The Usability of Description Logics

Understanding the Cognitive Difficulties Presented by Description Logics

Paul Warren, Paul Mulholland, Trevor Collins, and Enrico Motta

Knowledge Media Institute, The Open University
Milton Keynes, Buckinghamshire, MK7 6AA, U.K.
paul.warren@cantab.net,
{paul.mulholland,trevor.collins,enrico.motta}@open.ac.uk

Abstract. Description Logics have been extensively studied from the viewpoint of decidability and computational tractability. Less attention has been given to their usability and the cognitive difficulties they present, in particular for those who are not specialists in logic. This paper reports on a study into the difficulties associated with the most commonly used Description Logic features. Psychological theories are used to take account of these. Whilst most of the features presented no difficulty to participants, the comprehension of some was affected by commonly occurring misconceptions. The paper proposes explanations and remedies for some of these difficulties. In addition, the time to confirm stated inferences was found to depend both on the maximum complexity of the relations involved and the number of steps in the argument.

Keywords: #eswc2014Warren.

1 Introduction

During the past few decades the decidability and computational tractability of Description Logics (DLs) have been extensively studied. Baader et al. (2010) provide a comprehensive overview of the theory of DLs and also describe a number of applications. In particular, the properties of the various profiles of OWL, the Web Ontology Language, are well understood; see Motik et al. (2012). The usability of DLs has been much less investigated. This paper describes an experiment to understand the comprehensibility of the most commonly used features of DLs, as implemented in OWL. The study has revealed a number of misconceptions and the paper makes suggestions as to how these can be overcome. A particular goal of the study was to determine whether any of the psychological theories of reasoning could be used to explain the accuracy of human reasoning with DL statements and the time taken to undertake such reasoning. This theoretical approach has helped to explain the most significant of the misconceptions and also explained the time to confirm inferences.

Section 2 describes related work. This falls into two categories: work undertaken by computer scientists to understand the comprehensibility of DLs; and the work of

V. Presutti et al. (Eds.): ESWC 2014, LNCS 8465, pp. 550–564, 2014.

cognitive psychologists to understand the nature of reasoning in general. Section 3 then describes how the most commonly used DL features were identified. The study focuses on the features that are commonly used, rather than those features which, whilst useful in particular domains, are not extensively used. Section 4 explains how the study was designed and conducted. Section 5 presents the study questions and discusses the five which were found most difficult by participants. Where applicable we have used theories from cognitive science to help explain these difficulties. Some potential remedies for these problems are also suggested. The section also discusses participants' feedback, which confirm the usefulness of some of the suggestions. Section 6 provides a more detailed analysis, including some results relating the time taken to answer questions to psychological theories of reasoning. Finally, section 7 draws some conclusions and outlines areas for future study.

2 Related Work

2.1 Comprehensibility of Description Logics

There have been few studies of the comprehensibility of Description Logics. Rector et al. (2004) describe the difficulties experienced by newcomers to OWL, based on their experience in teaching the language. They provide a set of guidelines and also English paraphrases of some OWL expressions. Horridge et al. (2011) were interested in supporting the ontology developer during the debugging process. One way to offer such support is to display the minimal subset of the ontology that generates a particular entailment. Such a subset is termed a justification. Horridge et al. investigated the cognitive complexity of justifications for entailments of OWL ontologies. They developed a complexity model and compared the predictions of this model with the difficulty experienced by computer scientists in identifying correct entailments. Their model, which was not grounded in any psychological theory, "fared reasonably well". Commenting on a study by Newstead et al. (2006), Horridge et al. identified a strong advantage of studies within the psychological literature, i.e. that the problems considered "are very constrained and comparatively easy to analyse". The difficulty in studying DLs is the need to consider a wide range of commonly occurring constructs.

Nguyen et al. (2012) had a similar interest in assisting developers to debug ontologies. Their goal was to explain, in English, why an entailment follows from an ontology. In particular, they wished to predict the comprehensibility of alternative proof trees, when expressed in English. To do this they needed to first understand the comprehensibility of the individual deduction rules comprising a proof tree. They took 51 such deduction rules, expressed in English, and tested their comprehensibility on participants obtained through a crowdsourcing service. This enabled them to generate a facility index representing the ease of comprehensibility of each deduction rule, calculated as the proportion of participants who identified the deduction rule as being correct. Since their interest was in calculating these facility indices for future use, they did not attempt to create a model to predict and explain the indices.

2.2 Theories of Human Reasoning

There is a considerable psychological literature on human reasoning. One distinction is between rule-based and model-based theories. The rule-based approach is represented by the work of Rips, e.g. Rips (1983). The assumption is that the processes of human reasoning are akin to the steps executed by a logician in carrying out a proof. The model-based approach is represented by Johnson-Laird, e.g. Johnson-Laird and Byrne (1991), for whom mental models "have the same structure as human conceptions of the situations they represent". Johnson-Laird and his collaborators have built an extensive body of experimental evidence to support the view that at least 'naïve reasoners' (i.e. people not trained in logic), do use mental models in reasoning. It may be, though, that some people use a mixture of mental models and rules-based reasoning, depending upon the particular situation and their degree of training in logic. Our hypothesis is that when logicians are constructing a proof in a rule-based way, they use mental models at some or all of the deduction steps.

The mental model theorists propose that difficulties in reasoning often occur when several models need to be maintained in working memory. It is suggested that in certain situations people may ignore some of the possible models, thereby leading to errors. This may happen, for example, when a disjunction occurs. Moreover, mental model theory suggests that an inclusive disjunction will give rise to more errors than an exclusive disjunction, since the former requires three models to be held in working memory whilst the latter requires only two. This is borne out by experiment, e.g. see Johnson-Laird et al. (1992).

Relational complexity (RC) theory offers another approach to understanding performance in reasoning. Here complexity is defined "as a function … of the number of variables that can be related in a single cognitive representation", i.e. the number of arguments of a relation (Halford & Andrews, 2004). The theory proposes that it is the maximum relational complexity in a given process of reasoning which determines the difficulty of that reasoning. RC theory could be seen as compatible with either the rule-based or the model-based approach. Zielinski et al. (2010) have attempted to reconcile the mental model and RC theories for categorical syllogisms. Goodwin and Johnson-Laird (2005) have combined mental model theory and RC theory in their study of reasoning about relations. Apart from the number of arguments, they see depth of the relation as contributing to complexity. Relations between individuals are regarded as of first-order depth; relations between relations of second-order depth; relations between relations between relations of third-order depth.

It seems likely that the experimental success of the mental model theory, e.g. as described in Johnson-Laird and Byrne (1991) and Johnson-Laird et al. (1992), arises because the individual models were of broadly equal complexity. As a consequence, situations requiring two models created more difficulty than those requiring one, and situations requiring three models were even more difficult.

3 Identifying the Commonly Used Features

To identify the most commonly used features we drew on four sources. 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[1]. Power (2010) also used TONES to identify common axiom patterns. This identified the frequency of use of Boolean operators such as intersection, and also of the existential and universal restrictions. Khan and Blomqvist (2010) searched 682 online ontologies to determine the frequency of occurrence of content patterns from the ODP (ontology design pattern) portal[2]. This portal provides information on a variety of patterns, including around 100 content patterns. These are essentially small autonomous solutions to particular design problems. We then analyzed the 20 most frequent patterns that they detected, to determine the commonly occurring OWL features. In addition, Warren (2013) undertook a survey of ontology users which included a question about the usage of OWL features. Based on 47 responses to this question, these features were ranked by frequency of reported use.

The resultant lists were then compared. They identified broadly the same set of commonly occurring features. There were a few differences, e.g. Power and Third found class equivalence to be the second most commonly used functor; analysis of the common content patterns from Khan and Blomqvist identified class equivalence as the thirteenth most commonly used OWL feature, whilst it was not included in the survey by Warren. The set of features used in this study consisted of all those features which were relatively common in at least one of these lists, with two exceptions. The reason for these exceptions was that the study participants would not necessarily be familiar with OWL and they would need to be given information about the language which should be kept brief. Firstly, all features relating to datatype properties were ignored. It was felt that datatype properties would present no cognitive challenges that could not be represented with object properties. This is not to say that there might not be challenges arising from datatype properties during the learning process, but rather that subsequently, when working with ontologies, they do not give rise to any specific problems of cognition. Secondly, cardinality restrictions were not included. In fact, these did not occur in the list of most commonly used patterns in Power and were ranked relatively low in the survey by Warren. The 'min 1' cardinality restriction did occur moderately frequently in the patterns identified by Khan and Blomqvist. This states that a particular individual is the subject of at least one instance of a particular property. It is equivalent to an existential restriction, which was included in the study.

Table 1 shows the set of OWL features chosen for the study. In each case the Manchester OWL Syntax (MOS) representation of the language feature is also shown (Horridge et al., 2006). The features are grouped into those relevant to classes, those relevant to properties and the existential and universal restrictions. In each of these groupings, they are listed broadly in order of occurrence (i.e. with most commonly occurring at the top), although, as already noted, rankings differed across the various sources.

---

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

[2] http://ontologydesignpatterns.org/wiki/Main_Page

Table 1. Commonly used OWL features investigated in the study

Class features

language feature	MOS
subsumption	SubClassOf
class equivalence	EquivalentTo
disjoint classes	DisjointWith
class assertion	Type
conjunction	and
disjunction	or
complement	not

Property features

language feature	MOS
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

language feature	MOS
qualified existential restriction	some
universal restriction	only

4 The Study

In order to test comprehension of these OWL features, they were incorporated into a set of twenty-one questions based on three patterns from the the ODP portal (see beginning of previous section). The three patterns used were: Componency, Coparticipation, and Types of Entities; the second and third were modified to enable all the features in table 1 to be tested, with some simplification of the second to remove unnecessary statements. The three modified patterns were associated with ten, six and five questions. Each of the questions consisted of a set of statements and a proposed inference. The participant was required to indicate, by clicking on a button, whether the inference was or was not valid. In all there were thirteen questions with valid inferences and eight where the inference was not valid. The patterns and question statements were expressed in a simplified form of the Manchester OWL Syntax. Classes and properties were defined in the patterns and had intuitive names. Individuals were defined in the questions and were named A, B, C, D.

These patterns and questions were incorporated in a test, using the tool *SurveyExpression*[3]. The three patterns were ordered in all six permutations; each of these permutations existed in a form with the 'yes' option first and with the 'no' option first, to safeguard against any bias from the order of the possible answers. Thus there were twelve variants of the test. There were also twelve participants, i.e. one participant per variant. All the participants, of whom three were female, were researchers at the Centre for Research in Computing or the Knowledge Media Institute at the Open University, U.K. Screen capture software was used to record the participant's behaviour and in particular to provide the precise times spent in each question. The participants were observed as they took part in the study and any comments they made were noted.

The study was organized into five sections. In the first section participants were asked to rate their knowledge of logic and of OWL. The wording and the percentage of responses in each category for logic and OWL respectively were: no knowledge at all (0%,0%); a little knowledge, e.g. from an introductory course in formal logic (17%,25%); some knowledge, e.g. from a university course in formal logic

[3] http://www.surveyexpression.com/

(67%,42%); expert knowledge (17%,33%). The next three sections contained the three patterns and the questions. Each section began with a webpage displaying the pattern and then a series of pages, with each page repeating the pattern and containing one question. Participants were able to move through the pages at their own speed. The final section was an opportunity for the participants to provide feedback. Participants were provided with a handout containing all the necessary information about the OWL features and notation used. They were asked to read this at the beginning and it was available to them for reference during the session.

5 Study Results

5.1 The Difficult Questions

Of the 21 questions, eight were answered correctly by all of the participants, four were answered correctly by all but one of the participants, and a further four were answered correctly by all but two of the participants. The remaining five are discussed here in decreasing order of difficulty. Tables 2, 3 and 4 show the three patterns and the associated questions. The columns headed 'yes/no', 'MM' and 'RC' represent the correct answer, the maximum number of mental models and the maximum relational complexity associated with the question. The column headed 'num steps' shows the number of steps to arrive at a correct deduction for questions with answer 'yes'. The remaining two columns show the percentage of correct responses and the average time for each question.

Table 2. Questions based on the modified entity types 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

	Question	yes/ no	MM	RC	num steps	%age corr	av. time (secs)
1	A represents B; C represents D => A DifferentFrom C	no	1	2	n/a	83%	91.5
2	A Type Entity; A Type not (Event and Quality) => A Type (Abstract or Object)	no	4	3	n/a	25%	75.1
3	A represents B; C represent D; B Type Object; D Type Event => A DifferentFrom C	yes	1	4	2	50%	75.8
4	A Type Entity; A Type not (Event or Quality) => A Type (Abstract or Object)	yes	4	4	2	92%	44.0
5	A Type (Nonconceptual and Nontemporal) => A Type Object	yes	3	3	3	75%	63.1

Table 2: Q2 - Complementing the *and* operation (ans: no; 25% correct responses)
The most direct way of arriving at a 'no' conclusion for this question is to note that, since Event and Quality are disjoint classes, then *Event and Quality* must be Nothing (⊥). Hence *not (Event and Quality)* is Thing (⊤) and the statement *A Type not (Event and Quality)* is tautological.

The question should be contrasted with question Q4 in table 2, which is identical in form except for the replacement of the *and* with *or*, and has correct answer 'yes'. Q4 had 92% correct responses and was answered much more quickly; the average response time was 44 seconds for Q4 and 75 seconds for Q2. It is interesting to compare these results with those of Khemlani et al. (2012a), reporting on an experiment with 'naïve reasoners'. They investigated the comprehension of compound sentences of the form *not (A and B)* and *not (A or B)* and found a similar wide gap in accuracy of answering questions: 18% correct for the negated conjunction and 89% correct for the negated disjunction. They interpret these results in terms of the mental model theory. *not (A or B)* consists of only one mental model, i.e. *not A and not B*. However, *not (A and B)* consists of three mental models: *not A and not B*; *A and not B*; *not A and B*. They suggest that many people do not go beyond constructing the first of these, leading to erroneous reasoning. They elaborate this theory of negation in (Khemlani, Orenes, & Johnson-Laird, 2012b). In fact, in both Q2 and Q4, arguably four mental models are required, representing the decomposition of Entity (*Event or Abstract or Quality or Object*). The problem is not simply one of managing a number of different models, but of the difficulty of creating the full set of models in the negation process. In Q4 all the participant has to do is to erase Event and Quality from the decomposition of Entity, leaving Abstract and Object. In Q2, rather than evaluate *Event and Quality* and then *not (Event and Quality)* as proposed in the paragraph above, many participants may attempt to expand *not (Event and Quality)* and may arrive at a single mental model corresponding to the term *not Event and not Quality*.

Apart from emphasis during training, potential solutions to this problem are:

- automatic expansion of *not (A and B)* into its three atomic constituents;
- an automatically generated graphical representation.

There is also the additional possibility of confusion between the everyday use of *and* and its logical use. It may be that, when faced with the difficulty of negating a conjunction, participants take the easy option by interpreting *and* as equivalent to *or* (e.g. as in the English statement "the car is available in blue and silver"). The use of an alternative keyword to *and*, e.g. *intersection* or *int*, could avoid this linguistic confusion.

Whatever the reason for the erroneous treatment of Q2, it is striking that the results for naïve reasoners were so similar to those found amongst our participants. This suggests that both groups were using the same mental processes.

The complement operation was used by 22 of the 47 respondents to the question on DL feature usage in the survey by Warren (2013). However, it was not identified by Power (2010) as being commonly used and was not in the commonly used patterns identified by Khan and Blomqvist (2010). This relatively low usage may account for the low proportion of correct responses, since some participants may not have been familiar with the use of the complement operation.

Table 3. Questions based on the componency pattern

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

	Question	yes/ no	MM	RC	num steps	%age corr	av. time (secs)
1	A is_part_of B; C is_part_of B => A is_part_of C	no	1	2	n/a	100%	62.3
2	A is_part_of B; B is_part_of C => A is_part_of C	yes	1	2	1	100%	20.3
3	B is_part_of C; A is_part_of B => A is_part_of C	yes	1	3	1	100%	30.7
4	A has_component B; B has_component C => A has_component C	no	1	2	n/a	33%	62.8
5	A has_component B; B has_component C => A has part C	yes	1	2	3	83%	29.0
6	A has_component B; B is_part_of C => A has_part C	no	1	2	n/a	83%	57.9
7	A has_component B; C is_part_of B => A has_part C	yes	1	4	3	100%	37.4
8	A Type Object; A has_component B; C Type not Object => B DifferentFrom C	yes	1	3	2	100%	49.9
9	A Type Object; A has_part B; C Type Not Object => B DifferentFrom C	no	1	2	n/a	83%	47.5
10	A has_component B; C is_component_of B => C is_part_of A	yes	1	4	4	100%	54.2

Table 3: Q4 - Non-inheritance of transitivity (ans: no; 33% correct responses)
In the pattern, *has_component* is defined as a subproperty of *has_part*, which is defined to be transitive. For the deduction to be true it would be necessary for *has_component* to also be transitive. There are a number of reasons why this question might be answered incorrectly. It may be that participants forget which is the parent, transitive property, and which is the subproperty, i.e. that they confuse the two names. It might also be that the name *has_component* suggests transitivity. Alternatively, people may assume that property characteristics are necessarily inherited by subproperties. This would be natural for people coming from an object oriented background, or those chiefly used to thinking about class subsumption relations in ontologies. That this is not the case for transitivity was noted in the handout, which cited the example of the property *is_descendant_*of and its subproperty *is_child_of*[4]. In fact, a different choice of property name might guard against all these problems; *has_direct_part*, in place of *has_component*, could better convey the required meaning. Subproperties appear to be relatively frequently used, and the transitive characteristic is one of the most commonly used characteristics. When training ontology users, attention needs to be drawn to the fact that not all characteristics are inherited, perhaps spelling out those which are and those which are not.

[4] Similarly, the characteristic of symmetry is not inherited, as can be seen from the property *is_sibling_of* and its subproperty *is_brother_of*. On the other hand, functionality is inherited, since if a subproperty has two values for the same subject, then so will its superproperty.

Table 2: Q3 - The functional characteristic (ans: yes; 50% correct responses)
Since Object and Event are disjoint, B and D must be different. The functionality of *represents* then ensures that A and C are different. It may be that those who answered this question incorrectly did not fully understand the nature of a functional characteristic, despite it being explained in the handout. It may also be that the high relational complexity (i.e. RC = 4) of the question contributed to its difficulty. Here again, a diagrammatic representation would aid comprehension.

Table 4. Questions based on the 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

	Question	yes/ no	MM	RC	num steps	%age corr	av. time (secs)
1	A coparticipates_with B => A Type not Event	yes	1	2	2	92%	54.9
2	A is_participant_in B; C coparticipates_with D => A DifferentFrom C	no	1	2	n/a	92%	68.8
3	A is_participant_in B; C is_participant_in B => A is participant in C	no	1	2	n/a	100%	43.6
4	A has_participant B; C is_particpant_in D => B DifferentFrom D	yes	1	2	3	92%	44.6
5	B coparticipates_with A; B coparticipates_with C => C coparticipates_with A	yes	1	3	2	100%	34.8
6	A Type Game => A Type Event	yes	1	3	3	67%	47.6

Table 4: Q6 - The existential quantifier (ans: yes; 67% correct responses)
Each member of the class Game *has_participant some Player*; hence this is true of A. Since Player is a subclass of Object, *A has_participant some Object*. Since Event is defined as the set of individuals that *has_participant some Object*, A is in Event. The fact that a relatively large number of participants got this question right, despite its apparent complexity, may be due to the frequency of use of the existential quantifier, e.g. see Khan and Blomqvist (2010), Power (2010) and Warren (2013).

Table 2: Q5 - Superclasses (ans: yes; 75% correct responses)
Participants did relatively well on the question, only two providing incorrect answers and one not responding. In all the analyses the non-response is treated as an incorrect answer. The question is discussed here in part because it provides an example of a different approach to the use of mental models. Entity is composed of four disjoint subclasses. Nonconceptual and Nontemporal comprise two and three of these disjoint subclasses respectively. The only one they have in common is Object; hence their conjunction is equivalent to Object. A straightforward application of the mental model approach suggests that the maximum number of mental models is three, since Nontemporal is comprised of three disjoint classes. There is, however, little difficulty in formulating these models, unlike in the case of negation of a conjunction in

table 2, Q2. In fact, a quite natural way to think about this is as two overlapping superclasses, with Object constituting the overlapping portion. This also lends itself naturally to a graphical representation; some participants might even have visualized it. This only requires that two models be held in working memory, one representing Nonconceptual, and the other Nontemporal.

5.2 Participants' 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. 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 the formal axioms. The contrasting views were also present when considering class and property names. One participant (p3) commented that because the 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". Participant p1 also commented favourably on the lack of use of formal logic symbols, which is a feature of the Manchester OWL syntax.

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". Indeed, the automatic generation of diagrams is likely to have helped comprehension in all the questions discussed above. 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, see Q2 from table 1 discussed in subsection 5.1); and (participant p7) the effect of users' legacy, e.g. that of a database background. Each of these comments is relevant to one of the questions discussed in section 5.1.

6 Statistical Analysis

6.1 The Participants

The majority of participants achieved high scores; two achieved twenty out of twenty-one, whilst the lowest score was thirteen. Ranking on knowledge of logic and OWL significantly correlates with ranking on accuracy. The Spearman rank correlation coefficient between knowledge of logic and accuracy was 0.53, corresponding to $p = 0.038$ on a one-tailed t test. For the correlation between knowledge of OWL and

accuracy, the coefficient was 0.54, corresponding to p = 0.036 on a one-tailed t test. The effect was greater when we consider just the questions with correct answer 'yes'. For these questions, the correlation factor was 0.57 (p = 0.027) for knowledge of logic and 0.60 (p = 0.019) for knowledge of OWL. For the 'no' questions the rank correlation with knowledge of logic and knowledge of OWL were no longer significant (p = 0.052 and p = 0.102 respectively). The number of participants did not permit a statistical analysis on a per-question basis. None of the participants classifying themselves as having a little knowledge of either logic or OWL answered correctly the first two questions discussed in section 5.1, i.e. the two questions with fewest accurate responses. However, there were also some experts who got these questions wrong.

There was considerable variation in the total time taken to answer the questions, ranging from around thirteen minutes to around forty-two minutes. For our participants, knowledge of OWL had a much greater effect on the total time taken than did knowledge of logic. For knowledge of OWL the Spearman rank correlation coefficient was -0.65, with p = 0.011 on a one-tailed t test; for knowledge of logic the coefficient was -0.29, with p = 0.178. The low correlation in the case of knowledge of logic may have occurred because the majority (67%) of our participants ranked themselves in the same category ('some knowledge').

6.2 The Questions

Most of the questions were answered correctly by all or most of the participants, with an apparent tendency to achieve greater accuracy on the questions with correct answer 'yes'. Table 5 provides a breakdown of the responses showing how many were correct and incorrect for the two categories of questions; it also shows the average times for each combination. A Pearson χ^2 test confirmed the greater accuracy on the 'yes' questions (p = 0.005). The greater accuracy for the 'yes' questions occurs despite the fact that the average maximum relational complexity for these questions is greater than that for the 'no' questions, i.e. they appear on average to be harder.

Table 5. Breakdown of responses, also showing average times in each category

	Yes	No
Correct	138; 43.4 secs	72; 59.5 secs
Incorrect	18; 58.4 secs	24; 76.3 secs

The time spent by any participant answering a single question varied from 9 seconds to 208 seconds, the average time across all participants for each of the twenty-two question varied from 20.3 seconds to 91.5 seconds. A two sample one-sided unpaired t test indicated that the 'yes' questions were answered on average significantly more quickly than the 'no' questions (p < 0.001). This may represent a tendency to initially attempt to prove the validity of the deduction. After first such attempts fail, the participant then has two possible strategies: either to continue such attempts until convinced that a proof is not possible or to attempt to prove explicitly

that the deduction does not hold. The strategy adopted is likely to depend upon the person and the particular question. A one-sided unpaired t test also indicated that the correct responses were arrived at significantly more quickly than the incorrect responses ($p < 0.001$). Thus there was no trade-off between accuracy and speed of response. A two-way ANOVA indicated that the two factors did not interact ($p = 0.884$). Hence the correct responses to the 'yes' questions averaged the least time (43.4 seconds) whilst the incorrect responses to the 'no' questions averaged the greatest time (76.3 seconds).

A simple linear regression showed that, overall, questions with a large number of correct responses were answered more quickly than questions with fewer correct responses ($p < 0.001$). This was also true when analysis was restricted to those questions with correct answer 'yes' ($p < 0.001$) and also to all those questions correctly answered ($p = 0.001$). However, there was no significant relationship between time and number of correct responses for those questions where the correct answer was 'no' ($p = 0.349$), nor for the incorrectly answered questions ($p = 0.947$).

6.3 Theories of Reasoning

As already noted, an objective was to determine whether any of the psychological theories could be used to predict, in terms of accuracy and time, the behaviour of our participants, and thus whether any of these theories would be useful in understanding how people reason about DLs. Each question was analyzed to determine the maximum number of mental models and maximum relational complexity which it would entail; as shown in tables 2, 3 and 4. For the questions with correct answer 'yes', this was done by constructing the proof of the deduction, and determining the number of mental models and the relational complexity at each stage. The questions with correct answer 'no' were examined to determine the maximum number of mental models and maximum relational complexity which would be met in thinking about them.

Only three questions required more than one mental model, making any statistical analysis impossible. One (table 2, Q2) was a 'no' question requiring four mental models and was the least well answered, with only three correct responses; this question is discussed at the beginning of subsection 5.1. The other two were 'yes' questions requiring four (table 2, Q4) and three mental models (table 2, Q5) and with 11 and 9 correct responses; the second of these is also discussed in subsection 5.1. Moreover, whilst these three questions might be regarded as requiring more than one mental model, a participant with some knowledge of logic might well have used an alternative approach. The prevalence of questions requiring only one mental model seems to arise from the way in which Description Logics are used. The complexity is often in the relations, rather than in the existence of numerous possibilities.

This last statement might lead one to expect that relational complexity would be a better predictor of performance. However, a logistic regression of accuracy against maximum relational complexity did not provide a significant result ($p = 0.665$). This was also the case for a linear regression of time to answer each question versus maximum relational complexity ($p = 0.861$). When the latter regression was limited to the 'yes' questions, then there was a significant result ($p = 0.009$). The difference

between 'yes' and 'no' questions may be a feature of the way people approach 'no' questions, or it may arise from the design of questions; all but one 'no' question had relational complexity of 2.

The rule-based theory leads one to expect that, for the 'yes' questions, performance might be predicted by the number of steps in the reasoning chain. Whatever the validity of this theory is for naïve reasoners in everyday life, it could have some relevance to our participants, all of whom had at least a little knowledge of logic. This was investigated by looking at the 13 'yes' questions. It might be expected that, as the number of steps in the reasoning chain increases, the accuracy of answering will decline, as the possibility of error multiplies and fatigue sets in. A logistic regression of accuracy against number of steps did not provide a significant result ($p = 0.355$). However, all questions had one, two or three steps, with the exception of Q10 of table 3 which had four steps and was correctly answered by all participants. When this question is removed from the analysis, a significant result is achieved ($p = 0.046$).

A linear regression of time for each response against number of steps also provided a significant result ($p = 0.036$). This was slightly more significant when only the correct responses were analyzed ($p = 0.028$), but not significant for the incorrect responses ($p = 0.360$). The mean times for one, two, three and four step questions were 25.5, 51.9, 44.3 and 54.2 seconds respectively. A Tukey range test at the 95% level indicated a significant difference between the means of only the first two groups.

7 Conclusions and Future Work

This study represents an attempt to understand the cognitive difficulties of using Description Logics. A key message is that, despite training, users are prone to certain misconceptions. These include confusion about the combined use of *not* and *and*; about the inheritance of property characteristics; and to a lesser extent about the functional characteristic and also the existential quantifier. Confusion may also arise through choice of names, a point taken up in the comments made by participants. The use of realistic names can lead to erroneous intuitions; however the mnemonic advantage is likely to outweigh this disadvantage. The important thing is to use names which are not likely to create incorrect intuitions.

In the study, maximal relational complexity did not significantly affect accuracy but did significantly affect the time to confirm an inference. The number of steps in a reasoning chain affected the time to reason and also, when one question was removed from the analysis, affected the accuracy of reasoning. Given the participants' background and the nature of the questions, it seems likely that they did at the conscious level adopt a rule-based approach, as evidenced by the effect of number of steps on accuracy and time. The fact that one-third of the respondents commented on the value of drawing a diagram indicates that they were also thinking in terms of models.

There are a number of different areas for further investigation. The effect of alternative linguistic usages is a valuable area for study. This applies to keywords, entity names and the structure of logical statements. For the non-logician the Manchester

OWL Syntax appears to offer a considerable improvement over formal logical notation, as evidenced by one of the participants' comments. However, the choice of linguistic terms can have a significant effect on cognition, e.g. see Johnson-Laird and Byrne (1989); this has not been systematically studied in the context of Description Logics. Allied to this is the effect of the way the linguistic terms are displayed, e.g. with the use of colour coding.

The value of diagrams in reasoning has been noted by other researchers. Larkin and Simon (1987) have analysed the benefits of diagrams in terms of support for search, recognition and inference. The importance of the design of diagrams has been pointed out by Bauer and Johnson-Laird (1993) who noted the need to avoid arbitrary symbols and make explicit alternative states of affairs. In the context of DLs, diagrams offer a strategy to overcome misconceptions and generally support reasoning. In fact, a large number of tools have been created to display ontological structures, e.g. see Katifori et al. (2007). These are chiefly aimed at viewing the structure of the overall ontology or parts of the ontology, i.e. at the subsumption relations, rather than the more cognitively difficult features of Description Logics. An exception to this is the work of Howse et al. (2011) who use concept diagrams not only to view subsumption relations but also to view and reason about role restrictions. The most successful such representations may mirror the mental models we construct when reasoning, so this work should also draw on what is known about such mental models and what is known about the role of diagrams in complementing reasoning.

A valuable experimental strategy would be to compare the current approach with a modified linguistic approach and with a diagrammatic approach, and possibly with a combination of the last two.

Finally, a better understanding of the effect of complexity would require a controlled study in which questions of varying complexity are posed with a limited subset of OWL features which are known to be well understood, to avoid any confounding of complexity with differences in comprehension of different logical features.

In each case a better understanding of the participant's thought processes is needed. Follow-up questions to specifically elucidate this would be valuable.

References

1. Baader, F., Calvanese, D., McGuiness, D., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook: Theory, Implementation and Applications. In: CUP (2010)
2. Bauer, M.I., Johnson-Laird, P.N.: How diagrams can improve reasoning. Psychological Science 4(6), 372–378 (1993)
3. Goodwin, G.P., Johnson-Laird, P.N.: Reasoning about relations. Psychological Review 112(2), 468 (2005)
4. Halford, G.S., Andrews, G.: The development of deductive reasoning: How important is complexity? Thinking & Reasoning 10(2), 123–145 (2004)
5. Horridge, M., Bail, S., Parsia, B., Sattler, U.: The cognitive complexity of OWL justifications. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 241–256. Springer, Heidelberg (2011)
6. Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., Wang, H.H.: The manchester owl syntax. OWL: Experiences and Directions, 10–11 (2006)

7. Howse, J., Stapleton, G., Taylor, K., Chapman, P.: Visualizing ontologies: A case study. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 257–272. Springer, Heidelberg (2011)
8. Johnson-Laird, P.N., Byrne, R.M.: Only Reasoning. Journal of Memory and Language 28(3), 313–330 (1989)
9. Johnson-Laird, P.N., Byrne, R.M.: Deduction. Lawrence Erlbaum Associates, Inc. (1991), http://psycnet.apa.org/psycinfo/1991-97828-000 (retrieved)
10. Johnson-Laird, P.N., Byrne, R.M., Schaeken, W.: Propositional reasoning by model. Psychological Review 99(3), 418 (1992)
11. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods—a survey. ACM Computing Surveys (CSUR) 39(4), 10 (2007)
12. Khan, M.T., Blomqvist, E.: Ontology design pattern detection-initial method and usage scenarios. In: SEMAPRO 2010, The Fourth International Conference on Advances in Semantic Processing, pp. 19–24 (2010), http://www.thinkmind.org/index.php?view=article&articleid=semapro_2010_1_40_50071 (retrieved)
13. Khemlani, S., Orenes, I., Johnson-Laird, P.N.: Negating compound sentences. Naval Research Lab Washington DC Navy Center for Applied Research in Artificial inTELLigence (2012a), http://mindmodeling.org/cogsci2012/papers/0110/paper0110.pdf (retrieved)
14. Khemlani, S., Orenes, I., Johnson-Laird, P.N.: Negation: A theory of its meaning, representation, and use. Journal of Cognitive Psychology 24(5), 541–559 (2012b)
15. Larkin, J.H., Simon, H.A.: Why a diagram is (sometimes) worth ten thousand words. Cognitive Science 11(1), 65–100 (1987)
16. Motik, B., Grau, B., Horrocks, I., Wu, Z., Fokoue, A., Carsten, L.: OWL 2 Web Ontology Language Profiles, 2nd edn. W3C (2012), http://www.w3.org/TR/2012/REC-owl2-profiles-20121211/ (retrieved)
17. Newstead, S.E., Bradon, P., Handley, S.J., Dennis, I., Evans, J.S.B.: Predicting the difficulty of complex logical reasoning problems. Thinking & Reasoning 12(1), 62–90 (2006)
18. Nguyen, Power, Piwek, Williams: Measuring the understandability of deduction rules for OWL. Presented at the First International Workshop on Debugging Ontologies and Ontology Mappings, Galway, Ireland (2012), http://oro.open.ac.uk/34591/ (retrieved)
19. Power, R.: Complexity assumptions in ontology verbalisation. In: Proceedings of the ACL 2010 Conference Short Papers, pp. 132–136 (2010), http://dl.acm.org/citation.cfm?id=1858866 (retrieved)
20. Power, R., Third, A.: 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 (2010), http://dl.acm.org/citation.cfm?id=1944682 (retrieved)
21. Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wroe, C.: 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 (2004), http://link.springer.com/chapter/10.1007/978-3-540-30202-5_5 (retrieved)
22. Rips, L.J.: Cognitive processes in propositional reasoning. Psychological Review 90(1), 38 (1983)
23. Warren, P.: Ontology Users' Survey - Summary of Results (KMi Tech Report No. kmi-13-01) (2013), http://kmi.open.ac.uk/publications/pdf/kmi-13-01.pdf (retrieved)
24. Zielinski, T.A., Goodwin, G.P., Halford, G.S.: Complexity of categorical syllogisms: An integration of two metrics. European Journal of Cognitive Psychology 22(3), 391–421 (2010)