) showed that question position significantly affected response time, but not accuracy. Analysis from study 3 (section 8.9) confirmed the effect of question position on time, and also indicated an effect on accuracy. This effect was relatively small, but was significant when comparison was made between sections. These analyses indicate that compensating for order effects is prudent when considering accuracy and essential when considering response time.

In the studies reported here, this was done either with an experimental design which ensures that participants view the questions in different orders, or else by randomization of the question order. The former approach was adopted in studies 1 and 2. In study 1 only the order of the sections was permuted. In study 2 the order of the sections was permuted, and within each section the participants saw the questions in two reversed orders[164]. This latter strategy was not effective in compensating for the effect of position on time. It appeared that, in some sections, there was a large time penalty associated with the first, and sometimes also the second, question; and that this was not adequately compensated for by using two reversed orders. In the subsequent data analysis, where necessary any effect of this time penalty was avoided by not using questions which occupied first or second place in a section. Study 3 used randomization to compensate for the effect of section and question order. This also enabled a more rigorous analysis of the effect of question order, not confounded with the effect of any difference in the questions, with the results noted above.

Generally, randomization is an easier approach than attempting to compensate for question order with an experimental design, and there are a number of tools available which enable randomization. In the online alternative of study 2, the *surveygizmo* tool was used to provide question randomization. In study 3, the *MediaLab* application provided automatic randomization of section order and question order within sections.

In the absence of randomization, another approach would be to provide several 'warm-up' questions at the beginning of each section, the data from which would not be used in the analysis. This would avoid any initial effects. Presenting the remaining questions, i.e. the ones actually used in the analysis, in two reversed orders would then compensate for any longer-term effects. Indeed, in study 2, one participant did suggest using "a few examples at the start to warm up the brain" (see subsection 7.7.1). This might even be useful prior to randomized questions, to further eradicate any distortion of response time data associated with the initial questions.

[164] This was for the laboratory study. In the online study there was no control over the order in which participants approached each section, whilst the order of questions in each section was randomized.

In summary, avoiding bias caused by question order is an essential consideration, in particular where timing data is being collected. Whatever approach is used, it needs to be planned before the commencement of the experiments.

9.6.3. The interpretation of response time data

The theories of reasoning discussed in this dissertation are concerned with understanding and predicting reasoning accuracy. Less attention is given to the time taken to perform reasoning. As discussed in section 4.3.1, response time has been taken as a proxy for difficulty. With a few exceptions, there has been no attempt in the analysis to differentiate between response times for correct and incorrect responses. The first two studies were concerned with identifying difficulties, and the decision not to make that differentiation between correct and incorrect responses was justified on the grounds that, even when a question is incorrectly answered, its response time is likely to be a measure of its difficulty. The assumption is that people spend more time on difficult questions than easy questions, even when they get them wrong.

In study 3, the intention was to estimate the effect of changes to MOS syntax. The same approach was taken, i.e. to use response time as a measure of difficulty, regardless of the accuracy of the response. An alternative approach would be to use only the response times from the correct responses. The argument for doing this would be that syntax changes which lead to more rapid response are only of interest if those responses are also correct. The disadvantage is reduced sample size. Chapter 8 used between-participants comparisons, i.e. between studies 2 and 3. However, the approach of using only accurate responses for the time analysis would fit well with a within-participants study, where all participants answered questions with the original and changed syntax. Analysis of the accuracy of responses would identify whether a particular change led to an increase in accuracy. Analysis of response times could be restricted to responses to paired questions where both responses from a particular participant were correct. This would determine whether, even when the original answer was correct and hence could not be improved upon, the syntax change had an effect on the time to produce the correct answer.

10. Conclusions and future research

"Words are all we have. It is no good to find fault with them."
"And yet I do so. They are used as if they had some power. And how little they have!"

Ivy Compton-Burnett, 'A Heritage and Its History', 1959

The introduction to this dissertation presented two research questions, to which this chapter returns. Section 10.1 begins by reviewing to what extent they have been answered. Section 10.2 then suggests additional work to be done to answer them more fully. Section 10.3 expands the original questions into a wider domain. Section 10.4 provides some guidance for researchers undertaking similar studies to those reported in Chapters 6 to 8. Section 10.5 makes some final remarks.

10.1. The research questions – summarising progress

Chapter 1 posed two broad research questions:

> *(1) In what way can the difficulties experienced in using Description Logics be understood in terms of underlying theories, e.g., theories of reasoning, already developed within the cognitive psychology community?*
>
> *(2) In what way could such theories contribute to improving the usability of Description Logics?*

The next two subsections consider each of these questions in turn, reviewing how the questions have been addressed; and how improvements to MOS have been proposed and tested. Although the original focus was on using results from cognitive psychology, it became apparent during the research that insight from the philosophy of language was also valuable, and this will also be discussed.

10.1.1. The theories and their applicability to DLs

In Chapter 1, question 1 was expanded into two more detailed questions:

> *(1A) What theories are available, and what are their relevant strengths in the context of DLs?*
>
> *(1B) How can these theories explain how people understand DL statements?*

Section 9.1 has already reviewed the use of the theories of reasoning, and explained the motivation for concentrating on the rule-based and model-based theories, and the theory of relational complexity. The model-based theory has been particularly useful in explaining the difficulties experienced with negated conjunction, where the correspondence between the results in subsection 6.6.1 and those presented by Khemlani et al. (2012a) is striking. It has also been successful in explaining participant performance with questions involving universal and existential restrictions. With nested restrictions, for example, the mental model theory explains the difficulty caused when the less prominent model needs to be adopted in the first restriction. This leads to the need for backtracking, with consequent difficulties. The concept of the implicature, from the philosophy of language, augments mental model theory by suggesting why one model is more prominent than another. On the other hand, syntactic reasoning also appeared to be present, akin to the rule-based reasoning of Rips (1983). In any case, taken together, the model-based and the syntactic approach

offer a better explanation of performance than the use of alternative hypotheses put forward in Chapter 7, e.g. that reasoning performance with restrictions depends on the number of types of restriction (hypothesis H7.9), and the pattern of use of restrictions (H7.10).

Relational complexity theory proved useful as a yardstick for semantic complexity. It was most obviously applicable to reasoning about object properties, but was also applied elsewhere, e.g. to considering Boolean concept constructors as described in section 7.4 and Appendix B. Not all relations can be considered of equal difficulty. Functional relations gave greater difficulty than transitive ones; perhaps because of the need to remember whether subject or object was unique. However, care needs to be taken in interpreting this. People do not reason like computers, obtaining the characteristics of a property from a look-up table. They absorb the semantics of words, without even necessarily knowing the meaning of terms such as 'transitive' and 'functional'. It may be that the choice of object property names influences reasoning. This was suggested by participants' comments in subsection 7.7.3, commenting that *greater_than_or_equal* was easier than *has_nearest_neighbour*. When framing the questions, the choice of names was found difficult and likely to have not always been ideal. Again as noted in subsection 7.7.3, one participant had difficulty with *has_nearest_neighbour*, arguing quite reasonably that this was not necessarily functional. Yet the alternative, of using abstract or nonsense names, deviates a long way from natural usage. A better solution might be to work within the participant's own subject domain; but this is time-consuming and raises problems of comparability between participants.

10.1.2. Enhancing the usability of Description Logics

Chapter 1 also expanded on question 2:

> *How do these theories motivate:*
>
> *(2A) notational extensions;*
>
> *(2B) tool enhancements;*
>
> *(2C) enhancements to training?*

The preceding chapters have made suggestions to enhance the syntax of MOS. Some of these suggestions come from consideration of the theories discussed; others are more ad-hoc. The use of *union* and *intersection* rather than *or* and *and*, and the use of *including* and *noneOrOnly* for the existential and universal restrictions are in the former category. The use of *except* for *and not*, and the use of *solely* with functional properties are in the latter category; although identifying the need for additional support for functional properties was guided by the use of RC theory.

Apart from notational extensions to MOS, tool support could offer help to ontologists. The presentation of alternative syntactic forms could be valuable. Subsection 7.5.6 discussed pairs of questions which were semantically equivalent but which were answered with appreciably different degrees of accuracy. The equivalence of the questions arose from the equivalences:

has_child some (not MALE) ≡ not (has_child only MALE) (10.1)

has_child only (not MALE) ≡ not (has_child some MALE) (10.2)

Providing both equivalent forms in, e.g. a chain of deductions, might help readers to follow the reasoning, and also guard against the fallacies of equating *some not* with *not … some* and *only not* with *not … only*.

The quotation at the start of the chapter was deliberately provocative. Words are not all we have; we also have pictures, and pictures can be used to reason with. Larkin and Simon (1987) note the importance of diagrams for search, recognition and inference. As Ford (1995) has pointed out, some people appear to think more naturally in pictures than in words. Similarly, Petre and Blackwell (1999), in a study of mental imagery in programming, found evidence of visual, as well as textual and auditory, imagery. Moreover, diagrams have been part of mathematical reasoning at least since Euclid. Cheng (2012) discusses the use of diagrams in syllogistic reasoning; here the diagrams are not just used to visualize syllogisms, but also to reason about them. The use of diagrams in ontology engineering is an active research area, e.g. Stapleton et al. (2013; 2014). Their work concentrates on representing a whole ontology, or substantial parts of an ontology; the automatic creation of diagrams to aid individual reasoning steps would also be useful, e.g. to illustrate class membership and emphasize the direction of uniqueness in functional object properties.

Finally, the three studies suggest a number of areas worth particular emphasis in DL training. De Morgan's Laws are one; this should particularly mitigate the problem of negated conjunction. The analogues of De Morgan's Laws for restrictions, as exemplified by (10.1) and (10.2) above, are another; this should help when thinking about negation and restrictions, as in section 7.5. When introducing these equivalences, the false equivalences of *some not* with *not … some* and *only not* with *not … only* should be pointed out. Stress also needs to be placed on functional and inverse functional object properties, so that there is a clear awareness of the difference.

10.2. The research questions – further work

There remain some unresolved issues, within the scope of the research questions posed in Chapter 1. Subsection 10.2.1 describes some of these issues. Some have been studied but require further investigation; others were not examined in the current studies. The approach used here was to create controlled experiments based on commonly used language features. Other approaches could be applied to the research questions, and some of these alternative approaches are described in subsection 10.2.2.

10.2.1. Unresolved questions

There are a number of other commonly used OWL features which have not been investigated. Amongst these are the use of object properties defined to be inverse. This was relatively prominent in the survey described in Chapter 5 and also in the analysis of Power and Third (2010), as summarised in Table 6.1 in section 6.1. Some language features were investigated in the first study but were not prioritized for investigation in the subsequent studies. The use of object subproperties is an example; an alternative keyword to *SubPropertyOf*, as suggested in subsection 6.12.1, could avoid the erroneous 'inheritance' of property characteristics. The use of two restrictions for which the object properties are in a subsumption relation also comes into the category of language constructs examined in study 1 but not further investigated; in question 6 of section 6.5 difficulties seem to have been experienced with the combination of a subclass and the existential restriction.

There are also some outstanding issues arising from study 3. Whilst the use of *including* and *noneOrOnly* for the existential and universal restrictions achieved some benefit when a single restriction was used to define a class, this was not the case for nested restrictions.

More focussed research is needed to investigate this, and also to separate out the effects for the existential and universal restrictions. The limited amount of data for the separate cases made this difficult in study 3. Secondly, there remains the issue of how to improve performance with inverse functional object properties. As was suggested in subsection 8.3.3, possible areas to investigate would be the use of a different keyword, e.g. *alone* and moving the keyword to follow the subject, i.e. immediately before the object property.

10.2.2. Alternative methodologies

The starting point for study 1, which then influenced the subsequent two studies, was the identification of MOS language features most commonly used in creating ontologies. An alternative strategy would be to look at the kinds of justifications which arise in debugging ontologies. One approach to this would be to ask ontologists, e.g. through the appropriate lists as used for the survey of Chapter 5, to submit justifications which have created difficulty for them. Alternative approaches are to arrange focus groups with ontologists or to observe them as they work.

At the end of subsection 10.1.1, there was a discussion of the necessarily artificial nature of the kinds of studies described in this dissertation. The discussion was particularly related to object property names, but could also be widened to class and individual names. More broadly, an approach which abstracts from a real-world environment offers the advantage of focussing on a few particular factors at the expense of observing those factors removed from the context in which they normally have an effect. In the case of entity names, the strategies of using abstract names or 'nonsense' names, or even names chosen from a field unknown to the participants, may cause those participants to reason in a quite unnatural way. They may be obliged to explicitly refer to the entity name characteristics, rather than using an acquired understanding. From this viewpoint, the accepted practice of using an unknown or artificial domain may not be ideal. Rather, it would be better to use names from the participants' actual ontology usage. In doing this, care has to be taken to understand what semantics are normally associated with particular entity names e.g. which characteristics are associated with particular object properties, and whether particular classes are assumed to be disjoint.

Another approach to investigating how ontologists work in real-world situations is exemplified by Vigo et al. (2014b; 2015), who have instrumented Protégé to record and subsequently analyse user interactions. Whilst this approach can provide valuable insights into how people use tools such as Protégé, it may give limited access to thought patterns and difficulties of reasoning. Indeed, the recommendations in Vigo et al. (2015) are, quite reasonably, concerned with tool improvement.

Controlled experiments, as used here, can offer precise answers in well-defined situations[165]. They also have the advantage of being relatively easy to administer. Yet, they are to an extent unnatural; whilst that unnaturalness can be mitigated it can never be totally eradicated. Case studies (Easterbrook et al., 2008) offer insights based on observation. Whilst these are harder to interpret, they complement the results of controlled experiments. Indeed, Easterbrook et al. (2008, section 5.6) talk about "triangulation", i.e. "using different sources

[165] In comparing "the three fundamental paradigms … introspection, field studies, and controlled experiments", Shneiderman (1978) believes that "all three are useful, but that controlled experiments must ultimately be the basis for the most profound theoretical and practical conclusions." North (2006), discussing specifically visualization research whilst making some observations which are more generally applicable, provides a critique of benchmark tasks. He argues for more complex benchmark tasks and qualitative studies with open-ended tasks.

of data to confirm results and build a coherent picture". The research questions posed at the beginning of this dissertation could well be revisited through the technique of the case study.

Finally, the emphasis in this work has been on reasoning with DL statements. In line with this, study 3 investigated how well the proposed keywords improved reasoning. An additional question is how well any change to notation supports the creation of ontologies, i.e. the act of modelling using DLs. This would require an alternative experimental approach in which participants were required to undertake modelling exercises.

10.3. Expanding the research horizon

The original research questions were framed in the context of DLs, but they could be translated into the context of any other computing language, particularly those languages which are logic-based. This section looks at two areas in which analogous research questions could be posed. Subsection 10.3.1 looks at a long-standing area, that of database interaction; whilst subsection 10.3.2 looks at the more recent area of linked data querying and results presentation.

10.3.1. Interacting with databases

There is a long history of research into interaction with databases. A relatively early example is the work of Thomas and Gould (1975) who looked at a query system using a formal language and anticipated a number of the issues discussed in this dissertation. They reported that participants to their study had difficulty with universal quantification and with certain operators, e.g. COUNT and SUM. Conversely, they had little difficulty with conjunction and disjunction. The authors ascribe this to the fact that their system is "nearly wordless", using instead a tabular approach; i.e. there is no possibility of confusing *and* and *or* because these words are not used. Indeed, the paper starts with a critique of the use of natural language, arguing against its use. The authors also make some methodological comments. They note that studies such as theirs are concerned with translating test questions, whereas in real life users are concerned with generating their own questions; and they identify the need to understand how experts write queries, rather than the non-experts used in their study.

Subsection 2.5.3 has already noted Shneiderman's (1978) comparison of the use of natural and artificial language in database querying. As already noted, he found that whilst the number of valid queries was approximately the same in both cases, natural language participants posed more questions that could not be answered from the database. He includes a critique of natural language usage; amongst other issues pointing out that it may create an unrealistic expectation of what questions can be answered. Indeed, he includes the issue of natural versus artificial language as an important research question in database interaction. Shneiderman (1978) also includes a discussion of the difficulties of universal quantification, drawing on a number of previous authors.

More recently, Jagadish et al. (2007) discuss database usability, making comparisons between the differing expectations of database use and web search. Having previously been involved in a variety of other aspects of database research, they describe their motivation for becoming interested in usability research. This originated from observing the difficulties experienced by users who were professionals but not computer scientists, i.e. users analogous to domain experts in ontology design. They identify the need for database systems to be able to explain their results; a feature analogous to providing justifications in ontology debugging. In another echo of a problem experienced with ontology use, one of their observations was that users were having difficulty with "the mere complexity of the …

schema". Their solution was to develop schema representations "at different levels of detail", so that users could zoom in on the part of the schema of interest.

The previous paragraphs have provided inevitably a tiny snapshot from the history of database research. It is striking that many of the themes have also occurred in the context of ontology development. It is also striking that there is no reference to theories of reasoning or language, although the theme of using natural language does recur. On a specific point, Jagadish et al. (2007) observe that joins are hard to reason with; this could well be a topic to investigate in the context of theories of reasoning. Generally, insights from psychology and language could be applied to the usability of databases, both as experienced by experts and non-experts. Additionally, there is much that researchers investigating the usability of ontologies can learn from database usability research.

10.3.2. Interacting with linked data

The view is sometimes expressed that the Semantic Web has not developed as originally envisaged, in particular that there has been limited use of OWL. Instead we have seen the growth of the LOD cloud, which is chiefly based on RDFS plus a few OWL features such as *sameAs* and functional object properties. Chapter 2 has already described Hendler's (2015) distinction between 'big O' ontologies and 'small o' ontologies. The former find their application in specialized fields such as biological research. The latter are found at Web scale constituting the LOD cloud.

There is considerable use of 'big O' ontologies, as chapter 5 has demonstrated. The work reported in this dissertation is chiefly aimed at this world. Some of the work, e.g. relating to functional object properties, is certainly relevant also to the LOD world. However, in the LOD world, whilst there is limited complexity in the ontology design, SPARQL offers the possibility of complex queries. A number of studies have investigated to what extent that possibility has been taken up. Möller et al. (2010), looking at four datasets, found that most queries were SELECT queries and that almost all of those queries had at most three triple patterns, although some queries were found with up to 16 triple patterns. They also found that three pattern types constituted almost all the patterns for three of the datasets and around 86% for the other dataset. Gallego et al. (2011) looked at two datasets and reported similar results. They also found that the dominant query was SELECT. The great majority of queries had only one triple pattern. Similarly to Möller et al. (2010), Gallego et al. (2011) found a few pattern types constituting most of the patterns; however, there were some differences in the detail regarding the particular pattern types.

One approach to making SPARQL querying easier is by providing a visual editor for queries. One such is OWL2Query[166]. This is one of three ontology visualization tools compared specifically from the standpoint of information retrieval by Ramakrishnan and Vijayan (2014). In fact, in their study participants obtained the required information more quickly using a general ontology navigation tool, OntoGraf[167]. However, as the authors point out, OWL2Query had more features than OntoGraf, and this may make it more difficult to use, at least for novices.

There have been attempts to make querying easier through the use of controlled natural language. Rico et al. (2015) have developed such a query system. They note that there are two problems for those using SPARQL: mastering the syntax and knowing the vocabulary

[166] http://protegewiki.stanford.edu/wiki/OWL2Query
[167] http://protegewiki.stanford.edu/wiki/OntoGraf

of the queried dataset. Their solution to both problems is a system that guides users by proposing possible completions of a query in a controlled natural language. This depends on there being a relatively limited number of query patterns. In fact, similarly to Möller et al. (2010) and Gallego et al. (2011), they claim that, for two datasets investigated a very few pattern types account for most queries.

The requirement to know the vocabulary of a dataset can also be met by creating a CNL specific to an application. Chang et al. (2015) describe a French CNL designed for translation into SPARQL, for use in the analysis of computer logs represented as RDF. Consequently, it includes temporal operators such as the French language equivalent of *before*, *after*, *contains* and *overlaps*.

Research into SPARQL is following the same path as that trod by research into DLs. Usage studies are identifying what constructs are being used; natural language and graphical techniques are being investigated as alternatives to writing SPARQL. This opens up a number of research questions, some parallel to those investigated for DLs:

- Why are the great majority of queries relatively simple: because this is all users need; because this is all users can conceive; or because more complex queries are too difficult for users to implement?
- What are the relative advantages of formal language, natural language and graphical techniques for creating queries?
- What are the advantages, and best ways of using, application specific query language, e.g. as mentioned above for time-related queries (Chang et al., 2015)?

Answering these questions could draw upon insights from psychology and language, and also the work of database usability researchers. Additionally, two features differentiate working with the LOD cloud from working with databases: information is likely to be incomplete, and may be incorrect or even inconsistent. These features are shared with ontologies but are likely to be exaggerated with public LOD data. Supporting users in finding and interpreting data in the LOD cloud could well be helped by an understanding of how people reason with incomplete, potentially incorrect and inconsistent data. This might, for example, draw on some of the alternative theories of reasoning discussed in subsection 9.1.4.

10.4. Guidance for practitioners

This section draws on the experience of the studies reported in Chapters 6 to 8 to provide some guidance to those undertaking similar studies, not necessarily in the same area of research, but requiring similar methodological approaches.

10.4.1. Randomizing question order

It is clear from study 2 that question order needs to be randomized to obtain unbiased timing data. Failure to do this in study 2 meant that some data had to be discarded, reducing sample size and limiting the chance of achieving significant results. Moreover, the analysis in study 3 suggests that randomization is also prudent when only accuracy data is being collected. The availability of relatively cheap applications such as *MediaLab*, used in study 3, makes this randomization very easy. Moreover, this avoids the time-consuming manual recording of timing data, as was done for studies 1 and 2.

10.4.2. Avoiding confounding factors

Studies 2 and 3 were over-ambitious, attempting to investigate a considerable number of factors. This led to confounding factors. Separating those factors out, to the extent that it was possible, led to reduced sample sizes. For example, in study 3 the use of identical questions for all participants, rather than two variants, would have increased the sample size for many of the tests. This would have been at the expense of removing some of the possible areas of investigation, e.g. the use of *any* in place of *some* when preceded by a negation. There is a trade-off here. In an exploratory study, it may be justifiable to have many factors. In this case it needs to be understood that it will difficult to achieve significant results. Appreciable, but non-significant effects will need to be interpreted as indicative of the need for further study. For a definitive study, confounding effects need to be avoided.

10.4.3. Designing the study for an appropriate sample size

Subsection 9.6.1 has discussed the issue of achieving an adequate sample size. This will depend on the size of effect considered important. In some cases, the minimum effect of importance might be calculated by a formal cost-benefit analysis. More usually, it will depend on a subjective view. In the context of the work reported here, that might be a decision as to how big an increase in accuracy or reduction in reasoning time would make an appreciable difference to someone working with DLs. In any case, the decision as to what constitutes an appreciable effect needs to be made when the study is being designed. Once that decision is made, then it will be possible to calculate the appropriate sample size. Applications are available for such calculations. However, it may be useful to call on the assistance of a qualified statistician.

10.4.4. Achieving ecological validity

Chapter 6 discussed the need for ecological validity in terms of identifying the commonly used DL constructs, and placing those constructs in a commonly used context, as identified from a commonly used pattern. To generalise, this is about investigating features which occur frequently. However, this is not the only aspect of designing for ecological validity. Another aspect is making those investigations in an environment which is as close as possible to the working environment of the final user. This aspect of design may be interpreted in different ways, depending on the nature of the study. In the studies reported here, a relevant question was how much domain knowledge the participants should be able do draw on. Pre-existing domain knowledge should not permit the participant to know the answer, avoiding any reasoning. Depending on the goal of the study, it might include any knowledge which the participant would normally have available in real-life situations, e.g. knowledge that a particular object property has a particular characteristic, such as transitivity.

10.4.5. Non-valid conclusions

A more specific issue is the design of non-valid putative conclusions; or indeed any sort of answer which the participant should recognise as incorrect. If the question is not to be too easy, these conclusions need to be credible. More specifically, if two or more questions with non-valid conclusions are being used for comparison, e.g. to estimate the efficacy of a particular language change, then ideally the same non-valid conclusion should be used in both questions. If the nature of the questions is such that this is not possible, then the putative conclusions must be designed to have equal credibility. The latter may be extremely difficult. In some situations, it may be more appropriate to use the questions with non-valid conclusions as foils, and base all analysis on the questions with valid conclusions.

10.4.6. Achieving an absolutely fair comparison

In designing experiments, the researcher needs always to be on guard against any possible bias. For example, in studies 2 and 3 many of the participants were already familiar with MOS. When the two studies were being compared, this may have introduced a bias against the syntax changes. Indeed, in Section 8.7 one of the study 3 participants is reported as finding *not ... any* difficult because it was different from what the participant was used to. Ideally, both MOS and the modified MOS should have been unfamiliar to the participants. In practice, it would have been impossible to obtain a reasonable number of such particpants, within the constraint that participants were required with some knowledge of ontologies or computer science.

10.4.7. Alternative strategies

A final recommendation is to be prepared to consider alternative strategies. In these studies, obtaining participants with the required background was difficult. An alternative strategy would have been to have looked for participants who did not use ontologies, neither as ontology experts nor domain experts, but who were experts in particular domains and might potentially use ontologies. This is likely to require more preparation for the participant, but makes possible appreciably greater numbers of participants. It also avoids the kind of bias discussed in the preceding subsection.

10.5. Final remarks

This dissertation has put forward the thesis that insights from the psychological theories of reasoning, and from the philosophy of language can be of value in understanding the difficulties in working with computer languages, and can help mitigate those difficulties. Taking DLs as an example, the work described here has used those insights to explain reasoning performance and to suggest strategies for improving that performance. These suggested strategies have comprised notational extensions, tool enhancements, and enhancements to training. In a final study, certain notational extensions met with some success. At the same time, a number of unresolved questions have been identified. Whilst the work reported here has been limited to applying particular theories, the dissertation has also suggested how it might be expanded to take account of a broader range of theories. In addition, there are other areas of computer interaction to which these theories might be applied, specifically interaction with databases and LOD. Finally, in the course of three studies into reasoning with DLs, some lessons have been learned about how to conduct such studies. In particular, sample sizes should be appropriate for the effect size of interest; experimental design should be able to isolate individual factors, rather than have a number of factors confounded; and question order should be randomized or systematically permuted between participants. These lessons have wide applicability in human-computer interaction studies.

Glossary of acronyms

ACE	Attempto Controlled English (Fuchs et al. 2006)
CNL	Controlled Natural Language
CWA	Closed World Assumption
DAML	DARPA Agent Markup Language
DARPA	Defense Advanced Research Projects Agency
DL	Description Logic
DNF	Disjunctive Normal Form
F-Logic	Frame Logic
FOL	First Order Logic
HSD	Honest Significant Difference, as in 'Tukey HSD'
HTML	HyperText Markup Language
IHMC	Florida Institute for Human and Machine Cognition
LOD	Linked Open Data
MLL	Manchester-Like Language, a simplified form of MOS (Kuhn, 2013)
MOS	Manchester OWL Syntax
OBO	Open Biomedical Ontologies
ODP	Ontology Design Pattern
OIL	Ontology Inference Layer
OPPL	Ontology Pre-Processor Language (Aranguren et al. 2011)
OWA	Open World Assumption
OWL	Web Ontology Language
PL	Propositional Logic
RC	Relational Complexity
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
ROO	Rabbit to OWL Ontology authoring (Denaux et al., 2010)
SOS	Sydney OWL Syntax (Cregan et al. 2007)
SPARQL	SPARQL Protocol and RDF Query Language[168]
SWRL	Semantic Web Rule Language
UML	Unified Modelling Language
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	Extensible Markup Language

[168] SPARQL is an example of a recursive acronym, see http://lists.w3.org/Archives/Public/semantic-web/2011Oct/0041.html

References

Allemang, D., & Hendler, J. (2011). *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC.

Antoniou, G., & van Harmelen, F. (2004). *A semantic web primer*. MIT press.

Aranguren, M.E., Iannone, L., & Stevens, R. (2011). *Ontology Pre-Processor Language (OPPL): User's Manual*. Retrieved from http://oppl2.sourceforge.net/manual.pdf

Baader, F. (2003). *The description logic handbook: theory, implementation, and applications*. Cambridge: Cambridge University Press.

Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *An Introduction to Description Logic*. Cambridge University Press.

Baader, F., & Nutt, W. (2003). Basic description logics. In *Description logic handbook* (pp. 43–95).

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuiness, D., Patel-Schneider, P., & Stein, L. (2004). OWL Web Ontology Language Reference. Retrieved August 30, 2016, from https://www.w3.org/TR/owl-ref/

Berners-Lee, T., & Fischetti, M. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation.

Berners-Lee, T., Hendler, J., Lassila, O., & others. (2001). The semantic web. *Scientific American*, *284*(5), 28–37.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(3), 154–165.

Blake, A., Stapleton, G., Rodgers, P., Cheek, L., & Howse, J. (2012). Does the orientation of an Euler diagram affect user comprehension? In *DMS* (pp. 185–190).

Blake, A., Stapleton, G., Rodgers, P., Cheek, L., & Howse, J. (2014). The impact of shape on the perception of Euler diagrams. In *International Conference on Theory and Application of Diagrams* (pp. 123–137). Springer.

Blomqvist, E., Gangemi, A., & Presutti, V. (2009). Experiments on pattern-based ontology design. In *Proceedings of the fifth international conference on Knowledge capture* (pp. 41–48).

Blomqvist, E., Presutti, V., Daga, E., & Gangemi, A. (2010). Experimenting with eXtreme design. *Knowledge Engineering and Management by the Masses*, 120–134.

Braine, M. D. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*(1), 1.

Braine, M. D., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review*, *98*(2), 182.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*(3), 247–303.

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61–83.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. CRC.

Carston, R. (1998). Informativeness, relevance and scalar implicature. *PRAGMATICS AND BEYOND NEW SERIES*, 179–238.

Chang, B. K. W., Lefevre, M., Guin, N., & Champin, P.-A. (2015). SPARE-LNC: un langage naturel contrôlé pour l'interrogation de traces d'interactions stockées dans une base RDF. In *IC2015*. Rennes, France.

Chapman, P., Stapleton, G., Howse, J., & Oliver, I. (2011). Deriving sound inference rules for concept diagrams. In *Visual Languages and Human-Centric Computing (VL/HCC), 2011 IEEE Symposium on* (pp. 87–94). IEEE.

Cheng, P. C. (2012). Visualizing syllogisms: category pattern diagrams versus Venn diagrams. Retrieved from http://sro.sussex.ac.uk/41053/

Cockburn, A., & McKenzie, B. (2001). 3D or not 3D?: evaluating the effect of the third dimension in a document management system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 434–441). ACM.

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, *13*(6), 377–387. https://doi.org/10.1145/362384.362685

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., … others. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on the World Wide Web*, *17*, 25–32.

Costa, A., Foucart, A., Arnon, I., Aparici, M., & Apesteguia, J. (2014). "Piensa" twice: on the foreign language effect in decision making. *Cognition*, *130*(2), 236–254.

Craik, K. J. W. (1967). *The nature of explanation*. CUP Archive.

Cregan, A., Schwitter, R., & Meyer, T. (2007). Sydney OWL Syntax-towards a Controlled Natural Language Syntax for OWL 1.1. In *OWLED* (Vol. 258).

Dau, F., & Eklund, P. (2008). A diagrammatic reasoning system for the description logic ALC. *Journal of Visual Languages & Computing*, *19*(5), 539–573.

Denaux, R., Dimitrova, V., Cohn, A. G., Dolbear, C., & Hart, G. (2010). Rabbit to OWL: ontology authoring with a CNL-based tool. In *Controlled Natural Language* (pp. 246–264). Springer.

Dimitrova, V., Denaux, R., Hart, G., Dolbear, C., Holt, I., & Cohn, A. G. (2008). *Involving domain experts in authoring OWL ontologies*. Springer.

do Amaral, F. N. (2013). Model outlines: A visual language for DL concept descriptions. *Semantic Web*, *4*(4), 429–455.

Dodds, L., & Davis, I. (2012). *Linked Data Patterns*. Retrieved from http://patterns.dataincubator.org/book/

Easterbrook, S., Singer, J., Storey, M. A., & Damian, D. (2008). Selecting empirical methods for software engineering research. *Guide to Advanced Empirical Software Engineering*, 285–311.

Ehrlich, K., & Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 296–306.

Engelbrecht, P., Hart, G., & Dolbear, C. (2010). Talking rabbit: a user evaluation of sentence production. In *Controlled Natural Language* (pp. 56–64). Springer.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.

Ericsson, K., & Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Euler, L. (1843). *Lettres à une princesse d'Allemagne: sur divers sujets de physique & de philosophie*. PPUR presses polytechniques.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459.

Falbo, R., Guizzardi, G., Gangemi, A., & Presutti, V. (2013). Ontology patterns: clarifying conceptsa and terminology. Presented at the Workshop on Ontology and Semantic Web Patterns - WOP2013, Sydney, Australia. Retrieved from http://ontologydesignpatterns.org/wiki/WOP:2013

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition*, *54*(1), 1–71.

Fuchs, N. E., Kaljurand, K., & Schneider, G. (2006). Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces. In *FLAIRS Conference* (Vol. 12, pp. 664–669).

Gallego, M. A., Fernández, J. D., Martínez-Prieto, M. A., & de la Fuente, P. (2011). An empirical study of real-world SPARQL queries. In *1st International Workshop on Usage Analysis and the Web of Data (USEWOD2011) at the 20th International World Wide Web Conference (WWW 2011), Hyderabad, India*.

Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In *Handbook on ontologies* (pp. 221–243). Springer.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., … Tu, S. W. (2003). The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, *58*(1), 89–123.

Glimm, B., Hogan, A., Krötzsch, M., & Polleres, A. (2012). OWL: Yet to arrive on the Web of Data? *arXiv Preprint arXiv:1202.0984*.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., & Wang, Z. (2014). HermiT: an OWL 2 reasoner. *Journal of Automated Reasoning*, *53*(3), 245–269.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan, *Syntax and Semantics, Volume 3: Speech Acts* (pp. 41–58).

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220.

Halford, G. S, Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(06), 803–831.

Halford, Graeme S., & Andrews, G. (2004). : The development of deductive reasoning: How important is complexity? *Thinking & Reasoning*, *10*(2), 123–145.

Halford, Graeme S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*(1), 70–76.

Halford, Graeme S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences*, *11*(6), 236–242.

Halford, Graeme S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*(11), 497–505.

Halpin, T., & Curland, M. (2011). Enriched support for ring constraints. In *On the Move to Meaningful Internet Systems: OTM 2011 Workshops* (pp. 309–318). Springer.

Hammar, K. (2014). Ontology Design Patterns: Improving Findability and Composition. In *The Semantic Web: ESWC 2014 Satellite Events* (pp. 3–13). Springer.

Hart, G., Dolbear, C., & Goodwin, J. (2007). Lege Feliciter: Using Structured English to represent a Topographic Hydrology Ontology. In *OWLED*.

Hendler, J. (2015). *On beyond OWL*. Presented at the OWLED 2015. Retrieved from http://www.slideshare.net/jahendler/on-beyond-owl-challenges-for-ontologies-on-the-web

Hitzler, P., Kroetzsch, M., Parsia, B., & Rudolph, S. (2012). *OWL 2 Web Ontology Language Primer (Second Edition)*. Retrieved from http://www.w3.org/TR/owl2-primer/

Hoekstra, R. (2009). *Ontology Representation : design patterns and ontologies that make sense* (Ph.D. dissertation). Retrieved from http://dare.uva.nl/document/2/68623

Hopkins, W., Marshall, S., Batterham, A., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine+ Science in Sports+ Exercise*, *41*(1), 3.

Horridge, M., Aranguren, M. E., Mortensen, J., Musen, M., & Noy, N. F. (2012). Ontology Design Pattern Language Expressivity Requirements (Vol. 929). Presented at the WOP 2012 workshop on ontology patterns, Boston, MA.

Horridge, M., Bail, S., Parsia, B., & Sattler, U. (2011). The cognitive complexity of OWL justifications. *The Semantic Web–ISWC 2011*, 241–256.

Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., & Wang, H. H. (2006). The manchester owl syntax. *OWL: Experiences and Directions*.

Horridge, M., & Patel-Schneider, P. (2012). *OWL 2 Web Ontology Language Manchester Syntax (Second Edition)*. Retrieved from http://www.w3.org/TR/owl2-manchester-syntax/

Horridge, Matthew. (2011). A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools Edition1. 3. *The University of Manchester*.

Horridge, Matthew, & Patel-Schneider, P. F. (2008). Manchester syntax for OWL 1.1. *OWL: Experiences and Directions, Washington, DC*.

Horrocks, I. (2002). DAML+OIL: A Description Logic for the Semantic Web. *IEEE Data Eng. Bull.*, *25*(1), 4–9.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., Dean, M., & others. (2004). SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member Submission*, *21*, 79.

Howse, J., Stapleton, G., Taylor, K., & Chapman, P. (2011). Visualizing ontologies: a case study. *The Semantic Web–ISWC 2011*, 257–272.

Hussain, A., Latif, K., Rextin, A., Hayat, A., & Alam, M. (2014). Scalable visualization of semantic nets using power-law graphs. *Applied Mathematics & Information Sciences*, *8*(1), 355–367.

Iannone, L., Rector, A., & Stevens, R. (2009). Embedding knowledge patterns into owl. *The Semantic Web: Research and Applications*, 218–232.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Jagadish, H. V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., & Yu, C. (2007). Making database systems usable. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (pp. 13–24). ACM.

Johnson-Laird, P. N. (2004). The history of mental models. In K. Manktelow & M. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (p. 179). Psychology Press.

Johnson-Laird, P. N. (2010). Against logical form. *Psychologica Belgica*, *50*(3), 193–221.

Johnson-Laird, Philip N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*(1–3), 189–209.

Johnson-Laird, Philip N. (1999). Deductive reasoning. *Annual Review of Psychology*, *50*(1), 109–135.

Johnson-Laird, Philip N. (2005). Mental models and thought. *The Cambridge Handbook of Thinking and Reasoning*, 185–208.

Johnson-Laird, Philip N, & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61.

Johnson-Laird, Philip N., & Byrne, R. M. (1989). Only Reasoning. *Journal of Memory and Language*, *28*(3), 313–330.

Johnson-Laird, Philip N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646.

Johnson-Laird, Philip N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418.

Johnson-Laird, Philip N., Byrne, R. M., & Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. *Psychological Review*, *96*(4), 658.

Johnson-Laird, Philip Nicholas, & Byrne, R. M. (1991). *Deduction*. Lawrence Erlbaum Associates, Inc.

Jupp, S., Horridge, M., Iannone, L., Klein, J., Owen, S., Schanstra, J., … Stevens, R. (2012). Populous: a tool for building OWL ontologies from templates. *BMC Bioinformatics*, *13*(Suppl 1), S5.

Kaljurand, K. (2007). *Attempto controlled English as a Semantic Web language, Ph.D. thesis*. Faculty of Mathematics and Computer Science, University of Tartu.

Kaljurand, K. (2008). ACE View—an Ontology and Rule Editor based on Attempto Controlled English. In *OWLED*.

Kalyanpur, A., Parsia, B., Sirin, E., & Cuenca-Grau, B. (2006). *Repairing unsatisfiable concepts in OWL ontologies*. Springer.

Kalyanpur, A., Parsia, B., Sirin, E., Grau, B. C., & Hendler, J. (2006). Swoop: A web ontology editing browser. *Web Semantics: Science, Services and Agents on the World Wide Web*, *4*(2), 144–153.

Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods—a survey. *ACM Computing Surveys (CSUR)*, *39*(4), 10.

Kazakov, Y., Krötzsch, M., & Simancik, F. (2012). ELK Reasoner: Architecture and Evaluation. In *ORE*.

Khan, M. T., & Blomqvist, E. (2010). Ontology design pattern detection-initial method and usage scenarios. In *SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing* (pp. 19–24).

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427.

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012a). *Negating compound sentences*. Naval Research Lab, Washington DC, Navy Center for Applied Research in Artificial Intelligence. Retrieved from http://mindmodeling.org/cogsci2012/papers/0110/paper0110.pdf

Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012b). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*(5), 541–559.

Kifer, M., & Lausen, G. (1989). F-logic: a higher-order language for reasoning about objects, inheritance, and scheme. In *ACM SIGMOD Record* (Vol. 18, pp. 134–146). ACM.

Křemen, P., Šmíd, M., & Kouba, Z. (2011). OWLDiff: A practical tool for comparison and merge of OWL ontologies. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on* (pp. 229–233). IEEE.

Krötzsch, M., Simancik, F., & Horrocks, I. (2012a). A description logic primer. *arXiv Preprint arXiv:1201.4089*. Retrieved from http://arxiv.org/abs/1201.4089

Krötzsch, M., Simancik, F., & Horrocks, I. (2012b). A description logic primer. *arXiv Preprint arXiv:1201.4089*.

Krötzsch, M., Vrandečić, D., & Völkel, M. (2006). Semantic mediawiki. In *International semantic web conference* (pp. 935–942). Springer.

Kuhn, T. (2013). The understandability of OWL statements in controlled English. *Semantic Web*, *4*(1), 101–115.

Kuhn, T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics*, *40*(1), 121–170.

Lakoff, G. (1987). Women, fire, and dangerous things. *University of Chicago Press, Chicago*.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100.

Larson, S. D., & Martone, M. (2013). NeuroLex. org: an online framework for neuroscience knowledge. *Frontiers in Neuroinformatics*, *7*, 18.

Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition*, *136*, 123–134.

Levesque, H. J. (1988). Logic and the complexity of reasoning. *Journal of Philosophical Logic*, *17*(4), 355–389.

Lord, P. (2015). *Take Wing: building ontologies with Tawny-OWL*. Retrieved from http://homepages.cs.ncl.ac.uk/phillip.lord/take-wing/take_wing.pdf

Lord, Phillip. (2013). The Semantic Web takes Wing: Programming Ontologies with Tawny-OWL. *arXiv Preprint arXiv:1303.0213*.

Matassoni, M., Rospocher, M., Dragoni, M., & Bouquet, P. (2014). Authoring OWL 2 ontologies with the TEX-OWL syntax.

McDonald, J. (2014). Mutliple comparisons. In *Handbook of Biological Statistics*. Baltimore: Sparky House Publishing. Retrieved from http://www.biostathandbook.com/multiplecomparisons.html

Mendonça, E. A., Cimino, J. J., Campbell, K. E., & Spackman, K. A. (1998). Reproducibility of interpreting“ and” and“ or” in terminology systems. In *Proceedings of the AMIA Symposium* (p. 790). American Medical Informatics Association.

Minsky, M. (1974). A framework for representing knowledge.

Minsky, Marvin. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The Psychology of Computer Vision* (pp. 211–277). McGraw-Hill.

Möller, K., Hausenblas, M., Cyganiak, R., Grimnes, G., & Handschuh, S. (2010). Learning from Linked Open Data Usage: Patterns & Metrics. In *WebSci10: Extending the Frontiers of Society On-Line*. Raleigh, NC, US.

Motik, B., Cuenca-Grau, B., Horrocks, I., Wu, Z., Fokoue, A., & Lutz, C. (2012). OWL 2 Web Ontology Language Profiles (Second Edition). W3C. Retrieved from https://www.w3.org/TR/owl2-profiles/

Motik, P., Patel-Schneider, P., & Parsia, B. (2012). *OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition)*.

Nardi, D., & Brachman, R. J. (2003). An introduction to description logics. *The Description Logic Handbook: Theory, Implementation, and Applications*, 1–40.

Newstead, S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language*, *34*(5), 644.

Nguyen, Power, Piwek, & Williams. (2012). Measuring the understandability of deduction rules for OWL. Presented at the First international workshop on debugging ontologies and ontology mappings, Galway, Ireland.

Nguyen, T. A. T., Power, R., Piwek, P., & Williams, S. (2013). Predicting the understandability of OWL inferences. Retrieved from http://oro.open.ac.uk/36692/

Nickles, M. (2014). Functional-Logic Programming for Web Knowledge Representation, Sharing and Querying. In *Knowledge Engineering and Knowledge Management* (pp. 333–338). Springer.

North, C. (2006). Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, *26*(3), 6–9.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Ogbuji, C. (2008). InfixOWL: An Idiomatic Interface for OWL. In *OWLED* (Vol. 432).

Osborne, J. (2005). Notes on the use of data transformations. *Practical Assessment, Research and Evaluation*, *9*(1), 42–50.

Osherson, D. (1975). Logic and models of logical thinking. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 81–91).

Partee, B., & Rooth, M. (1983). Generalized conjunction and type ambiguity. *Formal Semantics: The Essential Readings*, 334–356.

Patel-Schneider, P. F., Hayes, P., Horrocks, I., & others. (2004). OWL web ontology language semantics and abstract syntax. *W3C Recommendation*, *10*.

Peirce, C. S., & Sowa, J. F. (2000). Existential Graphs: MS 514 by Charles Sanders Peirce with commentary by John Sowa,. Retrieved from http://www.jfsowa.com/peirce/ms514.htm

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14.

Petre, M., & Blackwell, A. F. (1999). Mental imagery in program design and visual programming. *International Journal of Human-Computer Studies*, *51*(1), 7–30.

Pinker, S. (2014). *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*. Penguin.

Poveda-Villalón, M., Suárez-Figueroa, M. C., & Gómez-Pérez, A. (2012). Validating ontologies with OOPS! In *Knowledge Engineering and Knowledge Management* (pp. 267–281). Springer.

Power, R. (2010). Complexity assumptions in ontology verbalisation. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 132–136).

Power, R., & Third, A. (2010). Expressing OWL axioms by English sentences: dubious in theory, feasible in practice. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 1006–1013).

Purchase, H. C. (2012). *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press.

R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*. ISBN 3-900051-07-0.

Ragni, M., Eichhorn, C., & Kern-Isberner, G. (2016). Simulating Human Inferences in the Light of New Information: A Formal Analysis. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI*.

Ramakrishnan, S., & Vijayan, A. (2014). A study on development of cognitive support features in recent ontology visualization tools. *Artificial Intelligence Review*, *41*(4), 595–623.

Rector, A. (2005). Representing Specified Values in OWL: "value partitions" and "value sets." Retrieved from http://www.w3.org/TR/swbp-specified-values/

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., … Wroe, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Engineering Knowledge in the Age of the Semantic Web* (pp. 63–81). Springer.

Rector, A. L. (2003). Defaults, context, and knowledge: Alternatives for OWL-indexed knowledge bases. In *Pacific Symposium on Biocomputing* (pp. 226–237).

Rico, M., Unger, C., & Cimiano, P. (2015). Sorry, I only speak natural language: a pattern-based, data-driven and guided approach to mapping natural language to SPARQL. In *Intelligent Exploration of Semantic Data (IESD) 2015*.

Rips, L. J. (2001b). Reasoning imperialism. *Common Sense, Reasoning, and Rationality*, 215–235.

Rips, L. J. (2001a). Two kinds of reasoning. *Psychological Science*, *12*(2), 129–134.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*(1), 38.

Ritchie, D. G. (1896). The Relation of Logic to Psychology. *The Philosophical Review*, *5*(6), 585–600.

Sánchez, J., Cañas, A. J., & Novak, J. D. (2010). The universality and ubiquitousness of concept maps.

Sato, Y., & Mineshima, K. (2016). Human Reasoning with Proportional Quantifiers and Its Support by Diagrams. In *International Conference on Theory and Application of Diagrams* (pp. 123–138). Springer.

Scheuermann, A., Motta, E., Mulholland, P., Gangemi, A., & Presutti, V. (2013). An empirical perspective on representing time. In *Proceedings of the seventh international conference on Knowledge capture* (pp. 89–96). ACM.

Schwitter, R., Kaljurand, K., Cregan, A., Dolbear, C., & Hart, G. (2008). A comparison of three controlled natural languages for OWL 1.1. In *4th OWL Experiences and Directions Workshop (OWLED 2008 DC), Washington*.

Scott, D. (2012). Tukey's ladder of powers. *Rice University*. Retrieved from http://onlinestatbook.com/2/transformations/tukey.html

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584.

Shams, Z., Jamnik, M., Stapleton, G., & Sato, Y. (2017). Reasoning with Concept Diagrams About Antipatterns in Ontologies. In *International Conference on Intelligent Computer Mathematics* (pp. 255–271). Springer.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 591–611.

Shaw, R. G., & Mitchell-Olds, T. (1993). ANOVA for unbalanced data: an overview. *Ecology*, *74*(6), 1638–1645.

Shneiderman, B. (1978). Improving the human factors aspect of database interactions. *ACM Transactions on Database Systems (TODS)*, *3*(4), 417–439.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3.

Smart, P. R. (2008). Controlled natural languages and the semantic web. Retrieved from http://eprints.soton.ac.uk/265735/

Sowa, J. F. (2000). *Knowledge representation: logical, philosophical, and computational foundations*. Brooks/Cole.

Stapleton, G., Howse, J., Bonnington, A., & Burton, J. (2014). A vision for diagrammatic ontology engineering. Retrieved from http://eprints.brighton.ac.uk/13046/

Stapleton, G., Howse, J., Taylor, K., Delaney, A., Burton, J., & Chapman, P. (2013). Towards Diagrammatic Ontology Patterns. Presented at the 4th Workshop on Ontology and Semantic Web Patterns, Sydney, Australia. Retrieved from http://ontologydesignpatterns.org/wiki/WOP:2013

Stenning, K., & Van Lambalgen, M. (2008). *Human reasoning and cognitive science*. MIT Press.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, *34*, 109–159.

Storey, M. A., Musen, M., Silva, J., Best, C., Ernst, N., Fergerson, R., & Noy, N. (2001). Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé.

Störring, G. (1908). Experimentelle Untersuchungen über einfache Schlussprozesse. *Archiv Für Die Gesamte Psychologie*, *11*, 1–27.

Strawson, P. F. (1952). *Introduction to logical theory*. Routledge.

Tempich, C., & Volz, R. (2003). Towards a benchmark for Semantic Web reasoners-an analysis of the DAML ontology library. In *EON* (Vol. 87).

Thomas, J. C., & Gould, J. D. (1975). A Psychological Study of Query by Example. In *Proceedings of the May 19-22, 1975, National Computer Conference and Exposition* (pp. 439–445). New York, NY, USA: ACM. https://doi.org/10.1145/1499949.1500035

TopQuadrant. (2014). *TopBraid Composer - Getting started guide version 4.5*. Retrieved from http://www.topquadrant.com/docs/TBC-Getting-Started-Guide40.pdf

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 99–114.

van Schaik, P., & Weston, M. (2016). Magnitude-based inference and its application in user research. *International Journal of Human-Computer Studies*, *88*, 38–50.

Vigo, M, Jay, C., & Stevens, R. (2014a). Design insights for the next wave ontology authoring tools. Presented at the ACM SIGCHI Conference on Human Factors in Computing Systems, CHI 2014, Toronto, Canada: ACM Press.

Vigo, M, Jay, C., & Stevens, R. (2014b). Protege4US: harvesting ontology authoring data with Protege. Presented at the HSWI2014 - Human Semantic Web Interaction Workshop, Crete.

Vigo, Markel, Jay, C., & Stevens, R. (2015). Constructing conceptual knowledge artefacts: activity patterns in the ontology authoring process. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3385–3394). ACM.

W3C. (2001). Web Ontology Language (OWL). Retrieved from http://www.w3.org/2001/sw/wiki/OWL

Warren, P. (2013). *Ontology Users' Survey - Summary of Results* (KMi Tech Report No. kmi-13-01). Retrieved from http://kmi.open.ac.uk/publications/pdf/kmi-13-01.pdf

Warren, P. (2014). *Ontology patterns: a survey into their use* (KMi Tech Report No. kmi-14-02). Retrieved from http://kmi.open.ac.uk/publications/pdf/kmi-14-02.pdf

Warren, P., Mulholland, P, Collins, T, & Motta, E. (2014). The usability of Description Logics: understanding the cognitive difficulties presented by Description Logics (pp. 550–564). Presented at the ESWC 2014, Crete: Springer.

Warren, P., Mulholland, P., Collins, T., & Motta, E. (2014). Using ontologies. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 579–590). Springer.

Warren, P., Mulholland, P., Collins, T., & Motta, E. (2015). Making sense of description logics. In *Proceedings of the 11th International Conference on Semantic Systems* (pp. 49–56). ACM.

Warren, P., Mulholland, P., Collins, T., & Motta, E. (2017). Improving the Comprehensibility of Description Logics - Applying insights from theories of reasoning and language. Presented at the ESWC 2017, Portoroz, Slovenia: Springer.

Wason, P. C. (1959). The processing of positive and negative information. *Quarterly Journal of Experimental Psychology*, *11*(2), 92–107.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology*, *23*(1), 63–71.

Wilkins, M. C. (1929). The effect of changed material on ability to do formal syllogistic reasoning. *Archives of Psychology*. Retrieved from http://psycnet.apa.org/psycinfo/1929-04403-001

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul.

Yamauchi, T. (2007). The Semantic Web and Human Inference: A Lesson from Cognitive Science. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, … P. Cudré-Mauroux (Eds.), *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings* (pp. 609–622). Berlin, Heidelberg: Springer Berlin Heidelberg.

Zielinski, T. A., Goodwin, G. P., & Halford, G. S. (2010). Complexity of categorical syllogisms: An integration of two metrics. *European Journal of Cognitive Psychology*, *22*(3), 391–421.

Appendix A: study 1 - reasoning steps for valid questions

This appendix describes the reasoning steps for each of the valid questions in study 1. Where a derivation requires more than one step, numbers in square brackets are used to indicate the provenance of statements derived from previous steps. All other statements are either from the pattern or the specific question.

A1 Componency pattern

Class Object	SubClassOf has_component only Object SubClassOf is_component_of only Object
Property has_part	Characteristics Transitive
Property is_part_of	Characteristics Transitive
InverseOf	has_part
Property has_component	SubPropertyOf has_part
Property is_component_of	SubPropertyOf is_part_of InverseOf has_component

Questions are numbered as in Table 6.7.

A1.1 Question 2

A is_part_of B
B is_part_of C
⇒ A is_part_of C

Step no.	Reasoning
1	is_part_of Characteristics Transitive; A is_part_of B; B is_part_of C ⇒ A is_part_of C

A1.2 Question 3

B is_part_of C
A is_part_of B
⇒ A is_part_of C

Step no.	Reasoning
1	is_part_of Characteristics Transitive; B is_part_of C; A is_part_of B ⇒ A is_part_of C

A1.3 Question 5

A has_component B
B has_component C
⇒ A has_part C

Step no	Reasoning
1	A has_component B; has_component SubPropertyOf has_part ⇒ A has_part B
2	B has_component C; has_component SubPropertyOf has_part ⇒ B has_part C
3	has_part Characteristics Transitive; A has_part B [1]; B has_part C [2] ⇒ A has_part C

A1.4 Question 7

A has_component B
C is_part_of B
⇒ A has_part C

Step no	Reasoning
1	A has_component B; has_component SubPropertyOf has_part ⇒ A has_part B
2	C is_part_of B; is_part_of InverseOf has_part ⇒ B has_part C
3	has_part Characteristics Transitive; A has_part B [1]; B has_part C [2] ⇒ A has_part C

A1.5 Question 8

A Type Object
A has_component B
C Type not Object
⇒ B DifferentFrom C

Step no	Reasoning
1	Class Object SubClassOf has_component only Object A Type Object; A has_component B ⇒ B Type Object
2	B Type Object [1]; C Type not Object ⇒ B DifferentFrom C

A1.6 Question 10

A has_component B
C is_component_of B
⇒ C is_part_of A

Step no	Reasoning
1	has_component SubPropertyOf has_part; A has_component B ⇒ A has_part B
2	is_part_of InverseOf has_part; A has_part B [1] ⇒ B is_part_of A
3	is_component_of SubPropertyOf is_part_of; C is_component_of B ⇒ C is_part_of B
4	is_part_of Characteristics Transitive; B is_part_of A [2]; C is_part_of B [3] ⇒ C is_part_of A

A2 Modified coparticipation pattern

Class Event	EquivalentTo has_participant some Object DisjointWith Object
Class Object	DisjointWith Event
Class Player	SubClassOf Object
Class Game	SubClassOf has_participant some Player
Property coparticipates_with	Domain Object, Range Object Characteristics Symmetric, Transitive
Property has_participant	Domain Event, Range Object InverseOf is_participant_in

Questions are numbered as in Table 6.9.

A2.1 Question 1

A coparticipates_with B
⇒ A Type not Event

Step no	Reasoning
1	A coparticipates_with B; coparticipates_with Domain Object ⇒ A Type Object
2	A Type Object [1]; Object DisjointWith Event ⇒ A Type not Event

A2.2 Question 4

A has_participant B
C is_participant_in D
⇒ B DifferentFrom D

Step no	Reasoning
1	A has_participant B; has_participant Range Object ⇒ B Type Object
2	C is_participant_in D; is_participant_in InverseOf has_participant ⇒ D has_participant C
3	D has_participant C [2]; has_participant Domain Event ⇒ D Type Event
4	B Type Object [1]; D Type Event [3]; Object DisjointWith Event ⇒ B DifferentFrom D

A2.3 Question 5

B coparticipates_with A
B coparticipates_with C
⇒ C coparticipates_with A

Step no	Reasoning
1	B coparticipates_with C; coparticipates_with Characteristics Symmetric ⇒ C coparticipates_with B
2	coparticipates_with Characteristics Transitive C coparticipates_with B [1] B coparticipates_with A ⇒ C coparticipates_with A

A2.4 Question 6

A Type Game
⇒ A Type Event

Step no	Reasoning
1	A Type Game; Game SubClassOf has_participant some Player ⇒ A Type has_participant some Player
2	A Type has_participant some Player [1]; Player SubClassOf Object ⇒ A Type has_participant some Object
3	A Type has_participant some Object [2] Event EquivalentTo has_participant some Object ⇒ A Type Event

A3 Modified types of entities pattern

Class Entity	EquivalentTo Event or Abstract or Quality or Object
Class Event	SubClassOf Entity
	DisjointWith Abstract, Quality, Object
Class Abstract	SubClassOf Entity
	DisjointWith Event, Quality, Object
Class Quality	SubClassOf Entity
	DisjointWith Event, Abstract, Object
Class Object	SubClassOf Entity
	DisjointWith Event, Abstract, Quality
Class Nonconceptual	EquivalentTo Event or Object
Class Nontemporal	EquivalentTo Abstract or Quality or Object
Property represents	Characteristic Functional

Questions are numbered as in Table 6.11.

A3.1 Question 3

A represents B
C represents D
B Type Object
D Type Event
A DifferentFrom C

Step no	Reasoning
1	B Type Object; D Type Event; Object DisjointWith Event ⇒ B DifferentFrom D
2	represents Characteristic Functional; B DifferentFrom D [1] A represents B; C represents D ⇒ A DifferentFrom C

A3.2 Question 4

A Type Entity
A Type not (Event or Quality)
⇒ A Type Abstract or Object

Step no	Reasoning
1	A Type Entity Entity EquivalentTo Event or Abstract or Quality or Object ⇒ A Type (Event or Abstract or Quality or Object)
2	A Type not (Event or Quality) A Type (Event or Abstract or Quality or Object) [1] ⇒ A Type (Abstract or Quality)

N.B. The treatment above reaches the conclusion in step 2 by removing *Event* and *Quality*, in the first line of step 2, from *Event or Abstract or Quality or Object*, in the second line of step 2. A more formal, algebraic, approach would be to expand *not (Event or Quality)* to

become *not Event and not Quality* and then form the conjunction of *not Event and not Quality* and *Event or Abstract or Quality or Object*. This algebraic approach would take several more steps.

A3.3 Question 5

A Type (Nonconceptual and Nontemporal)
⇒ A Type Object

Step no	Reasoning
1	A Type (Nonceptual and Nontemporal) Nonconceptual EquivalentTo Event or Object ⇒ A Type ((Event or Object) and Nontemporal)
2	A Type ((Event or Object) and Nontemporal) [1] Nontemporal EquivalentTo Abstract or Quality or Object ⇒ A Type ((Event or Object) and (Abstract or Quality or Object)
3	Event DisjointWith Abstract, Quality, Object Object DisjointWith Event, Abstract, Quality A Type ((Event or Object) and (Abstract or Quality or Object) [2] ⇒ A Type Object

The final step is dependent on *Event* being disjoint with *Abstract*, *Quality* and *Object*, and on *Object* being disjoint with *Abstract* and *Quality*.

Appendix B: study 2 – reasoning steps for Boolean operator questions

This appendix outlines the reasoning steps for each of the questions with valid conclusions in the section of study 2 relating to Boolean operators, see Section 7.4. The steps described here are intended to represent how participants might tackle the questions, bearing in mind that they are working without pen and paper. They do not correspond exactly to the manipulations which would be made by someone trained in Boolean logic and working with pen and paper. Moreover, they do not necessarily represent a unique way of arriving at the correct result. They do, however, illustrate the differing complexities of the questions.

In what follows, the symbol $\underline{\cup}$ (i.e. the symbol for union, underlined) is used to represent disjoint union. More usually, disjoint union is represented by $\cup$ with a dot over it or within it; the convention used here is for typographical convenience. The backslash (\) is used to denote relative complement, i.e. terms on the right-hand side of the backslash are removed from those on the left.

B1 Question 1

In question 1 we have:

TOP_CLASS	$\equiv A \underline{\cup} B \underline{\cup} C$	(1.1)
$A \underline{\cup} B \underline{\cup} C \setminus A$	$\equiv B \underline{\cup} C$	(1.2)
$B \underline{\cup} C \setminus B$	$\equiv C$	(1.3)

Thus, there are three steps. The first is simply the expansion of TOP_CLASS. The second is the most complex, requiring the manipulation of three classes and being of RC 3. The third is of RC 2.

B2 Question 2

This question can be approached very similarly to question 1. First, expand TOP_CLASS:

TOP_CLASS	$\equiv A \underline{\cup} B \underline{\cup} C$	(2.1)

Then remove the classes within the brackets:

$A \underline{\cup} B \underline{\cup} C \setminus (A \underline{\cup} B)$	$\equiv A$	(2.2)

Step 2.2 is of RC 3. Another way of representing the reasoning process would see step 2.2 broken down into two steps, identical to steps 1.2 and 1.3 above; and still with a maximum RC of 3.

B3 Question 3

In question 3 we need to start by expanding the inner bracketed expression in the first statement, which in turn first requires the expansion of A. After this A_1 needs to be removed:

A	$\equiv A\_1 \underline{\cup} A\_2$	(3.1)
$A\_1 \underline{\cup} A\_2 \setminus A\_1$	$\equiv A\_2$	(3.2)

TOP_CLASS then needs to be expanded, which is a two-stage process:

TOP_CLASS	≡ A ⋃ B	(3.3)
A ⋃ B	≡ A_1 ⋃ A_2 ⋃ B	(3.4)

Then finally, A_2 (from 3.2) needs to be removed from the expression in 3.4:

A_1 ⋃ A_2 ⋃ B \ A_2	≡ A_1 ⋃ B	(3.5)

In all, this takes five steps and the maximum RC of 3 occurs at 3.5.

B4 Question 5

One approach here would be to start by expanding TOP_CLASS:

TOP_CLASS	≡ A ⋃ B	(5.1)

Then, take account of the *not (A …* :

A ⋃ B \ A	≡ B	(5.2)

Finally, re-inserting the A_1 gives A_1 ⋃ B

This is simpler than for question 3, with a maximum RC of 2 at step 5.2.

B5 Question 7

The first step could be an expansion of TOP_CLASS:

TOP_CLASS	≡ A ⋃ B	(7.1)

There are then two parallel paths corresponding to the two bracketed expressions containing TOP_CLASS. The first bracketed expression gives:

A ⋃ B \ A	≡ B	(7.2)

The second bracketed expression starts with the expansion of A and then requires the intersection with A_1:

A ⋃ B	≡ A_1 ⋃ A_2 ⋃ B	(7.3)
(A_1 ⋃ A_2 ⋃ B) ⋂ A_1	≡ A_1	(7.4)

We then need to take the disjunction of these two bracketed expressions, as in the right-hand sides of 7.2 and 7.4, to give B ⋃ A_1. The maximum RC 3 at 7.4, where three classes need to be manipulated.

B6 Question 9

Question 9 is more complex. One strategy is to start by expanding the innermost bracket:

A_1	≡ A_1_X ⋃ A_1_Y	(9.1)
(A_1_X ⋃ A_1_Y) \ A_1_X	≡ A_1_Y	(9.2)

Working outwards to the next bracket, expanding A in two stages and then removing A_1_Y:

A	≡ A_1 ⋃ A_2	(9.3)
A_1 ⋃ A_2	≡ A_1_X ⋃ A_1_Y ⋃ A_2	(9.4)

$$A\_1\_X \underline{\cup} A\_1\_Y \underline{\cup} A\_2 \setminus A\_1\_Y \equiv A\_1\_X \underline{\cup} A\_2 \quad (9.5)$$

Then, TOP_CLASS needs to be expanded as far as is possible:

$$TOP\_CLASS \equiv A \underline{\cup} B \quad (9.6)$$

$$A \underline{\cup} B \equiv A\_1 \underline{\cup} A\_2 \underline{\cup} B \quad (9.7)$$

$$A\_1 \underline{\cup} A\_2 \cup B \equiv A\_1\_X \underline{\cup} A\_1\_Y \underline{\cup} A\_2 \cup B \quad (9.8)$$

Finally, the relative complement is then formed from the expanded TOP_CLASS in 9.8 and the expression in 9.5:

$$A\_1\_X \underline{\cup} A\_1\_Y \underline{\cup} A\_2 \cup B \setminus A\_1\_X \cup A\_2 \equiv A\_1\_Y \underline{\cup} B \quad (9.9)$$

Not only does this process involve more steps than the previous questions, but the final step (9.9) involves the manipulation of four classes and is of RC 4.

B7 Summary

Questions 1, 2, 3 and 7 have maximum RC 3. However, questions 3 and 7 do have more steps than questions 1 and 2. This is consistent with the greater syntactic complexity of questions 3 and 7, and also their greater complexity when represented as mental models. Question 5, although semantically equivalent to questions 3 and 7, requires fewer steps and has maximum RC 2. This comparative simplicity is not reflected in the accuracy of responses, consistent with the view that the complexity of the mental model is the major determinant of difficulty. Finally, question 9 is more complex both in the number of steps required and in having maximum RC 4. This additional complexity is reflected in reduced accuracy and increased response time; although only the latter effect was significant, see section 7.4.5.

Appendix C: effect of sample size on response time comparison

The analysis in this appendix is concerned with establishing an estimate for the number of samples required to achieve significance in a comparison of two groups, where the measurement is of a variate assumed to have a log-normal distribution with parameters μ and σ. These are the mean and standard deviation of the underlying normal distribution, i.e. of the variate after a logarithmic transformation.

The other assumptions made are:

- The sample mean and standard deviation of the log-normal distribution correspond exactly to the mean and standard deviation of the underlying distributions. These are represented by μ_L and σ_L.
- The ratio of standard deviation to mean (σ_L / μ_L) is assumed to be 0.75 for both groups.
- The number of samples in each group is the same. This number is represented by n, i.e. there are 2n samples in all.

C1 Parameters of the log-normal distribution

The mean and standard deviation of a log-normal distribution can be represented in terms of the parameters μ and σ by:

$$\mu_L = \exp(\mu + \sigma^2/2) \quad (1)$$

$$\sigma_L^2 = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$$

$$= \exp(2\mu)\exp(\sigma^2)(\exp(\sigma^2) - 1) \quad (2)$$

From (1),

$$\exp(\mu) = \mu_L \exp(-\sigma^2/2) \quad (3)$$

Substituting for $\exp(\mu)$ from (3) into (2) gives:

$$\sigma_L^2 = \mu_L^2 (\exp(\sigma^2) - 1) \quad (4)$$

This leads to:

$$\sigma^2 = \ln(1 + (\sigma_L / \mu_L)^2) \quad (5)$$

Taking logarithms on each side of (3):

$$\mu = \ln(\mu_L) - \sigma^2/2 \quad (6)$$

Substituting from (5) into (6):

$$\mu = \ln(\mu_L) - (\ln(1 + (\sigma_L / \mu_L)^2))/2$$

$$= \ln(\mu_L / \sqrt{(1 + (\sigma_L / \mu_L)^2)}) \quad (7)$$

Equations (5) and (7) give the underlying parameters of the log-normal distribution in terms of the distribution's mean and standard deviation.

C2 t-statistic

In Chapters 6, 7 and 8, the t-statistic was calculated after taking the logarithmic transformation. That is to say, the t-statistic was not calculated directly from the means and standard deviations of the response time distributions, but from the means and standard deviations of the distributions of the log of response time. These correspond to μ and σ in the previous section, and in what follows are written μ_1, σ_1 and μ_2, σ_2 for the two groups. On the assumption that the two groups have the same number of samples, n, then the t-statistic, with 2n-2 degrees of freedom, is given by:

$$t(2n-2) = (\mu_1 - \mu_2)\ \sqrt{n} / (\sqrt{(\sigma_1^2 + \sigma_2^2)}) \qquad (8)$$

Writing μ_{L1} and μ_{L2} for the means of response times for the two groups, again assuming that the sample means correspond to the population means, equation (6) gives:

$$\mu_1 = \ln(\mu_{L1}) - \sigma_1^2/2 \quad \text{and} \quad \mu_2 = \ln(\mu_{L2}) - \sigma_2^2/2 \qquad (9)$$

If we assume that the ratio of standard deviation to mean is the same for both groups, represented by r, then using equation (5) above:

$$\sigma_1^2 = \sigma_2^2 = \ln(1 + r^2) \qquad (10)$$

Using equations (9) and (10) to substitute for μ_1, μ_2, σ_1 and σ_2 in (8):

$$t(2n-2) = (\ln(\mu_{L1}) - \ln(\mu_{L2}))\ \sqrt{n} / \sqrt{(2\ \ln(1+r^2))} \qquad (11)$$

To investigate relative differences, write $\mu_{L1} = a\ \mu_{L2,}$ so that (11) becomes:

$$t(2n-2) = \ln(a)\ \sqrt{n} / \sqrt{(2\ \ln(1+r^2))} \qquad (12)$$

The t-statistic is now a function only of a, the ratio between the two means, and r, the ratio of standard deviation to mean, which is assumed to be the same for both groups. As discussed in Chapter 9, for the purposes of the example, this is taken to be 0.75.

On this basis, for any given value of a, the t-statistic can be calculated for a range of values of n so as to obtain, by search, the minimum value of n to achieve significance at the 0.05 level. Table C-1 shows these minimum values. This table is reproduced as Table 9-4 in subsection 9.5.2, where an alternative labelling in terms of percentage difference is used.

Table C-1 Minimum group size to achieve significance

a	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
min. no. samples	379	105	52	32	23	17	14	12	10	9

The table represents the minimum number of data points in each group. As noted in Chapter 9, this is based on a two-sided t-test; a one-sided test would require fewer points to achieve significance. For a comparison of one question between participants, then the required number of participants would be twice the number shown; for a within-participants study, then the total number of participants would be as shown. The number of participants required could be further reduced if more than one question was use for comparison.