The Cognitive Complexity of OWL Justifications

Matthew Horridge, Samantha Bail, Bijan Parsia, Ulrike Sattler

The University of Manchester
Oxford Road, Manchester, M13 9PL
{matthew.horridge|bails|bparsia|sattler@cs.man.ac.uk}

Abstract. In this paper, we present an approach to determining the cognitive complexity of justifications for entailments of OWL ontologies. We introduce a simple cognitive complexity model and present the results of validating that model via experiments involving OWL users. The validation is based on test data derived from a large and diverse corpus of naturally occurring justifications. Our contributions include validation for the cognitive complexity model, new insights into justification complexity, a significant corpus with novel analyses of justifications suitable for experimentation, and an experimental protocol suitable for model validation and refinement.

1 Introduction

A justification is a minimal subset of an ontology that is sufficient for an entailment to hold. More precisely, given $\mathcal{O} \models \eta$, $\mathcal{J}$ is a justification for η in $\mathcal{O}$ if $\mathcal{J} \subseteq \mathcal{O}$, $\mathcal{J} \models \eta$ and, for all $\mathcal{J}' \subsetneq \mathcal{J}$, $\mathcal{J}' \not\models \eta$. Justifications are the dominant form of explanation in OWL,[1] and justification based explanation is deployed in popular OWL editors. The primary focus of research in this area has been on explanation for the sake of debugging problematic entailments [8], whether standard "buggy" entailments, such as class unsatisfiability or ontology inconsistency, or user selected entailments such as arbitrary subsumptions and class assertions. The debugging task is naturally directed toward "repairing" the ontology and the use of "standard errors" further biases users toward looking for problems in the logic of a justification.

As a form of explanation, justifications are a bit atypical historically. While they present the ultimate, ontology specific reasons that a given entailment holds, they, unlike proofs, do not *articulate* how those reasons support the entailment, at least, in any detail. That is, they correspond to the *premises* of a proof, but do not invoke any specific proof calculus. Clearly, this brings advantages, as justifications are calculus independent, require nothing more than knowledge of OWL, and do not involve a host of knotty, unresolved issues of long standing (such as what to do about "obvious" steps [2]). Furthermore, justifications are highly manipulable: Deleting an axiom breaks the entailment, which allows for a very active, ontology related form of experimentation by users. However, in spite of their field success, justifications are held to be lacking *because* they don't articulate the connection and thus are too hard to understand.[2]

[1] Throughout this paper, "OWL" refers to the W3C's Web Ontology Language 2 (OWL 2).

[2] See, for example, a related discussion in the OWL Working Group http://www.w3.org/2007/OWL/tracker/issues/52. Also, in [1], the authors rule out of court justifications

The Description Logic that underpins OWL, $\mathcal{SROIQ}$, is N2ExpTime-complete [10], which suggests that even fairly small justifications could be quite challenging to reason with. However, justifications are highly successful in the field, thus the computational complexity argument is not dispositive. We do observe often that certain justifications are difficult and frustrating to understand for ontology developers. In some cases, the difficulty is obvious: a large justification with over 70 axioms is going to be at best cumbersome however simple its logical structure. However, for many reasonably sized difficult justifications (e.g. of size 10 or fewer axioms) the source of cognitive complexity is not clearly known.

If most naturally occurring justifications are easy "enough" to understand, then the need for auxilliary explanation faculties (and the concomitant burden on the user to master them and the tool developer to provide them) is reduced. In prior work [5,3,6], we proposed a predictive complexity model based on an exploratory study plus our own experiences and intuitions. However, in order to deploy this metric reliably, whether to assess the state of difficulty of justifications or to deploy an end-user tool using it, the model needed validation.

In this paper, we present the results of several experiments into the cognitive complexity of OWL justifications. Starting from our cognitive complexity model, we test how well the model predicts error proportions for an entailment assessment task. We find that the model does fairly well with some notable exceptions. A follow-up study with an eye tracker and think aloud protocol supports our explanations for the anomalous behaviour and suggests both a refinement to the model and a limitation of our experimental protocol.

Our results validate the use of justifications as the primary explanation mechanism for OWL entailments as well as raising the bar for alternative mechanisms (such as proofs). Furthermore, our metric can be used to help users determine when they need to seek expert help or simply to organise their investigation of an entailment.

2 Cognitive Complexity & Justifications

While there have been several user studies in the area of debugging [11,9], ontology engineering anti-patterns [16], and our exploratory investigation into features that make justifications difficult to understand [5], to the best of our knowledge there have not been any formal user studies that investigate the cognitive complexity of justifications.

Of course, if we had a robust theory of how people reason, one aspect of that robustness would be an explanation of justification difficulty. However, even the basic mechanism of human deduction is not well understood. In psychology, there is a long standing rivalry between two accounts of human deductive processes: (1) that people apply inferential rules [15], and (2) that people construct mental models [7].[3] In spite of a voluminous literature (including functional MRI studies, e.g., [14]), to date there is no scientific consensus [13], even for propositional reasoning.

as a form of explanation: "It is widely accepted that an explanation corresponds to a formal proof. A formal proof is constructed from premises using rules of inference".

[3] (1) can be crudely characterised as people use a natural deduction proof system and (2) as people use a semantic tableau.

Even if this debate were settled, it would not be clear how to apply it to ontology engineering. The reasoning problems that are considered in the literature are quite different from understanding how an entailment follows from a justification in a (fairly expressive) fragment of first order logic. For example, our reasoning problems are in a regimented, formalised language for which reasoning problems are far more constrained than deduction "in the wild." Thus, the artificiality of our problems may engage different mechanisms than more "natural" reasoning problems: e.g. even if mental models theory were correct, people *can* produce natural deduction proofs and might find that doing so allows them to outperform "reasoning natively". Similarly, if a tool gives me a justification, I can use my knowledge of justifications to help guide me, e.g., that justifications are minimal means that I must look at all the axioms presented and I do not have to rule any out as irrelevant. As we will see below, such meta-justificatory reasoning is quite helpful.

However, for ontology engineering, we do not need a *true account of human deduction*, but just need a way to determine how *usable* justifications are for our tasks. In other words, what is required is a theory of the *weak cognitive complexity* of justifications, not one of *strong cognitive complexity* [17].

A similar practical task is generating sufficiently difficult so-called "Analytical Reasoning Questions" (ARQs) problems in Graduate Record Examination (GRE) tests. ARQs typically take the form of a "logic puzzle" wherein an initial setup is presented, along with some constraints, then the examinee must determine possible solutions. Often, these problems involve positioning entities in a constrained field (e.g., companies on floors in a building, or people seated next to each other at dinner). In [13], the investigators constructed and validated a model for the complexity of answering ARQs via experiments with students. Analogously, we aim to validate a model for the complexity of "understanding" justifications via experiments on modellers.

In [13], Newstead et al first build a preliminary complexity model, as we did, based on a small but intense pilot study using think aloud plus some initial ideas about the possible sources of complexity. Then they validated their model in a series of large scale controlled experiments wherein a set of students were given sets of questions which varied systematically in complexity (according to their model) and in particular features used. One strong advantage Newstead el al have is that the problems they considered are very constrained and comparatively easy to analyse. For example, the form of ARQ question they consider have finite, indeed, easily enumerable, sets of models. Thus, they can easily determine how many possible situations are ruled out by a given constraint which is a fairly direct measure of the base line complexity of the problem. Similarly, they need merely to *construct* problems of the requisite difficulty, whereas we need to *recognise* the difficulty of arbitrary inputs. Finally, their measure of difficulty is exactly what proportion of a given cohort get the questions right, whereas we are dealing with a more nebulous notion of understanding.

Of course, the biggest advantage is that their problems are expressed in natural language and reasonably familiar to millions of potential participants, whereas our investigations necessarily require a fair degree of familiarity with OWL — far more than can be given in a study-associated training session. Nevertheless, the basic approach seems quite sound and we follow it in this paper.

3 A Complexity Model

We have developed a cognitive complexity model for justification understanding. This model was derived partly from observations made during an exploratory study (see [5,3,6] for more details) in which people attempted to understand justifications from naturally occurring ontologies, and partly from intuitions on what makes justifications difficult to understand.

Please note that reasonable people may (and do!) disagree with nigh every aspect of this model (the weights are particularly suspect). For each factor, we have witnessed the psychological reality of their causing a reasonably sophisticated user difficulty in our exploratory study. But, for example, we cannot warrant their orthogonality, nor can we show that some combinations of factors is easier than the sum of the weights would indicate. This should not be too surprising especially if one considers the current understanding of what makes even propositional formulae difficult for automated reasoners. While for extremely constrained problems (such as propositional kCNF), we have long had good predictive models for reasoning difficulty for key proving techniques, more unconstrained formulae have not been successfully analysed. Given that the complexity of a given algorithm is intrinsically more analysable than human psychology (consider simply the greater ease of controlled experiments), the fact that we do not have good predictive models for automated reasoners should be a warning for theorists of cognitive complexity. However, while daunting, these facts do not mean we should give up, as even a fairly crude model can be useful, as we have found. Furthermore, we can hope to improve the predictive validity of this model, even without determining the structure of the phenomena.

Table 1 describes the model, wherein $\mathcal{J}$ is the justification in question, η is the focal entailment, and each value is multiplied by its weight and then summed with the rest. The final value is a complexity score for the justification. Broadly speaking, there are two types of components: (1) structural components, such as **C1**, which require a syntactic analysis of a justification, and (2) semantic components, such as **C4**, which require entailment checking to reveal non-obvious phenomena.

Components **C1** and **C2** count the number of different kinds of axiom types and class expression types as defined in the OWL 2 Structural Specification.[4] The more diverse the basic logical vocabulary is, the less likely that simple pattern matching will work and the more "sorts of things" the user must track.

Component **C3** detects the presence of universal restrictions where *trivial satisfaction* can be used to infer subsumption. Generally, people are often surprised to learn that if $\langle x, y\rangle \notin R^{\mathcal{I}}$ for all $y \in \Delta^{\mathcal{I}}$, then $x \in (\forall R.C)^{\mathcal{I}}$. This was observed repeatedly in the exploratory study.

Components **C4** and **C5** detect the presence of synonyms of $\top$ and $\bot$ in the signature of a justification where these synonyms are *not explicitly* introduced via subsumption or equivalence axioms. In the exploratory study, participants failed to spot synonyms of $\top$ in particular.

Component **C6** detects the presence of a domain axiom that is not paired with an (entailed) existential restriction along the property whose domain is restricted. This

[4] `http://www.w3.org/TR/owl2-syntax/`

Table 1. A Simple Complexity Model

Name	Base value	Weight
C1 AxiomTypes	Number of axiom types in $\mathcal{J}$ & η.	100
C2 ClassConstructors	Number of constructors in $\mathcal{J}$ & η.	10
C3 UniversalImplication	If an $\alpha \in \mathcal{J}$ is of the form $\forall R.C \sqsubseteq D$ or $D \equiv \forall R.C$ then 50 else 0.	1
C4 SynonymOfThing	If $\mathcal{J} \models \top \sqsubseteq A$ for some $A \in \mathsf{Signature}(\mathcal{J})$ and $\top \sqsubseteq A \notin \mathcal{J}$ and $\top \sqsubseteq A \neq \eta$ then 50 else 0.	1
C5 SynonymOfNothing	If $\mathcal{J} \models A \sqsubseteq \bot$ for some $A \in \mathsf{Signature}(\mathcal{J})$ and $A \sqsubseteq \bot \notin \mathcal{J}$ and $A \sqsubseteq \bot \neq \eta$ then 50 else 0.	1
C6 Domain&NoExistential	If $\mathsf{Domain}(R, C) \in \mathcal{J}$ and $\mathcal{J} \not\models E \sqsubseteq \exists R.\top$ for some class expressions E then 50 else 0.	1
C7 ModalDepth	The maximum modal depth of all class expressions in $\mathcal{J}$.	50
C8 SignatureDifference	The number of distinct terms in $\mathsf{Signature}(\eta)$ not in $\mathsf{Signature}(\mathcal{J})$.	50
C9 AxiomTypeDiff	If the axiom type of η is not the set of axiom types of $\mathcal{J}$ then 50 else 0	1
C10 ClassConstructorDiff	The number of class constructors in η not in the set of constructors of $\mathcal{J}$.	1
C11 LaconicGCICount	The number of General Concept Inclusion axioms in a laconic version of $\mathcal{J}$	100
C12 AxiomPathLength	The number of maximal length expression paths in $\mathcal{J}$ plus the number of axioms in $\mathcal{J}$ which are not in some maximal length path of $\mathcal{J}$, where a *class (property) expression subsumption path* is a list of axioms of length n where for any $1 \leq i < n$, the axiom at position i is $C_i \sqsubseteq C_{i+1}$.	10

typically goes against peoples' expectations of how domain axioms work, and usually indicates some kind of non-obvious reasoning by cases. For example, given the two axioms $\exists R.\top \sqsubseteq C$ and $\forall R.D \sqsubseteq C$, the domain axiom is used to make a statement about objects that have R successors, while the second axiom makes a statement about those objects that do not have any R successors to imply that C is equivalent to $\top$. This is different from the typical pattern of usage, for example where $A \sqsubseteq \exists R.C$ and $\exists R.\top \sqsubseteq B$ entails $A \sqsubseteq B$.

Component **C7** measures maximum modal depth of sub-concepts in $\mathcal{J}$, which tend to generate multiple distinct but interacting propositional contexts.

Component **C8** examines the signature difference from entailment to justification. This can indicate confusing redundancy in the entailment, or synonyms of $\top$, that may not be obvious, in the justification. Both cases are surprising to people looking at such justifications.

Components **C9** and **C10** determine if there is a difference between the type of, and types of class expressions in, the axiom representing the entailment of interest and the types of axioms and class expressions that appear in the justification. Any difference can indicate an extra reasoning step to be performed by a person looking at the justification.

Component **C11** examines the number of subclass axioms that have a complex left hand side in a *laconic*[5] version of the justification. Complex class expressions on the left hand side of subclass axioms in a laconic justification indicate that the conclusions of several intermediate reasoning steps may interact.

Component **C12** examines the number of obvious syntactic subsumption paths through a justification. In the exploratory study, participants found it very easy to quickly read chains of subsumption axioms, for example, $\{A \sqsubseteq B, B \sqsubseteq C, C \sqsubseteq D, D \sqsubseteq E\}$ to entail $A \sqsubseteq E$. This complexity component essentially increases the complexity when these kinds of paths are lacking.

The weights were determined by rough and ready empirical twiddling, without a strong theoretical or specific experimental backing. They correspond to our sense, esp. from the exploratory study, of sufficient reasons for difficulty.

4 Experiments

While the model is plausible and has behaved reasonably well in applications, its validation is a challenging problem. In principle, the model is reasonable if it successfully predicts the difficulty an arbitrary OWL modeller has with an arbitrary justification sufficiently often. Unfortunately, the space of ontology developers and of OWL justifications (even of existing, naturally occurring ones) is large and heterogeneous enough to be difficult to randomly sample.

4.1 Design Challenges

To cope with the heterogeneity of users, any experimental protocol should require minimal experimental interaction, i.e. it should be executable over the internet from subjects' own machines with simple installation. Such a protocol trades access to subjects, over time, for the richness of data gathered. To this end, we adapted one of the experimental protocols described in [13] and tested it on a more homogeneous set of participants—a group of MSc students who had completed a lecture course on OWL. These students had each had an 8 hour lecture session, once a week, for five weeks on OWL and ontology engineering, and had completed 4 weeks of course work including having constructed several ontologies. The curriculum did not include any discussion of justifications or explanation per se, though entailment and reasoning problems had been covered.[6] Obviously, this group is not particularly representative of all OWL ontologists: They are young, relatively inexperienced, and are trained in computer science. However, given their inexperience, especially with justifications, things they find easy should be reliably easy for most trained users.

While the general experimental protocol in [13] seems reasonable, there are some issues in adapting it to our case. In particular, in ARQs there is a restricted space of possible (non-)entailments suitable for multiple choice questions. That is, the wrong answers can straightforwardly be made plausible enough to avoid guessing. A justification inherently has one statement for which it is a justification (even though it will

[5] Laconic justifications [4] are justifications whose axioms do not contain any superfluous parts.

[6] See `http://www.cs.manchester.ac.uk/pgt/COMP60421/` for course materials.

be a minimal entailing subset for others). Thus, there isn't a standard "multiple set" of probable answers to draw on. In the exam case, the primary task is successfully answering the question and the relation between that success and predictions about the test taker are outside the remit of the experiment (but there is an established account, both theoretically and empirically). In the justification case the standard primary task is "understanding" the relationship between the justification and the entailment. Without observation, it is impossible to distinguish between a participant who really "gets" it and one who merely acquiesces. In the exploratory study we performed to help develop the model, we had the participant rank the difficulty of the justification, but also used think aloud and follow-up questioning to verify the success in understanding by the participant. This is obviously not a minimal intervention, and requires a large amount of time and resources on the part of the investigators.

To counter this, the task was shifted from a justification understanding task to something more measurable and similar to the question answering task in [13]. In particular, instead of presenting the justification/entailment pair *as* a justification/entailment pair and asking the participant to try to "understand" it, we present the justification/entailment pair as a set-of-axioms/candidate-entailment pair and ask the participant to *determine* whether the candidate is, in fact, entailed. This diverges from the standard justification situation wherein the modeller knows that the axioms entail the candidate (and form a justification), but provides a metric that can be correlated with cognitive complexity: *error proportions*.

4.2 Justification Corpus

To cope with the heterogeneity of justifications, we derived a large sample of justifications from ontologies from several well known ontology repositories: The Stanford BioPortal repository[7] (30 ontologies plus imports closure), the Dumontier Lab ontology collection[8] (15 ontologies plus imports closure), the OBO XP collection[9] (17 ontologies plus imports closure) and the TONES repository[10] (36 ontologies plus imports closure). To be selected, an ontology had to (1) entail one subsumption between class names with at least one justification that (a) was not the entailment itself, and (b) contains axioms in that ontology (as opposed to the imports closure of the ontology), (2) be downloadable and loadable by the OWL API (3) processable by FaCT++.

While the selected ontologies cannot be said to generate a *truly representative* sample of justifications from the full space of possible justifications (even of those on the Web), they are diverse enough to put stress on many parts of the model. Moreover, most of these ontologies are actively developed and used and hence provide justifications that a significant class of users encounter.

For each ontology, the class hierarchy was computed, from which direct subsumptions between class names were extracted. For each direct subsumption, as many justifications as possible in the space of 10 minutes were computed (typically all justifications; time-outs were rare). This resulted in a pool of over 64,800 justifications.

[7] `http://bioportal.bioontology.org`
[8] `http://dumontierlab.com/?page=ontologies`
[9] `http://www.berkeleybop.org/ontologies/`
[10] `http://owl.cs.manchester.ac.uk/repository/`

While large, the actual logical diversity of this pool is considerably smaller. This is because many justifications, for different entailments, were of exactly the same "shape". For example, consider $\mathcal{J}_1 = \{A \sqsubseteq B, B \sqsubseteq C\} \models A \sqsubseteq C$ and $\mathcal{J}_2 = \{F \sqsubseteq E, E \sqsubseteq G\} \models F \sqsubseteq G$. As can be seen, there is an injective renaming from $\mathcal{J}_1$ to $\mathcal{J}_2$, and $\mathcal{J}_1$ is therefore *isomorphic* with $\mathcal{J}_2$. If a person can understand $\mathcal{J}_1$ then, with allowances for variations in name length, they should be able to understand $\mathcal{J}_2$. The initial large pool was therefore reduced to a smaller pool of 11,600 *non-isomorphic justifications*.

4.3 Items and Item Selection

Each experiment consists of a series of test items (questions from a participant point of view). A test *item* consists of a *set of axioms*, one *following axiom*, and a *question*, "Do these axioms entail the following axiom?". A participant *response* is one of five possible answers: "Yes" (it is entailed), "Yes, but not sure", "Not Sure", "No, but not sure", "No" (it is not entailed). From a participant point of view, any item may or may not contain a justification. However, in our experiments, every item was, in fact, a justification.

It is obviously possible to have non-justification entailing sets or non-entailing sets of axioms in an item. We chose against such items since (1) we wanted to maximize the number of actual justifications examined (2) justification understanding is the actual task at hand, and (3) it is unclear how to interpret error rates for non-entailments in light of the model. For some subjects, esp. those with little or no prior exposure to justifications, it was unclear whether they understood the difference between the set merely being entailing, and it being minimal and entailing. We did observe one person who made use of this metalogical reasoning in the follow-up study.

Item Construction: For each experiment detailed below, test items were constructed from the pool of 11,600 non-isomorphic justifications. First, in order to reduce variance due primarily to size, justifications whose size was less than 4 axioms and greater than 10 axioms were discarded. This left 3199 (28%) justifications in the pool. In particular, this excluded large justifications that might require a lot of reading time, cause fatigue problems, or intimidate, and excluded very small justifications that tended to be trivial.[11]

For each justification in the pool of the remaining 3199 non-isomorphic justifications, the complexity of the justification was computed according to the model presented in Table 1, and then the justification was assigned to a complexity bin. A total of 11 bins were constructed over the range of complexity (from 0 to 2200), each with a complexity interval of 200. We discarded all bins which had 0 non-isomorphic justifications of size 4-10. This left 8 bins partitioning a complexity range of 200-1800.

Figure 1 illustrates a key issue. The bulk of the justifications (esp. without the trivial), both with and without isomorphic reduction, are in the middle complexity range. However, the model is not sophisticated enough that small differences (e.g. below a difference of 400-600) are plausibly meaningful. It is unclear whether the noise from variance in participant abilities would wash out the noise from the complexity model.

[11] Note that, as a result, nearly 40% of all justifications have no representative in the pruned set (see Figure 3). Inspection revealed that most of these were trivial single axiom justifications (e.g. of the form $\{A \equiv B\} \models A \sqsubseteq B$ or $\{A \equiv (B \sqcap C)\} \models A \sqsubseteq B$, etc.

In other words, just from reflection on the model, justifications whose complexity difference is 400 or less do not seem reliably distinguishable by error rates. Furthermore, non-isomorphism does not eliminate all non-significant logical variance. Consider a chain of two atomic subsumptions vs. a chain of three. They have the same basic logical structure, but are not isomorphic. Thus, we cannot yet say whether this apparent concentration is meaningful.

Since we did not expect to be able to present more than 6 items and keep to our time limits, we chose to focus on a "easy/hard" divide of the lowest three non-empty bins (200-800) and the highest three non-empty bins (1200-1800). While this limits the claims we can make about model performance over the entire corpus, it, at least, strengthens negative results. If error rates overall do not distinguish the two poles (where we expect the largest effect) then either the model fails or error rates are not a reliable marker. Additionally, since if there is an effect, we expect it to be largest in this scenario thus making it easier to achieve adequate statistical power.

Each experiment involved a fixed set of test items, which were selected by randomly drawing items from preselected spread of bins, as described below. Please note that the selection procedure changed in the light of the pilot study, but only to make the selection more challenging for the model.[12]

The final stage of item construction was justification obfuscation. All non-logical terms were replaced with generated symbols. Thus, there was no possibility of using domain knowledge to understand these justifications. The names were all uniform, syntactically distinguishable (e.g. class names from property names) and quite short. The entailment was the same for all items, i.e. $\mathtt{C1} \sqsubseteq \mathtt{C2}$. It is possible that dealing with these purely symbolic justifications distorted participant response from response in the field, even beyond blocking domain knowledge. For example, they could be alienating and thus increase error rates or they could engage less error prone pattern recognition.

5 Results

The test items that were selected by the above sampling methodology are shown below. Every set of axioms is a justification for $\mathtt{C1} \sqsubseteq \mathtt{C2}$. There was no overlap in participants across the studies. For the main study, none of the authors were involved in facilitating the study, though Bail and Horridge participated in recruitment.

5.1 Pilot study

Participants: Seven members of a Computer Science (CS) Academic or Research Staff, or PhD Program, with over 2 years of experience with ontologies and justifications.
Materials and procedures: The study was performed using an in-house web based survey tool, which tracks times between all clicks on the page and thus records the time to make each decision.

The participants were given a series of test items consisting of 3 practice items, followed by 1 common easy item (**E1** of complexity 300) and four additional items,

[12] The selections are available from `http://owl.cs.manchester.ac.uk/research/publications/supporting-material/iswc2011-cog-comp`

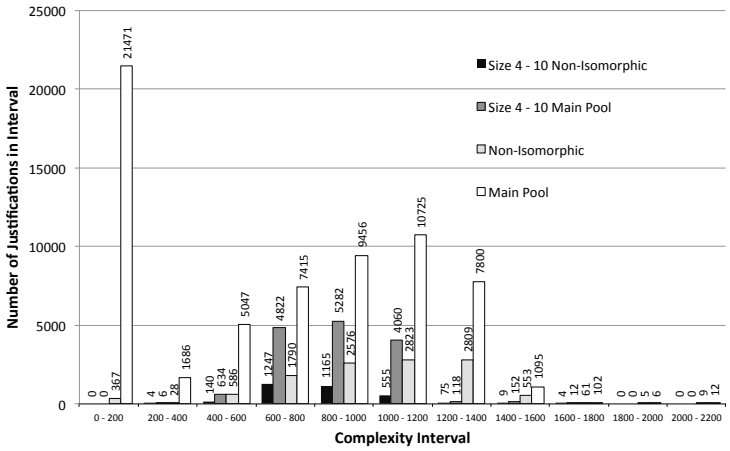

Fig. 1. Justification Corpus Complexity Distribution

2 ranked easy (**E2** and **E3** of complexities 544 and 690, resp.) and 2 ranked hard (**H1** and **H2** of complexities 1220 and 1406), which were randomly (but distinctly) ordered for each participant. The easy items were drawn from bins 200-800, and the hard items from bins 1200-1800. The expected time to complete the study was a maximum of 30 minutes, including the orientation, practice items, and brief demographic questionnaire (taken after all items were completed).

Results: Errors and times are given in Table 2. Since all of the items were in fact justifications, participant responses were recoded to success or failure as follows: Success = ("Yes" | "Yes, but not sure") and Failure = ("Not sure" | "No, Not sure" | "No"). Error proportions were analysed using Cochran's Q Test, which takes into consideration the pairing of successes and failures for a given participant. Times were analysed using two tailed paired sample t-tests.

Table 2. Pilot Study Failures and Response Times

Item	Failures	Mean Time (ms)	Time StdDev. (ms)
E1	0	65,839	39,370
E2	1	120,926	65,950
E3	2	142,126	61,771
H1	6	204,257	54,796
H2	6	102,774	88,728

An initial Cochran Q Test across all items revealed a strong significant difference in error proportions between the items [$Q(4) = 16.00, p = 0.003$]. Further analysis using

Cochran's Q Test on pairs of items revealed strong statistically significant differences in error proportion between: **E1/H1** [$Q(1) = 6.00$, $p = 0.014$], **E1/H2** [$Q(1) = 6.00$, $p = 0.014$] **E2/H2** [$Q(1) = 5.00$, $p = 0.025$] and **E3/H2** [$Q(1) = 5.00$, $p = 0.025$]. The differences in the remaining pairs, while not exhibiting differences above $p = 0.05$, were quite close to significance, i.e. **E2/H1** [$Q(1) = 3.57$, $p = 0.059$] and **E3/H1** [$Q(1) = 5.00$, $p = 0.10$]. In summary, these error rate results were encouraging.

An analysis of times using paired sample t-tests revealed that time spent understanding a particular item is not a good predictor of complexity. While there were significant differences in the times for **E1/H1** [$p = 0.00016$], **E2/H1** [$p = 0.025$], and **E3/H1** [$p = 0.023$], there were no significant differences in the times for **E1/H2** [$p = 0.15$], **E2/H2** [$p = 0.34$] and **E3/H2** [$p = 0.11$]. This result was anticipated, as in the exploratory study people gave up very quickly for justifications that they felt they could not understand.

5.2 Experiment 1

Participants: 14 volunteers from a CS MSc class on OWL ontology modelling, who were given chocolate for their participation.[13] Each participant had minimal exposure to OWL (or logic) before the class, but had, in the course of the prior 5 weeks, constructed or manipulated several ontologies, and received an overview of the basics of OWL 2, reasoning, etc. They did not receive any specific training on justifications.

Materials and procedures: The study was performed according to the protocol used in the pilot study. A new set of items were used. Since the mean time taken by pilot study participants to complete the survey was 13.65 minutes, with a standard deviation of 4.87 minutes, an additional hard justification was added to the test items. Furthermore, all of the items with easy justifications ranked easy were drawn from the highest easy complexity bin (bin 600-800). In the pilot study, we observed that the lower ranking easy items were found to be quite easy and, by inspection of their bins, we found that it was quite likely to draw similar justifications. The third bin (600-800) is much larger and logically diverse, thus is more challenging for the model.

The series consisted of 3 practice items followed by 6 additional items, 3 easy items(**EM1**, **EM2** and **EM3** of complexities: 654, 703, and 675), and 3 hard items (**HM1**, **HM2** and **HM3** of complexities: 1380, 1395, and 1406). The items were randomly ordered for each participant. Again, the expectation of the time to complete the study was a maximum of 30 minutes, including orientation, practice items and brief demographic questionnaire.

Results Errors and times are presented in Table 3. The coding to error is the same as in the pilot. An analysis with Cochran's Q Test across all items reveals a significant difference in error proportion [$Q(5) = 15.095, p = 0.0045$].

A pairwise analysis between easy and hard items reveals that there are significant and, highly significant, differences in errors between **EM1/HM1** [$Q(1) = 4.50$, $p = 0.034$], **EM1/HM2** [$Q(1) = 7.00$, $p = 0.008$], **EM2/HM1** [$Q(1) = 4.50$, $p = 0.034$], **EM2/HM2** [$Q(1) = 5.44$, $p = 0.02$], and **EM3/HM2** [$Q(1) = 5.44$, $p = 0.02$].

[13] It was made clear to the students that their (non)participation did not affect their grade and no person with grading authority was involved in the recruitment or facilitation of the experiment.

Table 3. Experiment 1 Failures and Response Times

Item	Failures	Mean Time (ms)	Time StdDev. (ms)
EM1	6	103,454	68,247
EM2	6	162,928	87,696
EM3	10	133,665	77,652
HM1	12	246,835	220,921
HM2	13	100,357	46,897
HM3	6	157,208	61,437

However, there were no significant differences between **EM1/HM3** [$Q(1) = 0.00$, $p = 1.00$], **EM2/HM3** [$Q(1) = 0.00$, $p = 1.00$], **EM3/HM3** [$Q(1) = 2.00$, $p = 0.16$] and **EM3/HM1** [$Q(1) = 0.67$, $p = 0.41$].

With regards to the nonsignificant differences between certain easy and hard items, there are two items which stand out: An easy item **EM3** and a hard item **HM3**, which are shown as the last pair of justifications in Figure 2.

In line with the results from the pilot study, an analysis of times using a paired samples t-test revealed significant differences between some easy and hard items, with those easy times being significantly less than the hard times **EM1/HM1** [$p = 0.023$], **EM2/HM2** [$p = 0.016$] and **EM3/HM1** [$p = 0.025$]. However, for other pairs of easy and hard items, times were not significantly different: EM1/HM1 [$p = 0.43$], **EM2/HM1** [$p = 0.11$] and **EM3/HM2** [$p = 0.10$]. Again, time is not a reliable predictor of model complexity.

Anomalies in Experiment 1: Two items (**EM3** and **HM3**) did not exhibit their predicted error rate relations. For item **EM3**, we conjectured that a certain pattern of superfluous axiom parts in the item (not recognisable by the model) made it harder than the model predicted. That is, that the *model* was wrong.

For item **HM3** we conjectured that the model correctly identifies this item as hard,[14] but that the MSc students answered "Yes" because of misleading pattern of axioms at the start and end of item **HM3**. The high "success" rate was due to an error in reasoning, that is, a *failure* in understanding.

In order to determine whether our conjectures were possible and reasonable, we conducted a followupup study with the goal of observing the conjectured behaviours in situ. Note that this study does *not* explain what happened in Experiment 1.

5.3 Experiment 2

Participants: Two CS Research Associates and one CS PhD student, none of whom had taken part in the pilot study. All participants were very experienced with OWL.

Materials and procedures: Items and protocol were exactly the same as Experiment 1, with the addition of the think aloud protocol [12]. Furthermore, the screen, participant vocalisation, and eye tracking were recorded.

[14] It had been observed to stymie experienced modellers in the field. Furthermore, it involves deriving a synonym for $\top$, which was not a move this cohort had experience with.

Results: With regard to **EM3**, think aloud revealed that all participants were distracted by the superfluous axiom parts in item **EM3**. Figure 3 shows an eye tracker heat map for the most extreme case of distraction in item **EM3**. As can be seen, hot spots lie over the superfluous parts of axioms. Think aloud revealed that all participants initially tried to see how the $\exists \mathsf{prop1}.\mathsf{C6}$ conjunct in the third axiom contributed to the entailment and struggled when they realised that this was not the case.

EM1

$\mathsf{C1} \sqsubseteq \exists\ \mathsf{prop1}.\mathsf{C3}$
$\mathsf{prop1} \sqsubseteq \mathsf{prop2}$
$\mathsf{prop2} \sqsubseteq \mathsf{prop3}$
$\mathsf{C3} \sqsubseteq \mathsf{C4}$
$\mathsf{C4} \sqsubseteq \mathsf{C5}$
$\mathsf{C5} \sqsubseteq \mathsf{C6}$
$\mathsf{C6} \sqsubseteq \mathsf{C7}$
$\mathsf{C7} \sqsubseteq \mathsf{C8}$
$\mathsf{C2} \equiv \exists\ \mathsf{prop3}.\mathsf{C8}$

HM1

$\mathsf{C1} \equiv \exists\ \mathsf{prop1}.\mathsf{C3}$
$\mathsf{prop1} \equiv \mathsf{prop2}^{-}$
$\mathsf{prop2} \sqsubseteq \mathsf{prop3}$
$\mathsf{prop3} \equiv \mathsf{prop4}^{-}$
$\mathsf{C3} \equiv (\exists\ \mathsf{prop5}.\mathsf{C4}) \sqcap (\exists\ \mathsf{prop2}.\mathsf{C1}) \sqcap (\forall \mathsf{prop5}.\mathsf{C4}) \sqcap (\forall\ \mathsf{prop2}.\mathsf{C1})$
$\mathsf{prop6} \equiv \mathsf{prop5}^{-}$
$\exists\ \mathsf{prop6}.\top \sqsubseteq \mathsf{C5}$
$\mathsf{C6} \sqsubseteq \mathsf{C7}$
$\mathsf{C6} \equiv (\exists\ \mathsf{prop5}.\mathsf{C5}) \sqcap (\forall \mathsf{prop5}.\mathsf{C5})$
$\mathsf{C2} \equiv \exists\ \mathsf{prop4}.\mathsf{C7}$

EM2

$\mathsf{C1} \equiv \mathsf{C3} \sqcap (\exists\ \mathsf{prop1}.\mathsf{C4}) \sqcap (\exists\ \mathsf{prop2}.\mathsf{C5})$
$\mathsf{C1} \sqsubseteq \mathsf{C6}$
$\mathsf{C6} \sqsubseteq \mathsf{C7}$
$\mathsf{C7} \sqsubseteq \mathsf{C8}$
$\mathsf{C8} \equiv \mathsf{C9} \sqcap (\exists\ \mathsf{prop1}.\mathsf{C10})$
$\mathsf{C2} \equiv \mathsf{C9} \sqcap (\exists\ \mathsf{prop1}.\mathsf{C4}) \sqcap (\exists\ \mathsf{prop2}.\mathsf{C5})$

HM2

$\mathsf{C3} \equiv (\exists\ \mathsf{prop1}.\mathsf{C5}) \sqcup (\forall\ \mathsf{prop1}.\mathsf{C5})$
$\mathsf{C3} \sqsubseteq \mathsf{C4}$
$\exists\ \mathsf{prop1}.\top \sqsubseteq \mathsf{C4}$
$\mathsf{C4} \sqsubseteq \mathsf{C2}$

EM3

$\mathsf{C1} \sqsubseteq \mathsf{C3}$
$\mathsf{C3} \sqsubseteq \mathsf{C4}$
$\mathsf{C4} \equiv \mathsf{C5} \sqcap (\exists\ \mathsf{prop1}.\mathsf{C6})$
$\mathsf{C5} \equiv \mathsf{C7} \sqcap (\exists\ \mathsf{prop2}.\mathsf{C8})$
$\mathsf{C1} \sqsubseteq \exists\ \mathsf{prop1}.\mathsf{C9}$
$\mathsf{C9} \sqsubseteq \mathsf{C10}$
$\mathsf{C2} \equiv \mathsf{C7} \sqcap (\exists\ \mathsf{prop1}.\mathsf{C10})$

HM3

$\mathsf{C1} \sqsubseteq \forall\ \mathsf{prop1}.\mathsf{C3}$
$\mathsf{C6} \equiv \forall\ \mathsf{prop2}.\mathsf{C7}$
$\mathsf{C6} \sqsubseteq \mathsf{C8}$
$\mathsf{C8} \sqsubseteq \mathsf{C4}$
$\mathsf{C4} \sqsubseteq \exists\ \mathsf{prop1}.\mathsf{C5}$
$\exists\ \mathsf{prop2}.\top \sqsubseteq \mathsf{C4}$
$\mathsf{C2} \equiv (\exists\ \mathsf{prop1}.\mathsf{C3}) \sqcup (\forall\ \mathsf{prop3}.\mathsf{C9})$

Fig. 2. Justifications Used in Experiment 1. All justifications explain the entailment $\mathsf{C1} \sqsubseteq \mathsf{C2}$

In the case of **HM3**, think aloud revealed that none of the participants understood how the entailment followed from the set of axioms. However, two of them responded correctly and stated that the entailment did hold. As conjectured, the patterns formed

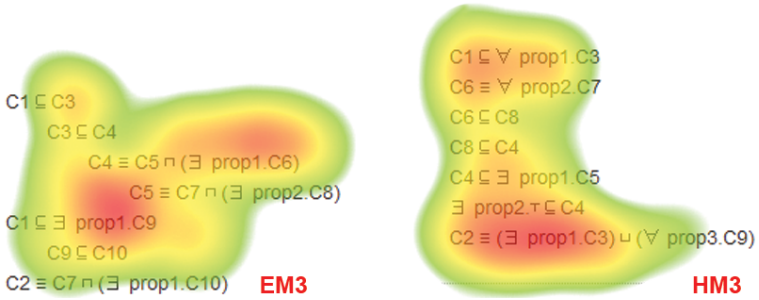

Fig. 3. Eye Tracker Heat Maps for **EM3** & **HM3**

by the start and end axioms in the item set seemed to mislead them. In particular, when disregarding quantifiers, the start axiom $\mathsf{C1} \sqsubseteq \forall \mathsf{prop1.C3}$ and the end axiom $\mathsf{C2} \sqsubseteq \exists \mathsf{prop1.C3} \sqcup \ldots$ look very similar. One participant spotted this similarity and claimed that the entailment held as a result. Hot spots occur over the final axiom and the first axiom in the eye tracker heat map (Figure 3), with relatively little activity in the axioms in the middle of the justification.

6 Dealing with Justification Superfluity

Perhaps the biggest issue with the current model is that it does not deal at all with superfluity in axioms in justifications. That is, it does not penalise a justification for having axioms that contain, potentially distracting, superfluous parts—parts that do not matter as far as the entailment is concerned. Unfortunately, without a deeper investigation, it is unclear how to rectify this in the model. Although it is possible to identify the superfluous parts of axioms using laconic and precise justifications [4], throwing a naive superfluity component into the model would quite easily destroy it. This is because there can be justifications with plenty of superfluous parts that are trivial to understand. For example consider $\mathcal{J} = \{A \sqsubseteq B \sqcap C\} \models A \sqsubseteq B$, where C is along and complex class expression, and yet there can be justifications with seemingly little superfluity (as in the case of **EM3**) which causes complete distraction when trying to understand an entailment. Ultimately, what seems to be important is the location and shape of superfluity, but deciding upon what "shapes" of superfluity count as non-trivial needs to be investigated as part of future work.

One important point to consider, is that it might be possible to deal with the problems associated with superfluity by presentation techniques alone. It should be clear that the model does not pay any attention to how justifications are presented. For example, it is obvious that the ordering (and possibly the indentation) of axioms is important. It can make a big difference to the readability of justifications and how easy or difficult they are to understand, yet the model does not take into consideration how axioms will be ordered when a justification is presented to users. In the case of superfluity, it is conceivable that *strikeout* could be used to cross out the superfluous parts of axioms and

this would dispel any problems associated with distracting superfluity. Figure 4 shows the helpful effect of strikeout on **EM3**. As can be seen, it immediately indicates that the problematic conjunct, $\exists$ prop1.C6, in the third axiom should be ignored. Some small scale experiments, carried out as part of future work, could confirm this.

Fig. 4. EM3 with and without strikeout

7 Discussion and Future Work

In this paper we presented a methodology for validating the predicted complexity of justifications. The main advantages of the experimental protocol used in the methodology is that minimal study facilitator intervention is required. This means that, over time, it should be possible to collect rich and varied data fairly cheaply and from geographically distributed participants. In addition to this, given a justification corpus and population of interest, the main experiment is easily repeatable with minimal resources and setup. Care must be taken in interpreting results and, in particular, the protocol is weak on "too hard" justifications as it cannot distinguish a model mislabeling from people failing for the wrong reason.

The cognitive complexity model that was presented in this paper fared reasonably well. In most cases, there was a significant difference in error proportion between model ranked easy and hard justifications. In the cases where error proportions revealed no difference better than chance, further small scale follow-up studies in the form of a more expensive talk-aloud study was used to gain an insight into the problems. These inspections highlighted an area for model improvement, namely in the area of superfluity. It is unclear how to rectify this in the model, as there could be justifications with superfluous parts that are trivial to understand, but the location and shape of superfluity seem an important factor.

It should be noted that the goal of the experiments was to use error proportion to determine whether two justifications come from different populations—one from the set of easy justifications and one from the set of hard justifications. This is rather different than being able to say, with some level of statistical confidence, that the model generalises to the whole population of easy or hard justifications. For the former the statistical toolbox that is used is workable with very small sample sizes. Ultimately

the sample size depends on the variance of the sample, but sample sizes of less than 10 can work, where sample size is the number of outcomes (successes or failures) per justification. For the latter, sample sizes must be much larger. For example, by rule of thumb, around 400 justifications would be needed from the hard category to be able say with 95% confidence that all of hard justifications are actually hard justifications. While being able to generalise to the whole population would be the best outcome, the fact that participants would have to answer 400 items means that this is not achievable, and so the focus is on using error proportion to determine the actually hardness of a justification.

The refinement and validation of our model is an ongoing task and will require considerably more experimental cycles. We plan to conduct a series of experiments with different cohorts as well as with an expanded corpus. We also plan to continue the analysis of our corpus with an eye to performing experiments to validate the model over the whole (for some given population).

References

1. A. Borgida, D. Calvanese, and M. Rodriguez-Muro. Explanation in the DL-lite family of description logics. In *OTM-08*, 2008.
2. M. Davis. Obvious logical inferences. In *IJCAI-81*, 1981.
3. M. Horridge and B. Parsia. From justifications towards proofs for ontology engineering. In *KR-2010*, 2010.
4. M. Horridge, B. Parsia, and U. Sattler. Laconic and Precise justifications in OWL. In *ISWC 2008*, 2008.
5. M. Horridge, B. Parsia, and U. Sattler. Lemmas for justifications in OWL. In *DL 2009*, 2009.
6. M. Horridge, B. Parsia, and U. Sattler. Justification oriented proofs in OWL. In *ISWC 2010*, 2010.
7. P. N. Johnson-Laird and R. M. J. Byrne. *Deduction*. Psychology Press, 1991.
8. A. Kalyanpur, B. Parsia, E. Sirin, and B. Grau. Repairing unsatisfiable concepts in OWL ontologies. In *ESWC 06*, 2006.
9. A. Kalyanpur, B. Parsia, E. Sirin, and J. Hendler. Debugging unsatisfiable classes in OWL ontologies. *Journal of Web Semantics*, 3(4), 2005.
10. Y. Kazakov. $\mathcal{RIQ}$ and $\mathcal{SROIQ}$ are harder than $\mathcal{SHOIQ}$. In *KR 2008*. AAAI Press, 2008.
11. S. C. J. Lam. *Methods for Resolving Inconsistencies In Ontologies*. PhD thesis, Department of Computer Science, Aberdeen, 2007.
12. C. H. Lewis. Using the thinking-aloud method in cognitive interface design. Research report RC-9265, IBM, 1982.
13. S. Newstead, P. Brandon, S. Handley, I. Dennis, and J. S. B. Evans. Predicting the difficulty of complex logical reasoning problems. *Psychology Press*, 12, 2006.
14. L. M. Parsons and D. Osherson. New evidence for distinct right and left brain systems for deductive versus probabilistic reasoning. *Cerebral Cortex*, 11(10):954–965, 2001.
15. L. J. Rips. *The Psychology of Proof*. MIT Press, Cambridge, MA, 1994.
16. C. Roussey, O. Corcho, and L. Vilches-Blázquez. A catalogue of OWL ontology antipatterns. In *Proc. of K-CAP-09*, pages 205–206, 2009.
17. G. Strube. The role of cognitive science in knowledge engineering. In *Contemporary Knowledge Engineering and Cognition*, 1992.