Testing the Adequacy of Automated Explanations of $\mathcal{EL}$ Subsumptions

Marvin Schiller[1], Florian Schiller[2], and Birte Glimm[1]

[1]University of Ulm, Germany
{marvin.schiller, birte.glimm}@uni-ulm.de
[2]Justus-Liebig-Universität Gießen, Germany
florian.schiller@psychol.uni-giessen.de

Abstract. Ontology verbalization techniques have been introduced to automatically translate description logic (DL) axioms and derivations to natural-language texts. This way, non-expert users can be offered explanations for subsumptions derived by systems using ontologies for knowledge-representation. We address the question of the readability and understandability of explanations generated from longer chains of inference steps, as they occur with non-trivial ontologies. An experimental design is presented to assess readers' understanding, the readability and the quality of the generated texts. The experiment tests verbalizations of derivations of different lengths, and assesses the effect of a strategy proposed to shorten explanations while retaining an adequate level of informativeness.

1 Introduction

Ontologies serve to organize concepts, terminology and relationships in a domain of interest, such as biology or medicine. Furthermore, logical consequences of this knowledge can be derived using automated ontology reasoners. To make this knowledge available to users that are not familiar with the employed formalisms (for instance, ontology languages such as OWL), verbalization techniques have been developed to automatically translate axioms and derivations to natural-language statements and explanations.

Example 1. As the running example throughout this paper, consider the subsumption

$$\mathit{EsophagealPathology} \sqsubseteq \mathit{DigestiveSystemPathology}$$

that can be derived from the following axioms:

$$\begin{aligned}
\mathit{EsophagealPathology} &\equiv (\mathit{PathologicalCondition} \sqcap \\
&\quad \exists \mathit{LocativeAttribute}.\mathit{Esophagus}) \\
\mathit{Esophagus} &\sqsubseteq \mathit{GastrointestinalTractBodyPart} \\
\mathit{DigestiveSystemPathology} &\equiv (\mathit{PathologicalCondition} \sqcap \\
&\quad \exists \mathit{LocativeAttribute}.\mathit{GastrointestinalTractBodyPart})
\end{aligned}$$

The verbalization approach presented in this paper constructs a step-wise argument for this derivation in natural language, in this case:

An esophageal pathology is defined as a pathological condition that is located in the esophagus. The esophagus is a part of the gastrointestinal tract, thus an esophageal pathology is located in a part of the gastrointestinal tract. Furthermore, since an esophageal pathology is a pathological condition, an esophageal pathology is a pathological condition that is located in a part of the gastrointestinal tract. A digestive system pathology is defined as a pathological condition that is located in a part of the gastrointestinal tract. Thus, an esophageal pathology is a digestive system pathology.

Such an explanation combines the information from the relevant axioms in a step-wise fashion and does not require readers to be familiar with the syntax of ontology languages or description logics. However, depending on the granularity at which knowledge is modeled in an ontology, the possible derivations, and consequently the generated explanations for these derivations, can grow very long. The state-of-the-art in ontology verbalization has so far concentrated on short inference problems (one or two inference steps), but not addressed the problems that arise when verbalizing more complex derivations (more than two inference steps). Ontologies do not need to be very expressive for long and sufficiently complex derivations to occur. All the derivations considered in the following remain within the OWL2 EL profile.

This paper is organized as follows. Building on a short introduction to the considered language fragment in Section 2, we present related work on verbalization in Section 3. In Section 4 we present our own approach, which extends previous approaches by focusing on the aspect of conciseness of the generated explanations. Section 5 presents an experiment that compares explanations generated by our setup in two different variants, leading to the conclusions in Section 6.

2 Preliminaries

This work remains within the DL fragment $\mathcal{EL}$ with some extensions that are common features of the OWL2 EL profile (which is based on the DL $\mathcal{EL}^{++}$). As usual, class names are denoted with capital letters $A, B, C, ...$, role names with small letters $r, s, ...$, individuals with small letters $a, b...$ and the universal concept with $\top$. Complex class expressions are formed by using conjunction ($C_1 \sqcap C_2$) and existential restriction ($\exists r.C$). Axioms that specify the subclass relationship between two class expressions C_1 and C_2, also known as subsumptions, are denoted as $C_1 \sqsubseteq C_2$. Besides these pure $\mathcal{EL}$ constructors, we consider further constructors that are common in the OWL2 EL profile. This includes the unsatisfiable concept $\bot$, nominals $\{a\}$ which are concepts consisting of a single individual a, domain axioms $dom(r, C)$ that are a shorthand for $\exists r.\top \sqsubseteq C$, equivalences between concepts (mutual subclass relationship), denoted as $C_1 \equiv C_2$, disjointness axioms specifying that two class expressions C_1 and C_2 are disjoint, as $disj(C_1, C_2)$, and role inclusion axioms $r \sqsubseteq s$. We further include role inclusions that use role composition $r_1 \circ ... \circ r_k \sqsubseteq s$ (called property chains in OWL2 EL).

A more comprehensive introduction to $\mathcal{EL}^{++}$ is provided by Baader et al. [2]. Despite its limited expressiveness, a number of practically relevant ontologies in numerous application domains fall into this language fragment. This includes, for instance, large biomedical ontologies such as SNOMED CT[1], the Gene Ontology (GO),[2] and large portions of the NCI Thesaurus[3] and the Galen ontology.[4]

3 Related Work

A correspondence between formal expressions in ontologies and natural-language has been proposed in the form of controlled languages (cf. [11]), for instance OWL Simplified English (OSE, [16]), Attempto Controlled English (ACE, [8]), Sydney OWL Syntax (SOS, [4]), CLOnE [6] and Rabbit [5]. Controlled languages define a (usually very restricted) subset of natural language that unambiguously corresponds to DL constructors and expressions. For example, the subsumption $C1 \sqsubseteq C2$ is represented in OSE as "A [C1] is a [C2]", where [C1] and [C2] are text strings to represent the concept descriptions for $C1$ and $C2$, respectively. For example, "A city is a place". Whereas controlled languages remain closely-tied to the corresponding formalism, some approaches have focused on text quality and support for different (natural) languages. This includes a tool developed by the SWAT project [18], the OntoVerbal verbalizer [12] and NaturalOWL [1]. Whereas these approaches verbalize axioms in a knowledge base, some approaches have considered explanations generated from derivations. These include the Classic system [13], the "tracing" facility of the ELK reasoner [9], and the approaches of Borgida et al. [3] and Nguyen et al. [14]. Whereas the explanation facilities of ELK and Classic do not use natural language, Borgida et al. use text patterns for inference rules, but retain formula language for axioms. The approach of Nguyen et al. [14] is similar to ours, since it employs rule-based proofs and patterns to produce natural-language explanations, such as, for example:

(a) Every A is a B.
(b) Every B is a C.
→ (c) Every A is a C.

The generated explanation combines such patterns using the text pattern "Statement (c) is implied because (a) ... and (b) ...". Thus, the structure remains quite close to how proofs are presented, but the formulae are replaced by more commonly-understandable text patterns. To test the understandability of these patterns, an experiment was conducted where the acceptance of these patterns was tested. Different rules and corresponding text patterns were found to vary greatly in whether they were accepted as correct by experiment participants (cf. [15]). The understandability of the individual verbalized inference rules was used to predict the understandability of verbalized two-rule inference problems, and was indeed found to be correlated with the empirically measured understandability [15].

[1] `http://www.snomed.org/snomed-ct`
[2] `http://www.geneontology.org`
[3] `https://ncit.nci.nih.gov/ncitbrowser/`
[4] `http://www.opengalen.org/`

4 Generating Verbalized Explanations for Derivations

The generation of explanations is based on two main components, a consequence-based reasoning system for generating step-wise derivations and a natural-language generation component to transform these formal derivations to text. In the following, these two components are introduced briefly. Based on this, we address the problem of the inconciseness of some of the generated explanations by introducing techniques and heuristics that shorten the explanations. The presented approach has been implemented as a prototype system and is available as a plugin[5] for the ontology editor Protégé.[6]

4.1 Reasoning

Derivations are constructed using a rule-based inference system with a custom set of inference rules. Using a custom system allows us to include inference rules that are logically redundant, but which help to obtain shorter derivations, and thus shorter explanations. The current implementation includes and modifies rules proposed by Nguyen et al. [14] and the rules employed in the ELK system [10] and incorporates a few additional inference rules. Fig. 1 shows some of the inference rules relevant for the remainder of this paper (together with verbalization patterns, as discussed further below). Note that the introduced modifications and additions do not impact the formal properties of the original rule systems, instead they introduce shortcuts (e.g. $\boldsymbol{R}_{\sqsubseteq\equiv}$) and n-ary versions of originally binary rules (e.g. $\boldsymbol{R}^{+}_{\sqcap}/\boldsymbol{R}_{\mathbf{5}}$). The full ruleset for the DL fragment considered in this paper is shown in [17, Appendix A].

Since the current implementation of the consequence-based reasoning procedure is not as performant on large ontologies as well-optimized tableau-based reasoners, proof search is not performed on an entire ontology. Rather, in a preprocessing step, a justification [7] (a minimal set of axioms required to prove a derived axiom) is obtained using an off-the-shelf tableau-based reasoner (such as FaCT++,[7] HermiT,[8] etc.). Then proof search is performed only on the set of relevant axioms. Such pre-processing is also used by related work [14]. In the example from the introduction, the following proof tree is obtained (with abbreviations EP: EsophagealPathology, DSP: DigestiveSystemPathology, GTP: GastrointestinalTractBodyPart, PC: PathologicalCondition, loc: locativeAttribute):

$$\boldsymbol{R}_{\sqsubseteq\equiv}\,\dfrac{\boldsymbol{R}^{+}_{\sqcap}/\boldsymbol{R}_{\mathbf{5}}\,\dfrac{\boldsymbol{R}_{\equiv-}\,\dfrac{\mathrm{EP}\equiv\mathrm{PC}\sqcap\exists\mathrm{loc.E}}{\mathrm{EP}\sqsubseteq\mathrm{PC}} \qquad \boldsymbol{R}_{\exists}/\boldsymbol{R}_{\mathbf{15}}\,\dfrac{\boldsymbol{R}_{\equiv-}\,\dfrac{\mathrm{EP}\equiv\mathrm{PC}\sqcap\exists\mathrm{loc.E}}{\mathrm{EP}\sqsubseteq\exists\mathrm{loc.E}} \qquad \mathrm{E}\sqsubseteq\mathrm{GTP}}{\mathrm{EP}\sqsubseteq\exists\mathrm{loc.GTP}}}{\mathrm{EP}\sqsubseteq\mathrm{PC}\sqcap\exists\mathrm{loc.GTP}} \qquad \mathrm{DSP}\equiv\mathrm{PC}\sqcap\exists\mathrm{loc.GTP}}{\mathrm{EP}\sqsubseteq\mathrm{DSP}}$$

---

[5] `https://verbalizer.github.io/`
[6] `http://protege.stanford.edu/`
[7] `http://owl.man.ac.uk/factplusplus/`
[8] `http://www.hermit-reasoner.com/`

$$\boldsymbol{R_1}\ \frac{C_1 \equiv C_2}{C_1 \sqsubseteq C_2}$$

—: `According to its definition,` $\mathrm{v}(C_1 \sqsubseteq C_2)$.

$$\boldsymbol{R_{\sqcap}^{-}}/\boldsymbol{R_2}\ \frac{C_1 \sqsubseteq C_2 \sqcap \ldots \sqcap C_n}{C_1 \sqsubseteq C_i}\ 2 \leq i \leq n$$

—: `Hence,` $\mathrm{v}(C_1 \sqsubseteq C_i)$.

$$\boldsymbol{R_{\sqcap}^{+}}/\boldsymbol{R_5}\ \frac{(1)\ C_1 \sqsubseteq C_2 \quad \ldots \quad (n)\ C_1 \sqsubseteq C_{n+1}}{C_1 \sqsubseteq C_2 \sqcap \ldots \sqcap C_{n+1}}$$

(1) ... (n): `Since` $\mathrm{v}(C_1 \sqsubseteq C_2)$ `and` ... `and` $\mathrm{v}(C_1 \sqsubseteq C_{n+1})$, $\mathrm{v}(C_1 \sqsubseteq C_2 \sqcap \ldots \sqcap C_{n+1})$.
(i), (j),...: `Furthermore, since` $\mathrm{v}(C_1 \sqsubseteq C_i)$ `and` $\mathrm{v}(C_1 \sqsubseteq C_j)$ `and` ...`,` $\mathrm{v}(C_1 \sqsubseteq C_2 \sqcap \ldots \sqcap C_{n+1})$.
— : `Therefore` $\mathrm{v}(C_1 \sqsubseteq C_2 \sqcap \ldots \sqcap C_{n+1})$.

$$\boldsymbol{R_{\sqsubseteq}}/\boldsymbol{R_{12}}\ \frac{(1)\ C_1 \sqsubseteq C_2 \qquad (2)\ C_2 \sqsubseteq C_3}{C_1 \sqsubseteq C_3}$$

(1)&(2): `Since` $\mathrm{v}(C_1 \sqsubseteq C_2)$ `and` $\mathrm{v}(C_2 \sqsubseteq C_3)$ `it follows that` $\mathrm{v}(C_1 \sqsubseteq C_3)$.
(1): $\mathrm{v}(C_1 \sqsubseteq C_2)$`, therefore being` $\mathrm{v}(C_3)$.
(2): `Given that` $\mathrm{v}(C_2 \sqsubseteq C_3)$, $\mathrm{v}(C_1 \sqsubseteq C_3)$.
— : `Thus, we have established that` $\mathrm{v}(C_1 \sqsubseteq C_3)$.

$$\boldsymbol{R_{\exists}}/\boldsymbol{R_{15}}\ \frac{(1)\ C_1 \sqsubseteq \exists r.C_2 \qquad (2)\ C_2 \sqsubseteq C_3}{C_1 \sqsubseteq \exists r.C_3}$$

(1)&(2): $\mathrm{v}(C_1 \sqsubseteq \exists r.C_2)$ `which [is]` $\mathrm{v}(C_3)$. `Therefore,` $\mathrm{v}(C_1 \sqsubseteq \exists r.C_3)$.
(1): $\mathrm{v}(C_1 \sqsubseteq \exists r.C_2)$`, thus` $\mathrm{v}(C_1 \sqsubseteq \exists r.C_3)$.
(2): $\mathrm{v}(C_2 \sqsubseteq C_3)$`, thus` $\mathrm{v}(C_1 \sqsubseteq \exists r.C_3)$.
— : `Therefore,` $\mathrm{v}(C_1 \sqsubseteq \exists r.C_3)$.

$$\boldsymbol{R_{\sqsubseteq\equiv}}\ \frac{(1)\ C_1 \sqsubseteq C_2 \quad (2)\ C_2 \equiv C_3}{C_1 \sqsubseteq C_3}$$

(1)&(2): $\mathrm{v}(C_2)$ `is defined as` $\mathrm{v}(C_3)$. `Thus,` $\mathrm{v}(C_1 \sqsubseteq C_3)$.
— : `Thus,` $\mathrm{v}(C_1 \sqsubseteq C_3)$ `according to the definition of` $\mathrm{v}(C_2)$.

$$\boldsymbol{R_{\equiv-}}\ \frac{C_1 \equiv C_2 \sqcap \ldots \sqcap C_n}{C_1 \sqsubseteq C_i}\ 2 \leq i \leq n$$

— : $\mathrm{v}(C_1)$ `is defined as` $\mathrm{v}(C_i \sqcap \ldots \sqcap C_n)$.

Fig. 1. Selected inference rules with verbalization patterns. Verbalization of formulas is denoted as v(...). Rules where premises are numbered have alternative different verbalization patterns (indicated by the corresponding numbers), which are selected according to which premises need to be mentioned at the position where the rule is applied within a proof. Numbered rule names refer to corresponding/similar inference rules in Nguyen et al. [14], $\boldsymbol{R_{\sqcap}^{-}}$, $\boldsymbol{R_{\sqcap}^{+}}$, $\boldsymbol{R_{\sqsubseteq}}$, and $\boldsymbol{R_{\exists}}$ refer to corresponding/similar rules in ELK.

4.2 Verbalization

First, the inference steps are ordered in a linear sequence for being output as text. For this, a post-order traversal of the proof tree (as seen from the root of the tree, which contains the conclusion of the derivation) is performed. For each inference rule it is specified in which order its children (i.e., premises) are being output, which corresponds to the order in which the rules are indicated in Fig. 1. Then, the text patterns in Fig. 1 are applied to transform the derivation into text. For each rule, the exact pattern that is applied depends on whether one or several of its premises have been presented immediately before in the generated text (for example, as a conclusion of a previous step), in which case they should not be repeated.

In the running example, the first step to be output is the application of $\boldsymbol{R}_{\equiv -}$(top left in the proof tree), with its template producing by default: "An esophageal pathology is defined as a pathological condition that is located in the esophagus". The second application of $\boldsymbol{R}_{\equiv -}$ (top center in the proof tree) produces no output, for being detected as identical to the previous output. The next rule application to be output is $\boldsymbol{R}_{\exists}/\boldsymbol{R}_{\mathbf{15}}$. Since the first premise is counted as being "covered" by the previous output, only the second premise (marked as (2)) is output, together with the conclusion, i.e. the second pattern is chosen: "The esophagus is a part of the gastrointestinal tract, thus an esophageal pathology is located in a part of the gastrointestinal tract." As can be seen in this example, some intermediate conclusions remain implicit (e.g. $\mathrm{EP} \sqsubseteq \mathrm{PC}$ and $\mathrm{EP} \sqsubseteq \exists \mathrm{loc.E}$ as conclusions of $\boldsymbol{R}_{\equiv -}$), a strategy we discuss below.

4.3 Techniques and Heuristics for Improving Text Quality

The aim of conciseness of the verbalized derivations is pursued at three levels. At the level of the generated proofs, inference rules are used that correspond to two or more applications of simpler rules and provide a kind of shortcut. In the running example, one application of $\boldsymbol{R}_{\sqsubseteq \equiv}$ (the last step) replaces the application of $\boldsymbol{R}_{\mathbf{1}}$ and $\boldsymbol{R}_{\sqsubseteq}/\boldsymbol{R}_{\mathbf{12}}$.

Secondly, some inference rules are considered to be trivial and their application is simply omitted altogether in the text output. Among the rules in Fig. 1, $\boldsymbol{R}^{-}_{\sqcap}/\boldsymbol{R}_{\mathbf{2}}$ represents such an inference rule. Furthermore, for some of the extra inference rules, the conclusion is not being output (again, for being considered obvious). As illustrated above, this is the case for $\boldsymbol{R}_{\equiv -}$.

Finally, at the level of individual statements, the text patterns are designed such that unnecessary repetitions are avoided. For example, the "middle term" in $\boldsymbol{R}_{\sqsubseteq}/\boldsymbol{R}_{\mathbf{12}}$ only needs to be mentioned once (in contrast to the pattern used by [14] and shown in Section 3). Furthermore, the verbalization mechanism uses annotations to replace class and role names with more readable names, where provided. This was used in the running example to supply the concept originally named *NAMEDGITractBodyPart* with a more readable label "part of the gastrointestinal tract".

When using more "complex" inference rules and hiding inference rule applications (referred to as "shortening" in the following), the question is whether the understandability of the resulting explanations is retained. This prompted an investigation presented in the following section.

5 Experiment

To test the understandability and text quality of generated explanations for derivations in ontologies, a questionnaire-based experiment was devised. This experiment allowed for a comparison between explanations in their unshortened form and their shortened form according to the presented heuristics. Since no such experiment has been conducted before, we explored with a small number of participants whether its design is suited as an instrument for assessing differences between shortened and unshortened explanations. We consider the results informative as a preparation for larger studies and also for the further development of the presented verbalization techniques.

Procedure Participants were randomly assigned to two groups. Eight explanations were shown to each participant. The first group received four explanations in their unshortened version and four in their shortened version. The second group received the corresponding shortened and unshortened alternatives of these explanations. The shortened version of an explanation uses logically redundant rules ($\boldsymbol{R}_{\sqsubseteq\equiv}$ and $\boldsymbol{R}_{\equiv-}$ in the running example), whereas the unshortened version uses only the most basic rules. In the unshortened case, verbalizations of $\boldsymbol{R}_{\sqcap}^{-}/\boldsymbol{R_2}$ are omitted. For comparison, Fig. 2 shows both versions for the running example used in the experiment.

As an objective test for participants' careful reading and understanding, participants were asked to indicate for each explanation whether it is logically correct. Two out of the eight explanations were manipulated to be erroneous by replacing one occurrence of a classname by a different one which was not part of the initial axioms. This manipulation was designed to ensure that participants read the text properly, but not to test their formal reasoning skills.[9] Note also that the two last sentences of Fig. 2 (b) are generated from one rule application of $\boldsymbol{R}_{\sqsubseteq\equiv}$. The employed pattern was different from the one presented in Fig. 1, in that it lacks the part "defined as" to make clear that it refers to an equivalence and not a subsumption. This deficiency was detected during the experiment and corrected in the verbalization system.

Understandability and readability of the explanations were assessed using a questionnaire to be answered on a 7-point scale. Participants were asked for ratings pertaining to key aspects of the text quality of the presented explanations; namely understanding/comprehension, conciseness and appreciation. The topic domains for which the explanations were generated were chosen to be relatively

[9]It is well known that untrained participants in reasoning experiments do not always apply classical logical reasoning. For example, consider the Wason selection task [19].

An esophageal pathology is defined as a pathological condition that is located in the esophagus.
Hence, an esophageal pathology is a pathological condition.
Additionally, an esophageal pathology is located in the esophagus.
The esophagus is a part of the gastrointestinal tract, thus an esophageal pathology is located in a part of the gastrointestinal tract.
Furthermore, since an esophageal pathology is a pathological condition, an esophageal pathology is a pathological condition that is located in a part of the gastrointestinal tract.
According to the definition of a digestive system pathology, a pathological condition that is located in a part of the gastrointestinal tract is a digestive system pathology.
Thus, we have established that an esophageal pathology is a digestive system pathology.

(a) Unshortened Explanation

An esophageal pathology is defined as a pathological condition that is located in the esophagus.
The esophagus is a part of the gastrointestinal tract, thus an esophageal pathology is located in a part of the gastrointestinal tract.
Furthermore, since an esophageal pathology is a pathological condition, an esophageal pathology is a pathological condition that is located in a part of the gastrointestinal tract.
A digestive system pathology is a pathological condition that is located in a part of the gastrointestinal tract. Thus, an esophageal pathology is a digestive system pathology.

(b) Shortened explanation

Fig. 2. Explanation for "An esophageal pathology is a digestive system pathology." employed in the experiment in unshortened and shortened form.

abstract and unfamiliar to most participants. Therefore, when judging readability one has to take into account that the domain itself may be challenging, an aspect for which questions were included under *appreciation*. The items are:

Understandability

- I can follow the reasoning steps presented in the explanation. (Question 1)

Conciseness

- I find that some steps in the explanation are so obvious that they should be skipped. (Question 2)
- The explanation conveys less information than I need to fully understand it. (Question 3)
- I find that the explanation should be made more concise. (Question 4)

Appreciation

- The text of the explanation is well-formed (according to writing conventions) (Question 5)
- The sentences are arranged such that they fit together well. (Question 6)
- I find the text easy to read. (Question 7)
- I am familiar with the technical terms in this text. (Question 8)
- The technical terms make it difficult for me to follow the text. (Question 9)
- The topic of the text makes it difficult for me to read the text. (Question 10)

At the beginning of the experiment, participants were informed that they will be asked to provide judgments for automatically generated explanations. An example was shown together with the correct answer and an explanation (reproduced in [17, Appendix B]). Then participants were asked for demographic data and prior experience with the following fields of science: computer science, artificial intelligence, mathematics/formal logic, philosophy, linguistics, physics, biology, medicine, chemistry. Participants received the eight explanations in random order, each with the same set of questions. The presentation of each explanation was preceded by a presentation of the axioms that were assumed to hold (in verbalized form) and the conclusion derived from them (also verbalized). A screenshot showing the running example together with the associated questionnaire is reproduced in [17, Appendix B]. After the experiment, participants were offered a free-text field for any comments on the explanations and were thanked for their participation. The questionnaire was administered using LimeSurvey.[10]

Participants Seven current and former members of Ulm University took part in the experiment. None of them was involved in the development of the presented verbalization techniques and the experiment. Participants included one female and six males and were aged between 20 and 34. Four indicated to be fluent in English, two indicated good English proficiency and one intermediate proficiency. All participants reported experience in computer science, four in Artificial Intelligence, and three in mathematics/formal logics.

Materials To generate a pool of verbalized derivations, the verbalization tool was run on the ontologies in the TONES repository.[11] To enable the retrieval of subcorpora of verbalizations according to a set of criteria (e.g. number of inference steps, employed rules), explanations and their properties were stored in a MySQL[12] database. The TONES ontology repository was chosen for including a good level of diversity regarding the included ontologies' domains and complexities. Since TONES includes ontologies of different levels of expressivity, only derivations that fell into the language fragment handled by the verbalization tool were generated. Still, a sufficiently large and diverse corpus of explanations in various domains (anatomy, chemistry, geology, physics, but also the "pizza" tutorial domain[13]) was obtained. Most importantly, this set represents "ordinary" ontologies that were not designed with verbalization in mind, but which are likely to be encountered by potential users of the verbalization tool. To restrict the scope of the experiment in a sensible way, the pool of considered explanations was further narrowed down according to the following criteria:

- The explanations are of medium length (3–5 inference steps when shortened).
- The verbalizations make use of additional rules or the skipping of rules in the presentation, as discussed in this paper, such that the effect of shortening can be studied.
- The axioms from which the verbalizations are generated are plausible according to common sense. Some of the included ontologies were found to contain unusual or erroneous formalizations of their domain. For example, some ontologies include axioms for a concept – which in the real world is known to have instances – such that the concept becomes unsatisfiable (probably

[10] https://www.limesurvey.org/de/
[11] https://zenodo.org/record/32717
[12] https://www.mysql.com/de/
[13] http://mowl-power.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf

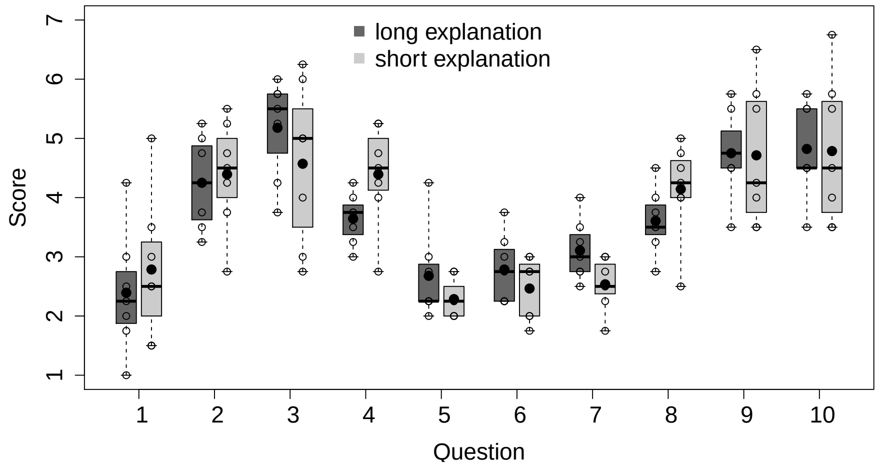

Fig. 3. Scores on the ten questions averaged across the four shortened and unshortened explanations on a 7-point scale from 1: *strongly agree* to 7: *strongly disagree*. Unfilled circles represent the average scores of a single participant. Filled black circles represent the average across all participants, the black horizontal bar represents the median. Whiskers indicate the the entire range of the individual means. The boundaries of boxes indicate the upper and lower quartile.

unintendedly) and therefore the concept can be shown to be a subconcept of any concept. Some ontologies also contain equivalences with apparently unintended consequences.

- Verbalizations that rely on equivalences between several concepts with the same classname in different ontologies (with different URLs) were excluded. Such equivalences result in statements such as "since a person is a person...".
- Explanations with excessively long concept names were also excluded. However, since long concept names are quite common in the investigated ontologies, compound concept names of up to three words were accepted.

In order to provide a realistic selection of explanations for the experiment (instead of hand-picking some "nice" explanations), and thus a realistic evaluation of the verbalization tool, the presented explanations were selected at random from the pool that fulfilled the criteria stipulated above. By default, the verbalization tool provides a simple highlighting of concept descriptions by displaying them in a different color (blue) than the surrounding text (black). Without such highlighting, the reading of the often long sentences containing long compound concept names is made unnecessarily tedious.

Results The two manipulated explanations were always correctly detected to be wrong by the participants except in one case. The non-manipulated explanations were mostly judged as correct, as predicted. One participant noticed the problem with the verbalization of $\boldsymbol{R}_{\sqsubseteq\equiv}$.

Figure 3 shows the subjects' averaged scores on Questions 1-10 for long and short explanations. The understandability of the explanations was judged favorably by the participants (Question 1), but not always found to be ideal. Participants' opinions on conciseness (Questions 2-4) turned out to be mixed and not too strong. Text quality (Questions 5-10) was also judged favorably. The answers of the participants to most questions are concentrated within relatively precise ranges and do not occupy the endpoints of the scale. Inter-rater reliability was assessed using Kendall's coefficient of concordance (corrected for ties) for the seven subjects' averaged scores on Questions 1-10 for long and short explanations (as shown in Fig. 3) and was found to be good ($W = 0.617$, $\chi^2(19) = 82.1$, $p < 0.001$). Concordance for individual answers measured separately for both experimental groups (with three and four participants each) was also good ($W = 0.759$, $\chi^2(79) = 180$, $p < 0.001$ and $W = 0.561$, $\chi^2(79) = 177$, $p < 0.001$).

This is a favorable outcome with respect to the question whether this experimental setup can be used as an instrument to detect differences in the scores related to the experimental manipulation, i.e. the shortening of the explanations, provided a larger number of participants to ascertain sufficient statistical power. For example, if a larger number of participants leads to a statistically significant difference in responses to Question 4 (the question whether the explanation should be made more concise), this would provide evidence for a positive effect of shortening on perceived conciseness of the explanations.

6 Conclusions

This work has, for the first time, investigated the generation of verbalized explanations for non-trivial derivations consisting of several inference steps. In the evaluation, verbalizations for up to seven inference steps (e.g. the running example in its unshortened form) were considered. To make such explanations more concise, we propose the use of an extended set of inference rules and the hiding of inference steps in the presentation. The presented experimental design can be used to test whether such adjustments impact the understandability and quality of the generated explanations. Our small-sample study provides a first indication of how these measures turn out, which will be helpful for conducting a larger follow-up study. Whereas the TONES repository was considered an instrumental choice for generating explanation material in a first step, more up-to-date corpora of ontologies should be considered in future studies. Furthermore, the experiment provided some first evidence supporting the understandability of such explanations in general, in spite of the considerable length of the derivations and the technical jargon used in the considered domains. In how far this is also the case for users with less background in technical domains than the participants of this study (in this case, all had a background in computer science) is to be investigated as part of future work.

Acknowledgments. We acknowledge the support of the Transregional Collaborative Research Center SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). We thank the three helpful reviewers and all experiment participants.

References

1. Androutsopoulos, I., Lampouras, G., Galanis, D.: Generating natural language descriptions from OWL ontologies: The NaturalOWL system. Journal of Artificial Intelligence Research 48, 671–715 (2013)
2. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope further. In: Clark, K., Patel-Schneider, P. (eds.) Proc. of the OWLED 2008 DC Workshop on OWL: Experiences and Directions. CEUR Workshop Proceedings, vol. 496 (2008)
3. Borgida, A., Franconi, E., Horrocks, I.: Explaining $\mathcal{ALC}$ subsumption. In: Horn, W. (ed.) Proc. of the 14th European Conf. on Artificial Intelligence, pp. 209–213. IOS Press (2000)
4. Cregan, A., Schwitter, R., Meyer, T., et al.: Sydney OWL syntax – Towards a controlled natural language syntax for OWL 1.1. In: Golbreich, C., Kalyanpur, A., Parsia, B. (eds.) Proc. of the OWLED 2007 Workshop on OWL: Experiences and directions, vol. 258 (2007)
5. Engelbrecht, P., Hart, G., Dolbear, C.: Talking rabbit: A user evaluation of sentence production. In: Fuchs, N.E. (ed.) Controlled Natural Language, pp. 56–64. LNAI 5972, Springer (2010)
6. Funk, A., Tablan, V., Bontcheva, K., Cunningham, H., Davis, B., Handschuh, S.: CLOnE: Controlled language for ontology editing. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) The Semantic Web (Proc. ISWC/ASWC 2007), pp. 142–155. LNCS 4825, Springer (2007)
7. Horridge, M.: Justification based explanations in ontologies. Ph.D. thesis, University of Manchester (2011)
8. Kaljurand, K., Fuchs, N.E.: Verbalizing OWL in Attempto Controlled English. In: Golbreich, C., Kalyanpur, A., Parsia, B. (eds.) Proc. of the OWLED 2007 Workshop on OWL: Experiences and directions, vol. 258 (2007)
9. Kazakov, Y., Klinov, P.: Goal-directed tracing of inferences in EL ontologies. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) The Semantic Web – ISWC 2014, LNCS, vol. 8797, pp. 196–211. Springer (2014)
10. Kazakov, Y., Krötzsch, M., Simancík, F.: The incredible ELK. Journal of Automated Reasoning 53(1), 1–61 (2014)
11. Kuhn, T.: The understandability of OWL statements in controlled English. Semantic Web 4(1), 101–115 (2013)
12. Liang, S.F., Scott, D., Stevens, R., Rector, A.: Ontoverbal: A generic tool and practical application to SNOMED CT. International Journal of Advanced Computer Science and Applications (IJACSA) 4, 227–239 (2013)
13. McGuinness, D.L.: Explaining Reasoning in Description Logics. Ph.D. thesis, Rutgers University (1996)
14. Nguyen, T., Power, R., Piwek, P., Williams, S.: Measuring the understandability of deduction rules for OWL. In: Lambrix, P., Qi, G., Horridge, M. (eds.) First International Workshop on Debugging Ontologies and Ontology Mappings (WoDOOM12), pp. 1–12. Linköping University Electronic Press (2012)
15. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Predicting the understandability of OWL inferences. In: The Semantic Web: Semantics and Big Data, pp. 109–123. Springer (2013)
16. Power, R.: OWL simplified English: A finite-state language for ontology editing. In: Kuhn, T., Fuchs, N.E. (eds.) Controlled Natural Language (Proc. CNL 2012), LNAI, vol. 7427, pp. 44–60. Springer (2012)

17. Schiller, M., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of $\mathcal{EL}$ subsumptions. Tech. rep., Universität Ulm (2017), https://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.090/Publikationen/2017/SSG17DL.pdf
18. Stevens, R., Malone, J., Williams, S., Power, R., Third, A.: Automating generation of textual class definitions from OWL to English. J. Biomedical Semantics 2(Suppl. 2), S5 (2011)
19. Wason, P.C.: Reasoning about a rule. The Quarterly Journal of Experimental Psychology 20(3), 273–281 (1968)