Making Sense of Description Logics

Paul Warren paul.warren@cantab.net
Paul Mulholland paul.mulholland@open.ac.uk
Trevor Collins trevor.collins@open.ac.uk
Enrico Motta enrico.motta@open.ac.uk
Knowledge Media Institute, The Open University, Milton Keynes, Buckinghamshire, MK7 6AA, U.K.

ABSTRACT

Description Logics are commonly used for the development of ontologies. Yet they are well-known to present difficulties of comprehension, e.g. when confronted with the justification for a particular entailment during the debugging process. This paper describes a study into the problems experienced in understanding and reasoning with Description Logics. In particular the study looked at: functionality in object properties; negation, disjunction and conjunction in Propositional Logic; negation and quantification; and the combination of two quantifiers. The difficulties experienced are related to theories of reasoning developed by cognitive psychologists, specifically the mental model and relational complexity theories. The study confirmed that problems are experienced with functional object properties and investigated the extent to which these difficulties can be explained by relational complexity theory. Mental model theory was used to explain performance with negation and quantifiers. This suggests that Boolean logic is easier to assimilate in Disjunctive Normal Form than in other forms and that particular difficulties arise when it is necessary to backtrack to form a mental model. On the other hand in certain cases syntactic clues seemed to contribute to reasoning strategies.

Categories and Subject Descriptors

F4.m [**Mathematical Logic and Formal Languages**]: Miscellaneous; H1.2 [**Models and Principles**]: User/machine systems – *human information processing, software psychology*; I2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *predicate logic, representation languages*

General Terms

Experimentation, Human Factors, Languages.

Keywords

Description Logics, Psychology of Reasoning.

1. INTRODUCTION

The problems experienced by non-logicians in understanding Description Logic (DL) statements is well-known, e.g. see [11]. However, little has been done to study these difficulties systematically. As a result there is little theoretical understanding of these difficulties and hence little guidance to overcome them. This paper, building on previous work [17], relates these difficulties to psychological theories of reasoning. Relating specific problems to general theories enables the creation of generalized guidelines which go beyond the particular examples discussed.

The previous work looked in particular at the difficulties arising with negated conjunction; with functional object properties; and with quantification, in that case existential quantification. The work reported here investigates these issues in more depth, besides also considering more complex situations involving quantification.

The next section describes related work, both by computer scientists and cognitive psychologists investigating human reasoning. Section 3 then describes the study. Sections 4 to 7 present and discuss the results of the study's four question sections. Finally, section 8 presents some conclusions and recommendations, and also discusses future work.

2. RELATED WORK

2.1 Comprehensibility of Description Logics

There has been a small amount of research looking at the comprehensibility of DLs. Rector et al. [11] describe the difficulties experienced by newcomers to OWL, based on their experience of teaching the language, and provided a set of guidelines. Horridge et al. [3] were concerned with supporting the ontology debugging process by presenting developers with the justification for a given entailment. They developed and validated a complexity model for justifications. Nguyen et al. [9] were concerned with selecting deduction rules to illustrate why an entailment follows from a justification. Their goal was to construct proof trees in English. They sought to establish the understandability of individual deduction rules in order to create the most comprehensible out of a number of possible alternative trees. In addition, some work has also looked at the potential to enhance DL comprehension with visualization, e.g. see Stapleton et al. [16]. This should be differentiated from techniques for visualizing all or part of the structure of an ontology, e.g. see [7]. None of this work has been informed by psychological theory. There has, however, been considerable research into human reasoning, which could be applied. This is discussed in the next sub-section.

2.2 Theories of human reasoning

Psychologists investigating human reasoning have developed a number of theoretical standpoints. One approach assumes that naïve reasoners i.e. people not trained in logic, execute rules similar to those carried out by a logician, e.g. see Rips [13]. By contrast, the model-based approach assumes that mental models are constructed to represent a given situation, e.g. see Johnson-Laird [5]. Table 1 illustrates how exclusive and inclusive disjunction would be interpreted as a mental model. The former has two

mental models, the latter three. Mental model theory suggests that inclusive disjunction will give rise to more reasoning errors than exclusive disjunction since in the former case the final disjunct (John is chairman and Tony is secretary) is often ignored. This is confirmed by experiment, e.g. see Johnson-Laird et al. [6]. It should be noted that for Propositional Logic this approach corresponds to representation in Disjunctive Normal Form (DNF), where each of the lines in a mental model is a disjunct [6].

Table 1 Illustrating mental model theory

	mental model
Exclusive disjunction *Either John is chairman or Tony is secretary*	John chairman Tony secretary
Inclusive disjunction *John is chairman or Tony is secretary or both*	John chairman Tony secretary John chairman Tony secretary

Sloman [15] has argued that both rule-based and model-based reasoning are present in normal reasoning. He suggests that "awareness provides … a fallible heuristic" for distinguishing between rule-based and model-based reasoning, with the former more likely to be associated with conscious reasoning and the latter to be associated with unconscious reasoning. It seems likely that people working with DLs will make more use of a rule-based approach than do naïve reasoners. Nevertheless, the former are also likely at times to make use of mental models, whether consciously or subconsciously.

Relational complexity (RC) theory represents a complementary approach. Here complexity is defined "as a function … of the number of variables that can be related in a single cognitive representation", see Halford and Andrews [2]. What is important in any reasoning step is the number of variables which need to be simultaneously manipulated. Proponents of the theory suggest that the accuracy of a chain of reasoning steps is dependent on the maximum RC of the individual steps.

As an example, Halford et al. [1] note that reasoning with transitivity has RC 3. A transitive relation, e.g. 'greater than', is binary since it relates two individuals. However, integrating two instantiations of a transitive relation in a deductive step requires concurrent attention to three individuals. This is exemplified by the following deductive step, where T is any transitive relation:

$$a\ T\ b;\ b\ T\ c \Rightarrow a\ T\ c$$

3. THE STUDY

The study consisted of 34 questions divided between the four sections. Each question consisted of a set of DL statements and a proposed inference, all written in a simplified form of Manchester OWL Syntax, see Horridge et al. [4]. The participant was required to indicate whether or not the proposed inference was valid. The study was created using the web survey tool, *SurveyExpression* (http://www.surveyexpression.com). Timing information was taken from screen recordings made using *Camtasia* (http://www.techsmith.com/camtasia.html).

Each permutation of the order of the four sections was used once, requiring 24 participants. For each section, half the participants saw the questions in one order, half in the reverse order. This was to compensate for the possibility that early questions in a section would be answered less accurately and more slowly than later ones. Neither of these two orders corresponds to the systematic order in which questions are described in sections 4 to 7 of this paper. Because of technical and experimenter error, there were an additional four participants for whom timing data was not available. For this reason accuracy data is based on a sample of 28, and timing data on a sample of 24.

Note that timing data is analysed using the t-test or ANOVA. These tests depend on approximately normal distributions of time data for each question. In fact, visualization revealed a positive skew for most questions. Further analysis suggested that the logarithmic transformation of time, selected from Tukey's ladder of powers [14], resulted in a distribution closer to the normal, and this transformation has been applied in all such tests reported in the paper. All statistical analysis was undertaken using the R statistical package [10].

Some of the studies were undertaken face-to-face, others making use of *Skype* (http://www.skype.com). At the face-to-face sessions participants were initially given a five page A4 handout which provided all the required information about the DL constructs and the particular syntax used. For the remote sessions, this handout was emailed beforehand. In both cases the handout was available for reference whilst answering the questions. The first section of the study gave some general information about the study and invited participants to provide information about their knowledge of logic, their knowledge and usage of OWL or other DLs, and their relationship to the English language. The next four sections were the question sections, in each case with an introductory page. Finally, there was a section which invited participants to provide feedback. Participants were required to complete the study without the use of pen and paper.

There was one occasion during a Skype session when transmission problems meant that the transition between two questions, in the section on Propositional Logic, was not available. In this case the total time for the two questions was apportioned according to the ratio of the mean times for these two questions for the other eleven participants for whom the section was presented in the same order. Apart from this instance, all the other data was complete.

Participants were from the authors' own and other universities, industrial laboratories and a research institute, all with a background in computer science or a related discipline. From the sample of 28, only 2 of the participants claimed no knowledge of logic and 3 no knowledge of OWL or another DL. Greater knowledge of logic did not significantly affect performance. Greater prior knowledge of DLs did significantly improve accuracy.

4. FUNCTIONAL OBJECT PROPERTIES

4.1 The questions

In the study reported in Warren et al. [17] only 50% of participants answered correctly a question featuring a functional object property. The question required a reasoning step of the form:

$$a\ F\ b;\ c\ F\ d;\ b\ \text{DifferentFrom}\ d \Rightarrow a\ \text{DifferentFrom}\ c$$

where a, b, c and d are individuals and F is a functional object property. This step has RC four, since it requires the concurrent attention to four individuals. It was not clear from this study to what extent the difficulty was caused by the relational complexity and to what extent it was inherent in the nature of functionality.

Table 2 Valid inference object property questions

N.B. (1) In this and subsequent similar tables 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 this latter is calculated on the basis of a slightly larger sample. (2) NA indicates insufficient data points to calculate a standard deviation.

	axioms	valid inference	RC	%age correct N = 28	mean time (SD), N = 24		
					overall - secs	correct - secs	incorrect - secs
1	a T b; b T c; c SameAs d; d T e	a T e	3, 2, 3	96%	34 (14)	34 (14)	40 (NA)
2	a F b; a F c; b F d; c F e	d sameAs e	3, 2, 3	75%	52 (36)	56 (39)	36 (9)
3	a F b; a F c; d F b; e F f; c DifferentFrom f	d DifferentFrom e	3, 2, 4	61%	84 (67)	90 (66)	73 (70)
4	a F b; c F d; b DifferentFrom d; e F a; f F g; c SameAs g	e DifferentFrom f	4, 2, 4	43%	109 (79)	96 (55)	119 (96)

A simpler reasoning step, also involving functionality, is of the form: a F b; a F c $\Rightarrow$ b SameAs c

This step has RC three, since it requires attention to three individuals. If difficulty were determined entirely by relational complexity, we would not expect to see any difference in performance between this step and the one involving transitivity shown in subsection 2.2. If functionality were inherently harder than transitivity, then the step shown here might be harder.

This was tested using the first two questions in Table 2. For brevity, T and F are used to represent transitive and functional properties respectively. In the study, the transitive property was named *greater_than_or_equal_to* and the functional property was *has_nearest_neighbour*. In the actual questions in this and all other sections, all the necessary declarative statements were included. Note that correct reasoning in both questions depends on three reasoning steps of RC 3, 2, 3. The middle step arises in question 2 because, once the equivalence of b and c has been established, one of these individuals must be substituted for the other in one of the final two expressions. A reasoning step of RC 2 was deliberately included in question 1 to achieve balance.

Question 3 replaces the last reasoning step in question 2 with one using functionality but of RC 4. From relational complexity theory we might expect question 3 to be answered less accurately than question 2, since the maximum RC has been increased.

Question 4 replaces the first reasoning step with one of RC 4. According to relational complexity theory, this should not make a difference to the accuracy of responses compared with question 3, as the maximum RC has not been increased.

There were four other questions in the section, as shown in Table 3. They repeated the axioms from the valid questions, but in each case with a non-valid proposed inference. Thus question 5 used the same axioms as question 1, question 6 the same as question 2, etc. The results of these questions were not used in the analysis discussed subsequently because the proposed inferences could be regarded as a confounding factor, i.e. some of these inferences might be more obviously non-valid than others.

Table 3 Non valid inference object property questions

	non valid inference	%age correct N = 28	mean time (SD), N = 24		
			overall - secs	correct - secs	incorrect - secs
5	d T b	86%	48 (34)	42 (14)	87 (95)
6	a sameAs c	96%	61 (46)	62 (47)	35 (NA)
7	a DifferentFrom d	79%	92 (66)	86 (62)	111 (80)
8	a DifferentFrom f	71%	96 (47)	93 (50)	101 (43)

4.2 Results

The percentage of correct responses for each valid question are shown in Table 2, along with the overall mean response time and the mean times for correct and incorrect responses. Note that questions 3 and 4 were not answered significantly better than chance ($p = 0.172$ and $p = 0.828$ on a one-sided test). Throughout this paper 'significant' is taken to mean at the 95% level.

Overall, comparing the four valid questions, accuracy is significantly dependent on question ($p = 0.000164$ on a χ^2 test). the same is true of time to respond ($p = 5.91 \times 10^{-8}$ using ANOVA).

A χ^2 test suggested that the difference in accuracy between questions 1 and 2 was approaching significance ($p = 0.0562$). Use of the chi-squared approximation in this test may not have been valid because of the small number of datapoints. However, Fisher's Exact Test gave a very similar result ($p = 0.05105$). A t-test revealed that question 1 was answered significantly more quickly than question 2 ($p = 0.0229$).

There was no significant difference in the accuracy of responses between questions 2 and 3 ($p = 0.391$ on a χ^2 test), but there was a significant difference in the time ($p = 0.0457$ on a t-test).

Similarly there was no significant difference in accuracy between questions 3 and 4 ($p = 0.285$ on a χ^2 test), whilst the difference in time was approaching significance ($p = 0.0866$ on a t-test).

4.3 Discussion

The difference in accuracy and time for questions 1 and 2 supports the view that functionality, as exemplified in question 2, is harder than transitivity, as exemplified in question 1. There are a number of possible factors contributing to this.

In the first place, in both the reasoning steps involving transitivity in question 1, the statements are presented in the most straightforward order (e.g. a T b; b T c). It might be that if this order were altered, participants would experience greater difficulty. This straightforward order better reflects the structure of the unified mental model, thereby reducing the cognitive effort required to build the unified mental model from the premises; compare the discussion in Knauff et al. (1998) on the continuity effect in spatial reasoning.

There are two other potential factors, not related to the ordering of the statements. Firstly, the reasoning step using transitivity in subsection 2.2 involves only the property: T. The reasoning steps using functionality in subsection 4.1 involve two properties: F and SameAs. Secondly, there may be confusion in some participants'

minds between functional and inverse functional. There is no analogous problem in the case of transitivity; the inverse of a transitive function is transitive.

Relational complexity theory is not supported by the failure to find a significant decrease in accuracy between question 2 and question 3. However, the introduction of the more complex reasoning step does significantly increase time to respond.

On the other hand, the lack of any significant decrease in accuracy between questions 3 and 4 is consistent with relational complexity theory's hypothesis that the relevant factor is the maximum complexity. That the increase in time was approaching significance further supports the view that the RC 4 reasoning step with functionality does take longer than the RC 3 step.

5. PROPOSITIONAL LOGIC

5.1 The questions

This section contained ten questions. Table 4 shows the six questions for which the proposed inferences were valid. In each case the axioms include an equivalence between the class Z and a Boolean expression. The inference is an equivalence between Z and a simpler Boolean expression. The Boolean expressions in questions 1 and 2 are logically equivalent, as are the Boolean expressions in questions 3, 4 and 5. The motivation for the form of expression in questions 1, 3 and 6 comes from Rector's discussion of the need to deal with exceptions [12]. The three groups of questions ({1,2}, {3,4,5} and {6}) are of increasing complexity. This is apparent from the three mental models, as shown in Table 4, where, for example, a and b represent typical members of A and B. For brevity, TOP_CLASS is not represented in the mental models since it would be present in each disjunct. The questions also display an increasing syntactic complexity, although there are a variety of ways of measuring this. For example, using the degree of nesting within brackets, question 1 is less complex than questions 3, 4, and 5, which in turn are less complex than question 6. On this measure, question 2 has the same complexity as questions 3, 4, and 5. There were three research questions to be investigated. Firstly, whether question 2, requiring an expansion of *not (A or B)* would prove any harder than question 1. The intention was to complement previous work which had shown that *not (A and B)* is harder than *not (A or B)*, e.g. see [8] and [17]. The second research question was whether there would be any difference in performance between the three different but logically equivalent Boolean expressions in questions 3, 4 and 5. As already noted, mental model representation corresponds to expressing Propositional Logic statements in DNF. It might be the case that logical expressions in, or close to, DNF, would be easier to interpret than other logical expressions. Question 5 is in DNF if we include TOP_CLASS and question 4 is in DNF if we take account that TOP_CLASS subsumes all classes, so we might expect these questions to be easier than question 3. The third research question was whether the increasing levels of complexity of mental models would lead to a significant decrease in performance.

The section contained four other questions, each with non-valid proposed inferences, as shown in Table 5. The axioms for questions 7, 8, 9 and 10 were as for questions 3, 4, 5 and 6 respectively. These questions were not used in the analysis because differences in credibility of the proposed non-valid inferences might act as a confounding factor.

5.2 Results

Table 4 shows the percentage of valid questions answered correctly, and the mean times overall and for the correct and incorrect responses. Questions 3, 4, 6 and 7 were not answered significantly better than chance (p = 0.172, 0.092, 0.425, 0.092).

There was no significant difference in accuracy between questions 1 and 2. Because of the small sample size, the Fisher Exact Test was used, giving a p-value not materially different from 1. A t-test also showed no significant difference in times for the two questions (p = 0.798). A χ^2 test revealed that that the variation in Boolean expressions in questions 3, 4 and 5 had no effect on accuracy (p = 0.856). An ANOVA revealed that the effect on time was approaching significance (p = 0.0575). A subsequent Tukey HSD analysis revealed that the difference in time between questions 3 and 4 was approaching significance (p = 0.0658).

Table 4 Valid inference Propositional Logic questions

	axioms	valid inference	Mental model	%age correct N = 28	mean time (SD), N = 24		
					overall - secs	correct - secs	incorrect - secs
1	Z EquivalentTo (TOP_CLASS and not A and not B); TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	¬a ¬b	82%	39 (26)	36 (23)	56 (37)
2	Z EquivalentTo (TOP_CLASS and not (A or B)); TOP_CLASS DisjointUnionOf A, B, C			86%	43 (29)	36 (24)	78 (27)
3	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	¬ a a_1	61%	96 (56)	99 (64)	89 (37)
4	Z EquivalentTo (TOP_CLASS and (not A or A_1)); TOP_CLASS DisjointUnionOf A, B; A DisjointUnionOf A_1, A_2			64%	65 (38)	61 (33)	74 (48)
5	Z EquivalentTo ((TOP_CLASS and not A) or (TOP_CLASS and A_1)); TOP_CLASS DisjointUnionOf A, B; A DisjointUnionOf A_1, A_2			68%	70 (45)	57 (31)	109 (59)
6	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 DisjointUniotOf A_1_X, A_1_Y	Z EquivalentTo B or A_1_Y	¬ a a_1 ¬a_1_x	54%	90 (48)	91 (49)	89 (51)

Table 5 Non valid inference Propositional Logic questions

	non valid inference	%age correct N = 28	mean time (SD), N = 24		
			overall - secs	correct - secs	incorrect - secs
7	Z EquivalentTo B	64%	105 (78)	112 (78)	89 (79)
8	Z EquivalentTo A_1	79%	58 (33)	54 (32)	70 (40)
9	Z EquivalentTo A_2	89%	65 (26)	66 (26)	59 (32)
10	Z EquivalentTo A_1_Y	68%	94 (47)	94 (50)	95 (45)

Complexity levels 1, 2 and 3 were assigned to questions 1, and 2, questions 3, 4, 5 and question 6 respectively. A χ^2 test revealed a significant dependency of accuracy on complexity level ($p = 0.00733$). A logistic analysis of deviance confirmed this ($p = 0.00567$). A subsequent Tukey HSD analysis revealed that the level 1 questions were answered more accurately than the level 2 questions ($p = 0.0346$) and the level 3 questions ($p = 0.0110$), whilst there was no significant difference between the level 2 and the level 3 questions ($p = 0.5701$). An ANOVA of time against level and a subsequent Tukey HSD analysis gave exactly parallel results. There was a significance dependence of time on level ($p = 9.34 \times 10\text{-}9$). The level 1 questions were answered significantly more quickly than the level 2 questions ($p = 0.0000005$) and the level 3 question ($p = 0.0000004$), whilst there was no significant difference in speed of response between the level 2 and level 3 questions ($p = 0.276$).

5.3 Discussion

The lack of any significant difference in accuracy or time between questions 1 and 2 indicates that participants could as equally well interpret negated disjunction, *not (A or B)*, as its expanded form, *not A and not B.*

Similarly, the lack of any significant difference in accuracy between questions 3, 4 and 5 indicates that our participants were able to interpret expressions of the form *not (A and B)* with accuracy not significantly different from when presented with the expanded form, *not A or not B*. However, the results of the ANOVA and Tukey HSD analysis do suggest that the unexpanded form takes longer. For some participants this may be because they are expanding the former expression algebraically, using the appropriate De Morgan's law. For other participants it may be that the construction of the mental model is quicker when the expression is in DNF.

The reduction in accuracy and increase in response time with complexity is to be expected. More interesting is the lack of a significant difference between the level 2 questions and the level 3 question. This may have been a result of the limited datapoints, specifically the fact that there was only one valid question at level 3. However, performance on level 2 questions was already quite close to chance and it may be that these questions were so difficult that increasing complexity made no difference. It is also the case that the level 3 question used the form, motivated by [12], that was least accurately and most slowly answered at level 2.

6. NEGATION AND QUANTIFICATION

6.1 The questions

The aim of this question section was to see what difficulties participants had in understanding the interaction between negation and quantification. After the necessary class and object property declarations, each question contained a statement constraining the class X. There were four variants on this statement, shown in questions 1 to 4 and then repeated in questions 5 to 8 in Table 6. The four statements were created by alternating the use of *some* and *only* and by alternating the position of *not* to be immediately preceding a class (MALE) or immediately preceding a property (*has_child*). A point to note is that the statements used in questions 1 and 5 are logically equivalent to the statements in questions 4 and 8, whilst the statements in questions 2 and 6 are logically equivalent to those in questions 3 and 7. To make this clear, the two sets of four statements are shown in different typefaces. There was then a statement constraining the class Y, with two variants: a variant using *only* in questions 1 to 4 and a variant using *some* in questions 5 to 8. All questions had the same proposed inference: *X Disjoint to Y.* Four of the questions were valid and four non-valid, as shown in Table 6.

6.2 Results

Table 6 shows the percentage of correct responses for each question and the mean times overall for the correct and incorrect responses. Only questions 4 and 7 were answered significantly better than chance ($p = 0.00626$ and $p = 0.00186$). The division of the first statement into two sets of logically equivalent statements, identified by the two typefaces in Table 6, in conjunction with the two variants of the second statement, means that there are four pairs of semantically equivalent questions: {1, 4}, {2, 3}, {5, 8}, and {6, 7}. A χ^2 test revealed that this categorization was not a significant determinant of accuracy ($p = 0.780$) and an ANOVA showed that it was also not a determinant of time to respond ($p = 0.357$). In addition, the form of the second statement does not significantly affect accuracy ($p = 0.889$ on a χ^2 test) or time ($p = 0.761$ on a t-test). Nor did participants perform significantly differently when the same quantifier was used in the statement constraining X and in the statement constraining Y (i.e. questions 2, 4, 5, 7) as when different quantifiers were used (questions 1, 3, 6, 8). A χ^2 test revealed no significant difference in accuracy ($p = 0.486$) whilst a t-test did reveal a significant difference in time to respond ($p = 0.0484$). In fact, the maximum mean response time for questions using the same quantifier (i.e. questions 2, 4, 5 and 7; 43 seconds) is less than the minimum mean time for questions using both quantifiers (i.e. questions 1, 3, 6 and 8; 44 seconds). Accuracy appears more determined by the structure of the first statement. All questions with a first statement in which negation preceded the property (questions 3, 4, 7, 8) were answered more accurately than those questions in which the negation preceded the class (questions 1, 2, 5, 6). This difference was significant ($p = 0.0178$ on a χ^2 test). However, overall the position of the negation did not significantly influence the time to respond ($p = 0.251$ on a t-test). In short, the combination of quantifiers affects the time whilst the position of the negation affects the accuracy.

Table 6 Negation and quantification: all questions have the proposed inference *X DisjointTo Y*

	First statement – constraining X	Second statement – constraining Y	Valid / Non valid	%age correct N = 28	mean time (SD), N = 24		
					overall - secs	correct - secs	incorrect - secs
1	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child only MALE	V	61%	52 (39)	38 (24)	80 (50)
2	X SubClassOf has_child only (not MALE)		N	50%	33 (18)	32 (14)	34 (22)
3	X SubClassOf not (has_child some MALE)		N	68%	45 (22)	43 (24)	49 (16)
4	***X SubClassOf not (has_child only MALE)***		V	75%	43 (25)	40 (24)	57 (24)
5	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child some MALE	N	64%	41 (30)	42 (32)	38 (27)
6	X SubClassOf has_child only (not MALE)		V	50%	44 (40)	38 (25)	52 (55)
7	X SubClassOf not (has_child some MALE)		V	79%	43 (37)	34 (26)	79 (53)
8	***X SubClassOf not (has_child only MALE)***		N	68%	60 (37)	61 (42)	58 (29)

6.3 Discussion

The fact that neither accuracy nor time to respond depended significantly on semantic structure may have been because participants were reasoning syntactically rather than semantically, or because the difficulty differentiating the questions was in translating from the syntax to the corresponding mental models. Indeed, a main thesis of the proponents of the mental model theory is that people make mistakes, usually of omission, in creating mental models. The likelihood of such a mistake can vary depending on the syntactic starting point.

A closer examination of the questions gives a more precise indication of what factors influence accuracy and time, and also suggest what strategies participants might be using. The questions answered most accurately, and in fact the only two to be answered significantly better than chance, were 4 and 7. These questions are uniquely characterized by the fact that the anonymous class negated in the first statement is also the anonymous class used in the second statement, i.e. *has_child only MALE* in question 4 and *has_child some MALE* in question 7. These questions are both valid, i.e. the participant is required to realize that X and Y are disjoint, and this should be apparent from the occurrence of the same anonymous class, in negated and non-negated form, in both statements.

The questions answered least accurately were questions 2 and 6. Question 6 is valid and participants are required to realize that an individual in X, who has no male child, cannot be a member of Y. It is not clear whether there is any particular mechanism contributing to low accuracy here, apart from the general comments that participants had difficulty when a quantifier occurred before a negated class and that there are no obvious syntactic clues.

Before discussing question 2 it is useful to make some general comments about the mental models associated with the universal and existential quantifiers. Table 7 illustrates this; here P is an arbitrary object property, X an arbitrary class, and x an arbitrary member of X. Note that the expression in brackets in the first disjunct in both cases (¬ P x) would normally be assumed in a mental model representation. In the case of *only*, the second disjunct corresponds to the trivial satisfaction of the universal quantifier. This is likely to cause confusion since it does not correspond to the normal usage of the English word 'only'. In effect, people may omit this second disjunct when creating their mental model of the expression. However, this is precisely the disjunct required to realize that question 2 is non-valid because an individual with no children can be a member of both X and Y. A similar problem may occur with *some*, and this will be discussed in section 7.

Table 7 Mental models for the two quantifiers

OWL expression	**mental model**
P only X	P x (¬ P ¬x) P ⊥
P some X	P x (¬ P ¬x) P x P ¬x

Another mechanism which may have led to fallacious reasoning in question 2 is suggested by the fact that this question possessed the minimum mean time for incorrect responses of all questions in the section, i.e. the participants who got this question wrong in general responded very quickly. The question which had the second smallest mean time for incorrect responses was question 5, for which the accuracy of response was also low. Both questions were non-valid, both used the same quantifier in the two statements, and both had a quantifier before a negated class. It is possible that some participants were equating these questions to questions 4 and 7, in which the negation preceded the object property. Rector et al. [11] have already noted that there can be confusion between "some not" and "not some", and this may also apply to "only not" and "not only". It appears that the use of the same quantifier in both questions can either encourage correct reasoning or fallacious reasoning, depending on the structure of the question. A possible hypothesis is that there is a category of participant who reason syntactically and hence will do well on one pair of questions and badly on the other. To test this, for each pair of questions each participant was scored 0, 1 or 2 depending on how many questions of each pair were correct. However, rather than a negative correlation between these scores, there was a Spearman's rank correlation of 0.524 which was significant on a one-sided test ($p = 0.00212$), i.e. those who did well on one pair of questions tended to do well on the other pair.

7. COMBINING QUANTIFIERS

7.1 The questions

The aim of this section was to see how well people coped with statements which combined two quantifiers.

Table 8 Combining quantifiers: all questions have proposed inference *a Type (not X)*

	First statement(s) – constraining X	Remaining statements – constraining a	Valid / Non valid	%age correct N = 28	mean time (SD), N = 24		
					overall - secs	correct - secs	incorrect - secs
1	X SubClassOf (has_child some (has_child some FEMALE))	a has_child b; b Type has_child some (not FEMALE)	N	71%	69 (45)	71 (46)	60 (44)
2	X SubClassOf has_child some Y; Y EquivalentTo has_child only FEMALE		N	57%	79 (53)	101 (55)	52 (37)
3	X SubClassOf (has_child only (has_child some FEMALE))		N	71%	63 (43)	63 (46)	65 (38)
4	X SubClassOf has_child only Y; Y EquivalentTo has_child only FEMALE		V	57%	63 (39)	56 (27)	72 (52)
5	X SubClassOf has_child some Y; Y EquivalentTo has_child some FEMALE	a has_child b; b Type (not (has_child some FEMALE))	N	54%	88 (62)	94 (68)	78 (53)
6	X SubClassOf (has_child some (only FEMALE))		N	64%	73 (45)	66 (44)	84 (46)
7	X SubClassOf has_child only Y; Y EquivalentTo has_child some FEMALE		V	71%	80 (36)	71 (34)	108 (29)
8	X SubClassOf (has_child only (has_child only FEMALE))		N	50%	55 (30)	50 (23)	59 (36)

The questions are shown in Table 8. Each question used two quantifiers to constrain the class X. Questions 1 to 4 use the four variants created from the permutations of *some* and *only* (i.e. *some … some*, *some …only*, *only … some*, and *only …* only); these are repeated in questions 5 to 8. A statement then related the individuals *a* and *b* through an object property (*has_child*). A third statement described *b* as a member of an anonymous class. There were two variants to this statement. Both variants used some, but in one negation occurred before a class identifier whilst in the other negation occurred before the property. All eight questions had the same proposed inference: *a Type (not X)*. In four of the questions the class defined by the second quantifier in the first statement was anonymous. In the other four questions the class was given the identifier Y, in order to investigate what difference the naming of a class made to accuracy or time to respond. To avoid confounding this factor with the structure of the first statement, each permutation of *some* and *only* occurs with and without the named class Y. As indicated in the Table 8, two of the questions are valid whilst six are non-valid.

7.2 Results

Table 8 shows the performance on each question. Only questions 1, 3 and 7 were answered significantly better than chance ($p = 0.0178$ in each case). An initial hypothesis was that there might be a variation in performance between those questions which used two different quantifiers in the first statement compared with those questions which repeated the same quantifier. A χ^2 test revealed that this was not the case for accuracy ($p = 0.271$) and a t-test showed this was not the case for time ($p = 0.262$). A χ^2 test revealed that the variation in the final statement i.e. the position of the negation, had no significant effect on accuracy ($p = 0.582$) and a t-test showed a similar result for time ($t = 0.246$). A χ^2 test also showed no significant difference in accuracy between the questions with Y and the questions in which this class was anonymous. A t-test indicated that the hypothesis that questions with Y took longer than questions with the anonymous class was approaching significance ($p = 0.0814$).

7.3 Discussion

Consider first the questions with valid inferences, questions 4 and 7. In question 7 it is clear from the syntax that Y is the complement of the anonymous class defined in the final statement, i.e. *not (has_child some FEMALE)*. In question 4 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 ($p = 0.0309$ on a t-test), suggesting that participants who answered this question correctly picked up the syntactic clue quickly.

Turning to the six non-valid questions, the least accurately answered was question 8. This question requires the second *only* in the first statement to be expanded using the second disjunct shown in Table 7. As discussed in section 6.2, this is the less favoured interpretation of the quantifier. The participant also has to realize that the statement *b Type (not (has_child some FEMALE))* is satisfied if *b* has no children.

The remaining five questions, in order of decreasing accuracy are: 1, 3, 6, 2 and 5. A point to note is that in question 5, the anonymous class defined in the third statement is quite clearly the negation of class Y defined in the first statement. This may have led participants to conclude correctly that *b* cannot be a member of Y, and then erroneously *a* cannot be a member of X. Note that this argument has nothing to do with the fact that both questions have a named class Y; the same argument would apply if Y were replaced with an anonymous class. There is also a clear distinction between how the mental models need to be formed in the questions 1, 3 and 6, and questions 2 and 5. Consider first question 1. Imagine the participant has read the first part of the first line *(X SubClassOf (has_child some …*) and the second statement (*a has_child b*). Then it is immediately apparent that *a* can be a member of X if *b* satisfies the constraints imposed by the remainder of the first statement as well as the constraint of the third statement. To understand that this is possible requires understanding that *has_child some FEMALE* and *has_child some (not FEMALE)* intersect. This requires the second disjunct in the mental model for

some shown in Table 7. As with *only*, but to a lesser extent, we hypothesize that there is a tendency initially to assume the first disjunct and ignore the second. However, many participants seemed able to develop the correct model in this question. For comparison, consider question 2. As before, the participant reads the first part of the first statement and the second statement and concludes that the response can be non-valid if *has_child only FEMALE* and *has_child some (not FEMALE)* intersect. Further consideration reveals that this is not the case, and the immediate conclusion is that the question is valid. However, all discussion so far has ignored the possibility of interpreting the first *some* using the second disjunct in Table 7. Specifically, if *a* is a member of X, and therefore has a child that is a member of Y, it may also have a child, *b*, which is not a member of Y and has a non-FEMALE child. Assuming that the participant starts with the more obvious interpretation of the first *some*, correctly answering the question requires that the participant, having reached a dead end, then backtracks to the first *some* and considers the second disjunct. This backtracking is also required for question 5, but not for the more accurately answered questions 1, 3 and 6. On this basis, one might expect that for questions 2 and 5 the mean time for the correct responses would be longer than the mean time for the incorrect responses. This is the case, significantly so for question 2 ($p = 0.0327$ on a t-test) but not significantly for question 5 ($p = 0.657$). The difference between correct and incorrect responses was not significant for any of the other non-valid questions.

8. CONCLUSIONS AND RECOMMENDATIONS

The foregoing discussions suggest that mental model theory and relational complexity theory can be used to understand how people understand and reason about DL statements. This has implications for those writing such statements and also for those responsible for how the deductive steps from justification to entailment are presented in the ontology debugging process. Mental model theory explains the difficulties experienced with negated conjunction. This construct should be avoided, e.g. by transforming with Boolean algebra. Mental model theory also explains some of the difficulties experienced with the existential and universal quantifier. It appears that, particularly for *only* but also in some cases for *some*, people may overlook the second disjunct in the mental model. Where possible, OWL should be written to avoid the need for the less favoured disjunct. Another tactic would be to expand quantifier expressions to make explicit the second disjunct, e.g. expand *P some X* as *P some X or (P some X and P some (not X))*, and expand *P only X* as *P only X or not (P some Thing)*. RC theory can be used to explain the variation in time for certain deduction steps involving object properties. Further investigation is required to determine whether RC theory can explain the greater apparent difficulty with functional properties relative to transitive or whether functional properties offer some inherent difficulty.

There is also evidence that participants were also reasoning syntactically. Hence syntax should be used to emphasize semantics, e.g. to make clear complementary classes. On the other hand, apparently similar syntax can confuse, e.g. *P some (not X)* and *not (P some X)*. One remedy for this might be teaching, and ensuring maintained awareness of, the duality rules for predicate logic expressed in Manchester OWL Syntax, i.e. *P some (not X)* ≡ *not (P only X)* and *not (P some X)* ≡ *P only (not X)*. This would have helped participants to reason about questions 1 and 6 discussed in section 7.3, and question 4 discussed in section 8.3.

Of the 34 questions, 17 were not answered significantly better than chance. This suggests a need to make DL statements more intelligible and easier to reason with. It seems likely that Manchester OWL Syntax is a significant improvement over notations from formal logic. However, it also seems likely that improved notations could further enhance comprehensibility. This will form the basis of future work, including investigating notations to draw attention to the 'second' disjunct in the universal quantifier and the uniqueness of the object entity in a functional property.

9. ACKNOWLEDGMENTS

The authors wish to thank all the study participants.

10. REFERENCES

[1] Halford, G. S., & Andrews, G. (2004). : The development of deductive reasoning: How important is complexity? *Thinking & Reasoning*, *10*(2), 123–145.

[2] 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.

[3] Horridge, M., Bail, S., Parsia, B., & Sattler, U. (2011). The cognitive complexity of OWL justifications. *The Semantic Web–ISWC 2011*

[4] Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., & Wang, H. H. (2006). The manchester owl syntax. *OWL: Experiences and Directions*, 10–11.

[5] Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, *50*(1), 109–135.

[6] Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418.

[7] Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods—a survey. *ACM Computing Surveys (CSUR)*, *39*(4), 10.

[8] Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). *Negating compound sentences*. Naval Research Lab Washington DC Navy Center for Applied Research in Artificial Intelligence.

[9] Knauff, M., Rauh, R., Schlieder, C., & Strube, G. (1998). Mental models in spatial reasoning. In *Spatial cognition*, Springer.

[10] Nguyen, Power, Piwek, & Williams. (2012). Measuring the understandability of deduction rules for OWL. 1st intnl. workshop on debugging ontologies & ont. mappings. http://oro.open.ac.uk/34591/

[11] R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria.: R Foundation for Statistical Computing,. Retrieved from http://www.R-project.org/

[12] Rector et al. (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.

[13] Rector, A. L. (2003). Defaults, context, and knowledge: Alternatives for OWL-indexed knowledge bases. In *Pacific Symposium on Biocomputing* (pp. 226–237).

[14] Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, *90*(1), 38.

[15] Scott, D. Tukey's Ladder of Powers. In *Online Statistics Education: A Multimedia Course of Study* (Version 2). Rice University, University of Houston-Clear Lake, and Tufts University. http://onlinestatbook.com/2/transformations/tukey.html

[16] Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3.

[17] Stapleton, G., Howse, J., Bonnington, A., & Burton, J. (2014). A vision for diagrammatic ontology engineering. Retrieved from http://eprints.brighton.ac.uk/13046/

[18] Warren, P., Mulholland, P, Collins, T, & Motta, T. (2014). The usability of Description Logics: understanding the cognitive difficulties presented by Description Logics. ESWC 2014, Crete.