Using Insights from Psychology and Language to Improve How People Reason with Description Logics

Conference or Workshop Item

Using insights from psychology and language to improve how people reason with Description Logics

Paul Warren, Paul Mulholland, Trevor Collins, 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. Inspired by insights from theories of human reasoning and language, we propose additions to the Manchester OWL Syntax to improve comprehensibility. These additions cover: functional and inverse functional properties, negated conjunction, the definition of exceptions, and existential and universal restrictions. By means of an empirical study, we demonstrate the effectiveness of a number of these additions, in particular: the use of *solely* to clarify the uniqueness of the object in a functional property; the replacement of *and* with *intersection* in conjunction, which was particularly beneficial in negated conjunction; the use of *except* as a substitute for *and not*; and the replacement of *some* with *including* and *only* with *noneOrOnly*, which helped in certain situations to clarify the nature of these restrictions.

Keywords: Description Logics, Psychology of Reasoning, Philosophy of Language, Empirical Studies

1 Introduction

The motivation for the research reported here is to mitigate the difficulties which occur when an ontologist reasons using DLs, in particular Manchester OWL Syntax (MOS). The ontologist needs to be able to perform such reasoning to understand the consequences of an ontology design, and also to understand why an entailment leads to a particular inference when debugging an ontology. In this context, the ontologist might be a computer scientist with a relatively deep knowledge of logic or a domain expert with less training in formal logic.

Previous studies by the authors have investigated the difficulties people experience in comprehending and reasoning with Description Logics (DLs). Warren et al. (2014) studied the difficulties experienced with DL features drawn from commonly used pat-

terns, using a simplified version of the Manchester OWL Syntax (MOS). They identified particular difficulties with negated conjunction and functional properties. A subsequent study (Warren et al., 2015) investigated these features in more detail, as well as looking at the effect of combining negation and a restriction and also nested restrictions. As a result, Warren et al. (2015) made some recommendations regarding training. Both these studies sought to explain participants' difficulties in terms of theories of human reasoning. Based on these studies, the authors have proposed some syntactic additions to MOS to mitigate the difficulties identified. The study reported here investigates the effect of these additions to MOS.

Section 2 describes related research on the difficulties experienced with DLs. Section 3 then describes the theories of human reasoning and language employed by the authors. Section 4 reviews the findings of the authors' two previous studies and their interpretation in terms of the theories discussed in section 3. Section 5 provides an overview of this study and makes some general observations. Section 6 investigates the use of an additional keyword, *solely*, to clarify the nature of functional and inverse functional properties. Section 7 is concerned with Boolean concept constructors. Specifically it investigates the effect of replacing *and* and *or* with *intersection* and *union*, the use of *except* as a substitute for *and not*, and also the use of prefix notation for conjunction and disjunction. Sections 8 and 9 investigate the use of *including* and *noneOrOnly* in place of *some* and *only*, with the intention of clarifying the nature of these restrictions. These two sections also investigate the effect of replacing the keyword *some* with *any* where the associated property is preceded by a negation. Finally, section 10 draws some conclusions and proposes some future work.

2 Related work

It has long been recognized that the original formal notation of DLs posed problems for those who were not logicians. This was the motivation for the Manchester OWL Syntax (Horridge et al., 2006). However, based on their experience of teaching DLs, Rector et al. (2004) observed that, even with a notation more akin to natural language, DLs still posed problems of comprehensibility. Indeed, Rector et al. (2004) point out that the use of natural language can create ambiguities, observing that *and* and *or* in everyday use do not always correspond to their meanings in logic.

There has been some empirical work investigating both the comprehensibility of DLs and the facility of human reasoning with them. Horridge et al. (2011) have investigated the difficulties experienced by users trying to understand how subsets of an ontology justified particular entailments, as is necessary when debugging an ontology. Participants were presented with a justification and an entailment, expressed in a formal notation, and were asked whether the entailment followed from the justification. The empirical results were related to an ad-hoc model of cognitive complexity with 12 parameters, e.g. number of axiom types, number of class constructors and maximum depth of class expressions in the justification.

Nguyen et al. (2012) were concerned with automatically creating proof trees, composed from deduction rules. When choosing between alternative possibilities they

needed a measure of the comprehensibility of each of 51 deduction rules. They created English equivalents of these deduction rules, using 'nonsense' words to avoid any effect of domain knowledge. Study participants, drawn from a crowdsourcing service, were required to confirm or reject the validity of the various deduction rules, thereby determining a comprehensibility rating for each rule. The intention was that these ratings could be used by an algorithm to create a proof tree, optimized for comprehensibility.

None of this work made any use of psychological theory. Our work goes beyond this previous work by looking more precisely at the difficulties experienced with specific OWL constructs, and interpreting those difficulties in terms of psychological theories.

3 Human reasoning and human language

There has been considerable research into how people reason. Two early opposing approaches were the rule-based and model-based approaches. The former assumes that 'naïve users', i.e. people not trained in logic, use rules similar to that of the logician (Rips, 1983). The model-based view, as argued by Johnson-Laird (2010), assumes that people create mental models of a given situation. In this view, any putative deduction is tested against the various models. If the deduction is true in every model, then it is a valid conclusion.

The mental model theory can be used to explain the mistakes that people make in reasoning. According to the theory, mistakes frequently arise when a situation requires more than one mental model. It may happen that some of these models are never formed, or get forgotten under situations of cognitive stress. For example, conjunction, exclusive disjunction and inclusive disjunction are represented by one, two and three mental models respectively[1]. Johnson-Laird et al. (1992) have confirmed that inclusive disjunction gives rise to more errors than exclusive disjunction. Khemlani et al. (2012) also demonstrated that people make more errors when reasoning about inclusive disjunction than conjunction.

Relational complexity (RC) theory complements these approaches by providing a measure of the complexity of a reasoning step. Complexity is defined "as a function … of the number of variables that can be related in a single cognitive representation" (Halford & Andrews, 2004). As an example Halford et al. (2004) note that reasoning with transitivity has an RC of 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 and hence has an RC of 3. Proponents of the theory argue that the likelihood of error in any chain of reasoning is determined by the maximum RC of the individual steps. In this study, RC theory is used to provide a measure of difficulty and enable comparison between different reasoning steps.

Besides theories of reasoning, studies of language offer useful insights. Of particular value is the concept of *implicature*, developed by Grice (1975) to describe

[1] conj: A and B; excl disj: not A and B, A and not B; incl disj: A and B, not A and B, A and not B.

a conclusion to which a speaker or writer leads an audience, but which is not a strictly logical implication of what has been said or written, e.g. "some of the students are industrious", which leads the reader to assume that not all the students are industrious. The existence of implicatures is of particular relevance in considering mental model theory. In certain situations language may lead people to form an incomplete set of mental models. This is discussed further in section 8.

4 Previous studies

In a previous study (study 1), Warren et al. (2014) identified a difficulty with functional object properties. A question requiring reasoning about a functional object property was only answered correctly by 50% of the participants, i.e. 6 out of 12. Since participants were presented with a binary choice between valid and non-valid, this is exactly equivalent to chance. The question required a reasoning step of RC 4 and it was not clear whether the difficulty was because of the complexity of this step or was a specific problem with functionality. In a subsequent study (study 2), Warren et al. (2015) compared reasoning steps of equal complexity using functional and transitive properties, which suggested that there was a problem specific to functional properties. This topic is returned to in section 6.

In addition, study 1 identified a difficulty with negated conjunction. Only 25% of participants (3 out of 12) correctly answered a question with negated conjunction, compared with 92% (11 out of 12) who correctly answered a question with negated disjunction. The difference in ability to handle negated conjunction and negated disjunction has also been observed by Khemlani et al. (2012), who interpret this in terms of the mental model theory. Negated conjunction requires three mental models (*not A and not B; not A and B; A and not B*), whereas negated disjunction requires only one model (*not A and not B*). This topic is returned to in section 7.

Study 2 also identified problems with understanding universal and existential restrictions, using the MOS keywords *only* and *some*. These difficulties were interpreted in terms of mental model theory. Table 1 shows the mental models corresponding to the two restrictions. In each case there are two models, the first of which is the more obvious and the second of which may be overlooked. This is well known for the case of *only* (Rector et al., 2004) but was found also to be the case in certain situations for *some*. This topic is returned to in sections 8 and 9.

Table 1. Mental models for universal and existential restrictions

	universal restriction	**existential restriction**
MOS	P only X	P some X
mental models	P x P ⊥	P x P x P ¬x

5 Overview of current study

The principal objective of the current study was to determine whether certain additions to MOS would lead to improved human performance with the features found difficult in the previous two studies. Consequently, the majority of questions were isomorphic to questions in the previous studies, in particular study 2.

Each question followed the same pattern as in the previous studies. A set of statements were provided, plus a putative conclusion; participants had to indicate whether the conclusion was valid or not valid. The language features used in this and the previous studies were chosen because of their relatively common use; see the discussion in Warren et al. (2014). Performance was measured by accuracy of response and time taken to respond. MediaLab[2] was used to collect the responses and measure response time. The assumption is that response time can be regarded as a proxy for difficulty. All statistical analysis was undertaken using the R statistical package (R Core Team, 2014).

The study used a simplified form of MOS, with additions as explained below. At the beginning of the study each participant was given a handout that explained the syntax used. Participants retained this handout during the study and could refer to it at any time.

The study comprised four sections, each containing eight questions. There were two variants of the study, referred to as variant 1 and variant 2, permitting some of the questions to be different in the two variants. There were 30 participants, 15 for each variant, drawn from the authors' own university, another U.K. university and an industrial research laboratory. At the beginning of the study participants were asked about their knowledge of formal logic, and of OWL or other DL formalisms. The breakdown for knowledge of formal logic was: none 3%; little 23%; some 47%; expert 27%. The breakdown for knowledge of OWL or other DL formalisms was: none 13%; little 47%; some 27%; expert 13%. Thus, the majority of the participants had a reasonable knowledge of logic but fewer had much knowledge of DL; in the latter respect the majority represented occasional users of ontologies.

Examination of the distribution of the response times revealed a positive skew, indicating a considerable deviation from normality. This phenomenon has been reported elsewhere (Blake et al., 2012; Warren et al., 2015). Further analysis suggested that the logarithmic transformation of time, selected from Tukey's ladder of powers (Scott, 2012), resulted in a distribution closer to the normal. Since ANOVA and the t-test require approximately normal populations, this transformation has been applied prior to all such tests on time data reported in this paper[3].

To prevent any bias due to question position, the order of the sections and of the questions within each section were randomized, using a randomization feature provided within MediaLab.

---

[2] Provided by Empirisoft: http://www.empirisoft.com/

[3] An alternative would have been to use a non-parametric test. Hopkins et al. (2009) note that a transformation to reduce skewness followed by a parametric test provides greater statistical power at small sample sizes than does a non-parametric test.

In reporting all statistical results, the convention adopted is the usual one of taking $p < 0.05$ as representing significance. For significant results, p is reported as being < 0.05, < 0.01, < 0.001 etc. Non-significant results are identified either by explicitly stating $p \geq 0.05$ or using the abbreviation n.s. (not significant).

6 Functional and inverse functional object properties

The motivation for this part of the study came from a comparison in study 2 between functional object properties and transitive object properties. The comparison concluded that, under conditions of equal relational complexity, the proportion of correct responses to the transitive questions was not significantly different from that for the functional questions (Fisher's Exact Test), but the latter took significantly longer ($t(87.484) = 2.2376$, $p < 0.05$), implying that the participants were finding the functional questions harder. A possible explanation for this is confusion between functionality and inverse functionality, i.e. whether it is the subject or object of the property that is unique. In this study the keyword *solely* was added after the property name and before the object to indicate that it is the object which is unique. This keyword was chosen because of its anticipated power to convey uniqueness. A possibly more natural choice, *only*, was rejected because of its existing use in MOS.

This leads to the following hypothesis:

H1 The introduction of the additional keyword *solely* between a functional property and its object will improve participant performance.

Previous studies did not investigate inverse functional properties. However, it is possible that similar difficulties will arise as with functional properties. This study investigates the effect of introducing the keyword *solely* before the subject of an inverse functional property, to indicate that in this case it is the subject which is unique. This leads to a second hypothesis:

H2 The introduction of the additional keyword *solely* before the subject of an inverse functional property will improve participant performance.

These two hypotheses are investigated in the following two subsections.

6.1 Functional object properties – comparison with study 2; hypothesis H1

In this study six questions were created isomorphic to the six functional questions in study 2, with the additional keyword, *solely*, being used as described above. Table 2 shows the six questions. For brevity F is used to represent the property. In practice, this and study 2 used the object property *has_nearest_neighbour*. Note that consecutive questions share the same axioms; but have different putative conclusions. Three reasoning steps are required to arrive at each of the valid conclusions. The table shows the RC of each step. For example, question 1 starts by using the first two axioms in an inference of RC 3 to deduce that *s* and *t* are identical. A step of RC 2 replaces *s* with *t* in axiom 3 and, finally, another inference of RC 3 concludes that *v* and *w* are identical.

Table 2. Functional object property questions

	axioms (F = *has_nearest_neighbour*)	**putative conclusion**	**validity**	**relational complexity**
1	r F solely s; r F solely t;	v sameAs w	valid	3,2,3
2	s F solely v; t F solely w	r sameAs t	not valid	
3	r F solely s; r F solely t; v F solely s;	v DifferentFrom w	valid	3,2,4
4	w F solely x; t DifferentFrom x	r DifferentFrom v	not valid	
5	r F solely s; t F solely v;	w DifferentFrom x	valid	4,2,4
6	s DifferentFrom v; w F solely r; x F solely z; t SameAs z	r DifferentFrom x	not valid	

Table 3 shows the percentage of correct responses and the mean time to respond for each question in the two studies. The data is also provided aggregated over all six questions and over the three valid questions and the three non-valid questions. For four of the six questions the percentage of correct responses was greater for the current study than for study 2; these included the two questions with the worst performance in study 2. However a Fisher's Exact test indicated no significant difference between the two studies ($p \geq 0.05$). This was also the case when the comparison was limited to the valid questions and to the non-valid questions.

For each question the mean response time was less for the current study. A two-factor ANOVA indicated that response time varied significantly between the studies ($F(1, 320) = 7.559$, $p < 0.01$) and between the valid and non-valid questions ($F(1, 320) = 4.928$, $p < 0.05$), with no significant interaction ($F(1, 320) = 1.761$, n.s.). A subsequent Tukey Honest Significant Difference (HSD) analysis revealed a significant difference in response time between the two studies for the valid questions ($p < 0.05$) but not for the non-valid.

In summary, the study partially supports hypothesis H1. The introduction of *solely* significantly reduces response time for the valid questions, although it has no significant effect on the non-valid questions, nor on the accuracy of responses.

Table 3. Functional object property questions: accuracy and response times

	study 2 – without *solely*		**current study** – with *solely*	
	%age correct N = 28	mean time (SD) – secs; N = 24	%age correct N = 30	mean time (SD) – secs; N = 30
1	75%	52 (36)	83%	39 (31)
2	96%	61 (46)	83%	50 (29)
3	61%	84 (67)	70%	58 (27)
4	79%	92 (66)	83%	78 (49)
5	43%	109 (79)	63%	73 (37)
6	71%	96 (47)	70%	90 (46)
All six questions	71%	83 (61)	76%	65 (41)
valid questions	60%	81 (67)	72%	57 (35)
non-valid questions	82%	83 (55)	79%	73 (45)

6.2 Inverse functional object properties; hypothesis H2

As neither of the previous studies included inverse functional properties, comparison of any syntactic changes could not be made between studies. Consequently, participants in this study were given two questions employing inverse functional properties. In variant 1, the questions used our simplified version of MOS. In variant 2, the keyword *solely* was included before the subject to indicate its uniqueness. Table 4 shows the format of the questions in the two variants. For brevity, I is used to represent the property. In the study *is_nearest_neighbour_of* was used. Note that questions 7 and 8 were created from questions 1 and 5 in Table 2 by restating the axioms using *is_nearest_neighbour_of* rather than its inverse *has_nearest_neighbour* and by interchanging the individual names.

Table 5 shows the results of the study. Considering the two questions aggregated, there was no significant difference in accuracy of response between the two variants (Fisher's Exact Test). In addition, a two-way ANOVA showed that there was a significant difference in response time between the questions ($F(1, 56) = 26.836$, $p < 0.00001$), reflecting their difference in complexity, but not between the two variants ($F(1, 56) = 0.640$, n.s.). There was no interaction effect ($(F1, 56) = 0.382$, n.s.).

Thus, the study offered no support for hypothesis H2. In the case of inverse functional properties, the use of the keyword *solely* has no significant effect on performance. It may be that *solely* was taken to refer to the whole <subject, predicate, object> triple without making clear the uniqueness of the subject rather than object. Other keywords and other choices of position might improve performance. One could, for example, experiment with the use of *alone* after the subject, e.g. *s alone has_nearest_neighbour r*.

Table 4. Inverse functional object property questions

	axioms (I = *is_nearest_neighbour_of)*	**putative conclusion**	**validity**	**relational complexity**
variant 1				
7	r I s; t I s; v I r; w I t	v SameAs w	valid	3,2,3
8	r I s; t I v; r DifferentFrom t; s I w; x I z; v SameAs x	w DifferentFrom z	valid	4,2,4
variant 2				
7	solely r I s; solely t I s; solely v I r; solely w I t	v SameAs w	valid	3,2,3
8	solely r I s; solely t I v; r DifferentFrom t; solely s I w; solely x I z; v SameAs x	w DifferentFrom z	valid	4,2,4

Table 5. Inverse functional object property questions: accuracy and response times

	variant 1 – without *solely*; N = 15		**variant 2** – with *solely*; N = 15	
	% corr	mean time (SD) - secs	% corr	mean time (SD) - secs
question 7	73%	38 (18)	87%	48 (23)
question 8	73%	105 (92)	73%	90 (43)
both questions	73%	72 (74)	80%	69 (40)

7 Boolean concept constructors

The eight questions in this part of the study were designed to test out whether amendments to MOS could reduce the difficulties experienced with negated conjunction. Study 1 noted that negated conjunction was significantly harder than negated disjunction, as has also been observed by Khemlani et al. (2012). At the same time, it is known that *and* and *or* in everyday language are used ambiguously, e.g. see Mendonça et al. (1998). It was thought that performance might be improved by the use of unambiguous terminology for conjunction and, for consistency, disjunction. In particular, the use of *intersection* rather than *and* may avoid the implicature, based on normal usage, that *and* represents union. This leads to the hypothesis:

H3 The use of the keyword *intersection* in place of *and* will improve performance for negated conjunction.

In some syntaxes, OWL already uses the terms *intersection* and *union* as part of prefix operators. A relevant question is how such a prefix notation compares with the infix used in MOS. The hypothesis was proposed:

H4 There will be a difference in performance between the prefix notation *IntersectionOf()* and *UnionOf()*, and the infix notation *intersection* and *union*.

Study 2 included some questions which contained the consecutive keywords *and not*. The original motivation came from the discussion of exceptions in Rector (2003), where exceptions were defined using *and not*. For this study, it was thought that *except* might be more intuitively understandable, including within nested exceptions, i.e. *and not (… and not …)*, which give rise to negated conjunction. This leads to:

H5 The use of *except* in place of *and not* will improve performance.

Table 6 shows the original questions from the two previous studies, used to generate the questions in this study.

Table 6. Boolean concept constructor questions as used in the study

	axioms	**putative conclusion**	**validity**
study 1			
1	Entity DisjointUnionOf Event, Abstract, Quality, Object; A Type Entity; A Type not (Event and Quality);	A Type (Abstract or Object)	not valid
2	Entity DisjointUnionOf Event, Abstract, Quality, Object; A Type Entity; A Type not (Event or Quality);	A Type (Abstract or Object)	valid
study 2			
3	Z EquivalentTo (TOP_CLASS and not A and not B); TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	valid
4	Z EquivalentTo (TOP_CLASS and not (A or B)); TOP_CLASS DisjointUnionOf A, B, C	Z EquivalentTo C	valid
5	Z EquivalentTo (TOP_CLASS and not (A and not A_1)); TOP_CLASS DisjointUnionOf A, B; A DisjointUnionOf A_1, A_2	Z EquivalentTo (B or A_1)	valid
6	As for question 5	Z EquivalentTo B	not valid
7	Z EquivalentTo (TOP_CLASS and not (A and not (A_1 and not A_1_X))); TOP_CLASS DisjointUnionOf A, B; A DisjointUnionOf A_1, A_2; A_1 DisjointUnionOf A_1_X, A_1_Y	Z EquivalentTo (B or A_1_Y)	valid
8	As for question 7	Z EquivalentTo A_1_Y	not valid

Table 7. Boolean concept constructor questions: accuracy and response times

	%age correct	mean times (SD) - secs	%age correct	mean time (SD) - secs	%age correct	mean time (SD) - secs
	study 1 *and, or, not*		**current study variant 1** *intersection, union, not*		**current study variant 2** *IntersectionOf, UnionOf, not*	
	N = 12	N = 12	N = 15	N = 15	N = 15	N = 15
1	25%	75 (48)	80%	53 (41)	67%	47 (16)
2	92%	44 (19)	87%	42 (28)	100%	39 (25)
	study 2 *and, or, not*		**current study variant 1** *intersection, union, except*		**current study variant 2** *IntersectionOf, UnionOf, not*	
	N = 28	N = 24	N = 15	N = 15	N = 15	N = 15
3	82%	39 (26)	100%	39 (25)	100%	47 (30)
4	86%	43 (29)	100%	35 (26)	93%	46 (35)
5	61%	96 (56)	53%	61 (37)	53%	65 (38)
6	64%	105 (78)	100%	44 (25)	73%	82 (57)
7	54%	90 (48)	60%	97 (75)	40%	156 (126)
8	68%	94 (47)	80%	88 (60)	60%	93 (51)
Q1 and 2	58%	60 (39)	83%	47 (35)	83%	43 (21)
Q3 to 8	69%	78 (56)	82%	61 (50)	70%	82 (74)
Q1 to 8	68%	75 (54)	82%	57 (47)	73%	72 (67)

In the case of study 1, the two questions originally made use of an ontology pattern. In the table, the essence of the questions has been extracted out. In this study, variant 1 consists of questions isomorphic to the eight questions shown, with *and* replaced by *intersection*, *or* replaced by *union*, and in the case of questions 3 to 8, *and not* replaced by *except*. Variant 2 consists of questions isomorphic to the questions shown, with *and* replaced by the prefix form *IntersectionOf()*, and *or* replaced by the prefix form *UnionOf()*. As a result, H3 can be tested by comparing performance on variant 1 of question 1 with performance on the analogous question in study 1. Question 2 was included to determine that the change in terminology did not have a deleterious effect on negated disjunction. H4 can be tested by comparing the two variants of this study, using questions 1 and 2. Finally, H5 can be tested by comparing the two variants of the study, using questions 3 to 8. Table 7 shows the relevant data.

7.1 Negated conjunction and disjunction; hypotheses H3

Table 7 shows that question 1 was answered much more accurately in variant 1 of this study, with the use of *intersection*, than in study 1 which used *and* ($p < 0.01$, Fisher's Exact Test). However, there was no significant difference in response times ($t(23.988) = 1.6031$, n.s.). Thus, hypothesis H3 is supported with regard to accuracy, but not response time. For question 2, with negated disjunction, there was no significant difference in accuracy between variant 1 and study 1 ($p \geq 0.05$, Fisher's Exact Test); nor was there a significant difference in response time ($t(24.433) = 0.5372$, n.s.). Thus, the change in notation had no effect on negated disjunction.

7.2 Use of prefix notation; hypothesis H4

The two variants of this study enable a comparison of infix and prefix notation. However, if all questions were used this comparison would be confounded with the effect

of using *except* in variant 1. For a more controlled comparison, only questions 1 and 2 are used. Taking these two questions together, the percentage of correct responses is the same for both variants. Moreover, there was no significant difference in response time between the two variants ($t(56.443) = 0.21532$, n.s.). In summary, there is no evidence of a difference in performance between infix and prefix notation.

7.3 Use of *except* in place of *and not*; hypothesis H5

Questions 3 to 8 in variant 1 were intended to test the effect of replacing *and not* with *except*. For technical reasons concerned with the fact that the order of questions in study 2 was not fully randomized, it is not possible to compare study 2 with variant 1 of this study. Instead the two variants of this study are compared on the assumption, supported by the evidence of the previous subsection, that the use of infix and prefix notation makes no significant difference. For questions 3 to 8 aggregated, there was no significant difference in accuracy ($p \geq 0.05$, Fisher's Exact Test). However, there was a significant difference in response time ($t(176.22) = 2.3962$, $p < 0.05$), with variant 1, using *except*, having the lower response time. Thus, hypothesis H5 is not supported in respect of accuracy but is supported in respect of time.

8 Negation and restriction

As already noted, difficulties with the universal and existential restrictions may in part be due to a failure to form both the required mental models. This part of the study investigated whether the replacement of *only* with *noneOrOnly* and *some* with *including* would improve performance with these restrictions. *noneOrOnly* was intended to draw attention to the fact that, e.g. the class *has_child noneOrOnly MALE* includes those individuals who have no children at all, i.e. to avoid the implicature that *only MALE* suggests the existence of some male child, and thereby help participants to form both the necessary mental models, as shown in Table 1. *including* was intended to draw attention to the fact that, e.g. the class *has_child including MALE* may contain individuals who have a female child in addition to a male one, i.e. again to help participants form both the mental models. The hypothesis to be investigated is:

H6 The use of *noneOrOnly* in place of *only* and *including* in place of *some* will lead to improved participant performance.

In addition, it was thought that the use of *not ... some* might read unnaturally, and that *not ... any* corresponds to more normal English usage; this led to the hypothesis:

H7 The use of *any* to indicate the existential restriction, when the corresponding property is preceded by a negation, will lead to improved performance.

The original questions, as in study 2, are shown in Table 8. Each question is comprised of two axioms. For the first axiom there are four variants; for the second axiom two variants. All questions have the same putative conclusion. In Table 8, two different typefaces are used for the first axioms to indicate semantic equivalence. The questions can be grouped into four semantically equivalent pairs: {1, 4}; {2, 3}; {5, 8}; {6, 7}. For this study these questions were modified by the replacements described above. For six of the questions there was no difference between the two variants. However, in

their original form questions 3 and 7 contain the class description *not (has_child some MALE)*. In variant 1 *some* was replaced with *including* as in the other questions. In variant 2 *any* was used, i.e. *not (has_child any MALE).* Table 9 shows the accuracy and response times for the questions in study 2 and the current study. For the latter, separate data are shown for the two variants when the questions differ.

Table 8. Questions employing negation and restrictions; form as in study 2. N.B. the putative conclusion in each case was *X DisjointWith Y.*

	first axiom	**second axiom**	**validity**
1	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child only MALE	valid
2	X SubClassOf has_child only (not MALE)		not valid
3	X SubClassOf not (has_child some MALE)		not valid
4	***X SubClassOf not (has_child only MALE)***		valid
5	***X SubClassOf has_child some (not MALE)***	Y SubClassOf has_child some MALE	not valid
6	X SubClassOf has_child only (not MALE)		valid
7	X SubClassOf not (has_child some MALE)		valid
8	***X SubClassOf not (has_child only MALE)***		not valid

8.1 *noneOrOnly* and *including*; hypothesis H6

To avoid the confounding effect of the introduction of *not ... any* in variant 2 for questions 3 and 7, an analysis was conducted based on the six questions excluding questions 3 and 7. Table 9 shows the mean results for these six questions. The use of *noneOrOnly* and *including* led to a significant increase in accuracy (Fisher's Exact Test, $p < 0.01$) and reduction in response time ($t(284.57) = 2.7897$, $p < 0.01$), supporting H6. This suggests that the new keywords do support the creation of the two mental models necessary for each of the restrictions. It is also noteworthy that there is an appreciable increase in accuracy for the two questions which were answered worst in study 2, i.e. questions 2 and 6. The former requires the second model for the universal restriction.

Table 9. Negation and restriction questions: accuracy and response times.

	study 2 *only, some*		**current study: overall;** N = 30		**current study: variant 1;** N = 15		**current study: variant 2;** N = 15	
			noneOrOnly, including				*not ... any*	
	%age correct N = 28	mean time (SD) – secs N = 24	%age correct	mean time (SD) – secs	%age correct	mean time (SD) – secs	%age correct	mean time (SD) – secs
1	61%	52 (39)	80%	42 (33)				
2	50%	33 (18)	73%	29 (20)				
3	68%	45 (22)	70%	69 (127)	67%	55 (49)	73%	84 (175)
4	75%	43 (25)	90%	41 (32)				
5	64%	41 (30)	70%	30 (21)				
6	50%	44 (40)	70%	33 (33)				
7	79%	43 (37)	80%	29 (16)	73%	26 (14)	87%	33 (18)
8	68%	60 (37)	67%	38 (24)				
Exc Q3 and 7	61%	45 (33)	75%	35 (28)				
Q3 and 7	73%	44 (30)	75%	49 (92)	70%	40 (38)	80%	58 (125)

8.2 *not … any*; hypothesis H7

Questions 3 and 7 provide an opportunity to investigate the effect of using *not ... any* in variant 2. There was no support for H7, i.e. no significant difference between the variants, in accuracy ($p \geq 0.05$, Fisher's Exact Test) or time ($t(57.832) = 0.79496$, n.s.).

9 Nested restrictions

Study 2 included eight questions making use of nested restrictions, as shown in Table 10; Table 11 shows the associated data. Analogous questions in variant 1 of this study were created by replacing *only* with *noneOrOnly* and *some* with *including*, enabling a further investigation of hypothesis H6. For variant 2, questions 5 to 8 were as for variant 1, except that in the final axiom, *not ... some* in study 2 has been replaced by *not ... any*, enabling a further investigation of hypothesis H7

Table 10. Questions employing nested restriction; form as in study 2. N.B. the putative conclusion in each case was *a Type (not X)*.

	first axiom(s)	**final axiom**	**validity**
1	X SubClassOf (has_child some (has_child some FEMALE))	a has_child b; b Type has_child some (not FEMALE)	not valid
2	X SubClassOf has_child some Y; Y EquivalentTo has_child only FEMALE		not valid
3	X SubClassOf (has_child only (has_child some FEMALE))		not valid
4	X SubClassOf has_child only Y; Y EquivalentTo has_child only FEMALE		valid
5	X SubClassOf has_child some Y; Y EquivalentTo has_child some FEMALE	a has_child b; b Type (not (has_child some FEMALE))	not valid
6	X SubClassOf (has_child some (only FEMALE))		not valid
7	X SubClassOf has_child only Y; Y EquivalentTo has_child some FEMALE		valid
8	X SubClassOf (has_child only (has_child only FEMALE))		not valid

Table 11. Nested restriction questions: accuracy and response times

	study 2		**current study: var 1**		**current study: var 2**	
	%age correct	mean time (SD) - secs	%age correct	mean time (SD) - secs	%age correct	mean time (SD) - secs
	N = 28	N = 24	N = 15	N = 15	N = 15	N = 15
	only, some		*noneOrOnly, including*			
1	71%	69 (45)	80%	47 (30)	60%	73 (59)
2	57%	79 (53)	40%	65 (20)	67%	68 (41)
3	71%	63 (43)	60%	54 (31)	53%	79 (46)
4	57%	63 (39)	53%	100 (87)	47%	66 (49)
					not ... any	
5	54%	88 (62)	40%	64 (30)	67%	74 (67)
6	64%	73 (45)	73%	85 (82)	73%	99 (90)
7	71%	80 (36)	53%	64 (39)	47%	97 (102)
8	50%	55 (30)	60%	83 (35)	53%	63 (37)
Mean for all questions	62%	71 (45)	58%	70 (51)		
Mean for Q5 to 8	60%	74 (46)	57%	74 (51)	60%	83 (77)

9.1 *noneOrOnly* and *including*; hypothesis H6

For the eight questions aggregated there was no significant difference between study 2 and variant 1, neither in accuracy ($p \geq 0.05$, Fisher's Exact Test) nor in response time ($t(288.79) = 0.40724$, n.s.). Thus, unlike the questions discussed in subsection 8.1, these questions offered no support for hypothesis H6.

9.2 *not … any*; hypothesis H7

For questions 5 to 8, there was no significant difference in accuracy ($p \geq 0.05$, Fisher's Exact Test) between the two variants of this study, nor in response time ($t(106.86) = 0.19408$, n.s.), i.e. as with subsection 8.2, there was no support for H7.

10 Conclusions and future work

Table 12. Summary of findings

Hypothesis (results section shown in brackets)	**Accuracy**	**Response time**
H1 – *solely* with functional properties (6.1)	no advantage	advantage for valid questions
H2 – *solely* with inverse funct. properties (6.2)	no advantage	
H3 – *intersection* in place of *and* (7.1)	advantage	no advantage
H4 – prefix versus infix notation (7.2)	no difference	
H5 – *except* in place of *and not* (7.3)	no advantage	advantage
H6 – *noneOrOnly* and *including* (8.1, 9.1)	advantage with single restriction but not with nested restrictions	
H7 – *not ... any* (8.2, 9.2)	no advantage	

Table 12 summarizes the findings from the study. Based on the empirical evidence, the following extensions to MOS can be recommended: *intersection* in place of *and*, avoiding the associated ambiguity; *except* as an alternative to *and not*, providing a natural way to think about exceptions; *solely* with functional properties, identifying that the object of the property is unique. Whilst *noneOronly* and *including* improve reasoning in some cases, more research is needed to investigate under what circumstances these keywords are beneficial, and whether alternatives might be preferable. Research is also needed to improve performance with inverse functional properties.

This work has shown how theory can be used to guide language development. An understanding of the mental models associated with logical constructs can help choose keywords which emphasize all the models, as was done here with the universal and existential restrictions. An understanding of the implicatures present in natural language can help avoid the use of words which, seemingly user friendly, are ambiguous or even misleading, as is the case for *and*. However, when ambiguities and misleading implicatures are avoided, the use of natural language can aid human reasoning, as was shown with *except*. More generally, it is proposed that theories of reasoning and language will be able to provide support in the development of a range of computer languages.

Acknowledgements

The authors would like to thank all the study participants, and in particular Dr. Gem Stapleton, of Brighton University, and Dr. John Davies, of BT Research, for facilitating experimental sessions at their respective institutions.

References

1. Blake, A., Stapleton, G., Rodgers, P., Cheek, L., & Howse, J. (2012). Does the orientation of an Euler diagram affect user comprehension? In *DMS* (pp. 185–190).
2. Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan, *Syntax and Semantics, Volume 3: Speech Acts* (pp. 41–58).
3. Halford, G. S., & Andrews, G. (2004). : The development of deductive reasoning: How important is complexity? *Thinking & Reasoning*, *10*(2), 123–145.
4. Hopkins, W., Marshall, S., Batterham, A., & Hanin, J. (2009). Progressive statistics for studies in sports medicine and exercise science. *Medicine+ Science in Sports+ Exercise*, *41*(1), 3
5. Horridge, M., Bail, S., Parsia, B., & Sattler, U. (2011). The cognitive complexity of OWL justifications. *The Semantic Web–ISWC 2011*, 241–256.
6. Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., & Wang, H. H. (2006). The manchester owl syntax. *OWL: Experiences and Directions*.
7. Johnson-Laird, P. N. (2010). Against logical form. *Psychologica Belgica*, *50*(3), 193–221.
8. Johnson-Laird, P. N., Byrne, R. M., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*(3), 418.
9. 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. http://mindmodeling.org/cogsci2012/papers/0110/paper0110.pdf
10. Mendonça, E. A., Cimino, J. J., Campbell, K. E., & Spackman, K. A. (1998). Reproducibility of interpreting‘ and’ and‘ or’ in terminology systems. In *Proceedings of the AMIA Symposium* (p. 790). American Medical Informatics Association.
11. Nguyen, Power, Piwek, & Williams. (2012). Measuring the understandability of deduction rules for OWL. Presented at the First international workshop on debugging ontologies and ontology mappings, Galway, Ireland.
12. R Core Team. (2014). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*. ISBN 3-900051-07-0.
13. Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., … Wroe, C. (2004). OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. *Engineering Knowledge in the Age of the Semantic Web* (pp. 63–81). Springer.
14. Rector, A. L. (2003). Defaults, context, and knowledge: Alternatives for OWL-indexed knowledge bases. In *Pacific Symposium on Biocomputing* (pp. 226–237).
15. Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, 90(1), 38.
16. Scott, D. (2012). Tukey’s ladder of powers. *Rice University*. http://onlinestatbook.com/2/transformations/tukey.html
17. Warren, P., Mulholland, P., Collins, T., & Motta, E. (2015). Making sense of description logics. *Proceedings of the 11th International Conference on Semantic Systems* (pp. 49–56). ACM.
18. Warren, P., Mulholland, P., Collins, T., & Motta, E. (2014). The usability of Description Logics: understanding the cognitive difficulties presented by Description Logics (pp. 550–564). Presented at the ESWC 2014, Crete: Springer.