In the Eye of the Beholder: Which Proofs are Best?

(Extended Abstract)*

Stefan Borgwardt[1], Anke Hirsch[2], Alisa Kovtunova[1], and Frederik Wiehr[2]

[1] Institute of Theoretical Computer Science, TU Dresden, Germany
firstname.lastname@tu-dresden.de
[2] German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany
firstname.lastname@dfki.de

We conducted an experiment where we assessed participants' understanding of different representations of proofs for description logics [2]. The main formats we used are full formal proofs in a tree-shaped representation, e.g. based on consequence-based reasoning procedures [9, 18], and linearized translations of these proofs into text, e.g. as produced by various verbalization techniques [1, 11,16]. In addition, to find out how detailed proofs should be, we added shortened representations for each of these two versions, in which some (easy) reasoning steps were omitted or merged. We chose these four combinations since they are representative of the state-of-the-art in DL explanations. Differently from previous studies [13–16], we directly compared textual and formal proof formats.

Participants. 16 participants (four female) were assessed with a mean age of 23 (standard deviation = 1.71). Our international participants were recruited from undergraduate and graduate university students with basic knowledge of logic, which was required to understand the proofs. Participants were recruited via advertisements on mailing lists. Screening criteria were familiarity with first-order logic (e.g. through a lecture), a stable Internet connection, installing a video conference app with video access (on their mobile device or computer) and the permission to record their handwriting and voice during the experiment.

Conditions and Design. We used two different conditions with two levels each. One condition was the representational form of the proof, which was either textual or formal. The other condition was the length of the proof, which was either short or long. Thus, there were the four following condition combinations: *Long Text*, *Short Text*, *Long Formal*, and *Short Formal*. We used a 2×2 within-subjects design, which means that each participant saw all four combinations.

Material. The proofs were chosen such that they represent an unintuitive consequence, e.g. the unsatisfiability of a concept name, or that any amputation of a finger is also an amputation of the whole hand [3]. All four examples were chosen from the literature on DL explanations, in particular [3, 7, 12, 17]. For each of them, four different proof representations were manually created, not automatically generated, to make them comparable in difficulty.

* This is an abstract of the paper [4] which will appear at DL 2020.

$$\dfrac{\dfrac{\mathsf{CCl t} \sqsubseteq \exists \mathsf{ct}.\mathsf{C} \sqcap \forall \mathsf{ct}.\mathsf{C} \qquad \dfrac{\mathsf{CCl t} \sqsubseteq \mathsf{MaObj} \qquad \mathsf{MaObj} \sqsubseteq \exists \mathsf{ct}.\mathsf{At}}{\mathsf{CCl t} \sqsubseteq \exists \mathsf{ct}.\mathsf{At}}}{\mathsf{CCl t} \sqsubseteq \exists \mathsf{ct}.(\mathsf{At} \sqcap \mathsf{C})} \qquad \dfrac{\dfrac{\mathsf{C} \sqsubseteq \mathsf{Cmp}}{\mathsf{At} \sqcap \mathsf{C} \sqsubseteq \mathsf{At} \sqcap \mathsf{Cmp}} \qquad \mathsf{At} \sqcap \mathsf{Cmp} \sqsubseteq \bot}{\mathsf{At} \sqcap \mathsf{C} \sqsubseteq \bot}}{\mathsf{CCl t} \sqsubseteq \bot}$$

Since every cell culture is a material object and every material object contains an atom, every cell culture contains an atom. From the facts that every cell culture contains an atom and that every cell culture contains a cell and contains only cells, it follows that every cell culture contains something which is both an atom and a cell.

Every cell is a compound. Thus, any object which is an atom and a cell at the same time is also an atom and a compound. There is no object which is an atom and a compound at the same time. Therefore, there is no object which is both an atom and a cell.

Furthermore, since every cell culture contains something which is both an atom and a cell and there is no object which is both an atom and a cell, there is no cell culture.

Fig. 1. A formal and a textual representation of a proof. For the sake of presentation, in the formal proof we abbreviate the words "Atom", "Cell", "CellCulture", "MaterialObject", "Compound" and "contains" to "At", "C", "CClt", "MaObj", "Cmp" and "ct". For the experiment, the formal version was printed without abbreviation.

Figure 1 depicts a short formal and a short textual representation for one of the examples. Each of them (as well as their longer versions) were shown below a list of the involved TBox axioms ($\mathsf{Cell} \sqsubseteq \mathsf{Compound}$ etc.), a textual translation of these axioms (e.g. "Every cell is a compound."), as well as a short statement of the entailment ("The ontology above implies that there is no cell culture."). The full details of the experiment are available online[3].

To make sure the participants really understood the proofs a logic expert reviewed the video of each participant after each session. Due to the think-aloud technique the expert was able to follow the participant's thought and rated the video based on the participant's understanding on a scale from 1 (no understanding) to 3 (complete understanding).

Independent and Dependent Variables. To assess participants' experience we asked them how they would rate their experience with propositional, description, first order logic on a Likert-like scale from 1 (no knowledge) to 5 (expert). We evaluated how they rated the difficulty of each proof on a Likert-like scale from 1 (very easy) to 5 (very difficult). To compare the proof representations, we asked the participants to rank the proofs at the end of the experiment based on their comprehensibility (first rank = very easy, fourth rank = very difficult). It was possible to give several proofs the same rank. They were also asked to comment on the ranking and on what they liked and disliked about the proofs.

Hypotheses. We stated three hypotheses concerning the participants' self-rating of the difficulty of the proofs and their self-rated experience with logic.

[3] https://cloud.perspicuous-computing.science/s/Wmtmyp8JQNaF2AD

Hypothesis 1: It is easier to understand a short, concise explanation than a longer version (in the same representation format).

Hypothesis 2: Users with less experience in logic can understand the longer text better than a short formal proof.

Hypothesis 3: Users with more experience in logic can understand a long formal proof better than a long text.

Results. To compute the quantitative analyses IBM SPSS Statistics (Version 26) predictive analytics software for Windows [6] and the Macro PROCESS [5] was used. For all hypotheses, we used a p-value threshold of 0.05.

For *Hypothesis 1*, a multiple linear regression with contrast coding (K1, K2, K3) was conducted. K1 contrasted the textual representation against the formal one. K2 contrasted the short vs. long proofs and K3 the interaction between the two general conditions. The three contrasts explained 14.2% of variance in the rating after each proof, $R^2 = .14$, $F(3,\ 60) = 3.30$, $p < .05$. Only K2 was found to be a significant predictor in the linear regression, $\beta = -.29$, $t(60) = -2.42$, $p < .05$. This means that the participants rated the shorter proofs as being easier than the longer ones, which was independent of the presentation format. Thus, *Hypothesis 1* could be supported by our data.

For *Hypotheses 2* and *3*, we computed moderator analyses with the two condition combinations as a predictor, the experience as a moderator variable and the rating after each proof as the criterion. However, neither *Hypothesis 2* nor *3* was supported by our data. Experience with logic did not make a difference on the understanding of the different proof representations.

Additionally to the three hypotheses, we used Friedman's ANOVA for comparing the comprehensibility ranking of the proof representations at the end of the experiment (first rank = very easy, fourth rank = very difficult). It revealed a significant difference in the ranking of the condition combinations, $\chi^2(3) = 15.29$, $p < .01$ with a moderate effect size (Kendall's $W = .32$). For the post-hoc pairwise comparisons Bonferroni correction was used which resulted in a p-threshold of 0.008, resulting in only two significant comparisons.

The participants' ranking of condition combinations is shown in Figure 2. The combination *Short Text* was preferred over *Long Text*, $Z = 1.53$, $p < .008$. The median ranking for *Short Text* and *Long Text* was 2 and 3.5, respectively. Additionally, *Short Formal* was preferred over *Long Text*, $Z = 1.50$, $p < .008$. *Short Formal* had the lowest median ranking with 1.50. Both comparisons showed moderate effect sizes, $r = 0.38$ for both. Median ranking for *Long Formal* was 2.

Only one participant chose *Long Text* on the first rank. However, nobody put *Long Text* on the second rank, but 15 chose the third or fourth rank for it. Thus, most participants ranked it as (very) difficult. *Short Text* was never assigned the fourth rank, but by 13 participants it was considered very easy or easy.

Discussion. Short proofs were rated as being easier than long proofs, independent of the presentation format. Thus, future experiments and theoretical approaches should focus on shortening proofs. With our data, *Hypotheses 2* and *3* could not be supported. However, the rankings and discussions showed individual user preferences between formal and textual representations. One possibility

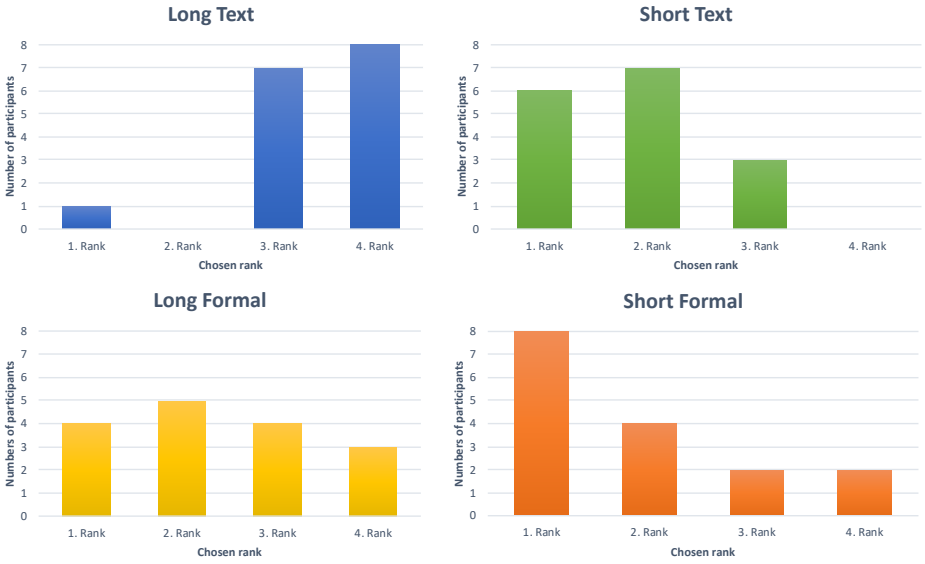

Fig. 2. The participants' ranking of conditions with 1 = very easy and 4 = very difficult

to further assess this could be to include experts working in the field of logic, like computer scientists and mathematicians teaching logic at a university. This way, there could be a clearer distinction between novices, e.g. students having attended a single lecture about logic, vs. experts. Maybe then one could find an influence of experience on the perception of difficulty of the proofs.

On the other hand, the ultimate goal is to explain logical conclusions to domain experts who are not familiar with logic. Here, an interesting direction of study is to generate (concise) textual explanations [1, 11, 16], or perhaps a combination of graphical and textual elements to better convey the structure of a proof while still providing each (derived) axiom in a readable form.

From a procedural point of view, it would be preferable to use a between-subjects design (different people test each condition) instead of within-subjects (when the same person tests all the conditions), to minimize learning effects, which however requires more participants. Of course we would also like to compare other proof representations, e.g. pure justifications, linear vs. non-linear formats, mixed formal/textual presentations as mentioned above, incorporating annotations such as axiom numbering or coloring, and most importantly interactive approaches such as the proof plugin for the Protégé ontology editor [8]. The main goal with these different representations should always be usability, which has to be assessed experimentally.

As was demonstrated by the participants' different opinions and preferences about proof representations, it makes sense to incorporate the user as an active element in the design of a suitable presentation. User modeling [10] can help make automatic design decisions, by taking into account user preferences or the

user's existing knowledge, e.g. in the form of a *background ontology* that the user is assumed to know intimately.

Acknowledgements This work was partially supported by DFG grant 389792660 as part of TRR 248 (https://perspicuous-computing.science). We are also grateful to our colleagues who forwarded the advertisement for the experiments to their students, and of course to the 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). https://doi.org/10.1613/jair.4017
2. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017). https://doi.org/10.1017/9781139025355
3. Baader, F., Suntisrivaraporn, B.: Debugging SNOMED CT using axiom pinpointing in the description logic $\mathcal{EL}^+$. In: Proc. of the 3rd Conference on Knowledge Representation in Medicine (KR-MED'08): Representing and Sharing Knowledge Using SNOMED. CEUR-WS, vol. 410 (2008), http://ceur-ws.org/Vol-410/Paper01.pdf
4. Borgwardt, S., Hirsch, A., Kovtunova, A., Wiehr, F.: In the Eye of the Beholder: Which Proofs are Best? In: Borgwardt, S., Meyer, T. (eds.) Proc. of the 33th Int. Workshop on Description Logics (DL 2020) (2020), to appear
5. Hayes, A.F.: Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Publications (2017)
6. IBM Corp.: IBM SPSS Statistics for Windows [computer software], https://www.ibm.com/products/spss-statistics
7. Kalyanpur, A.: Debugging and Repair of OWL Ontologies. Ph.D. thesis, University of Maryland, College Park, USA (2006), http://hdl.handle.net/1903/3820
8. Kazakov, Y., Klinov, P., Stupnikov, A.: Towards reusable explanation services in Protege. In: Artale, A., Glimm, B., Kontchakov, R. (eds.) Proc. of the 30th Int. Workshop on Description Logics (DL'17). CEUR Workshop Proceedings, vol. 1879 (2017), http://www.ceur-ws.org/Vol-1879/paper31.pdf
9. Kazakov, Y., Krötzsch, M., Simancik, F.: The incredible ELK – from polynomial procedures to efficient reasoning with $\mathcal{EL}$ ontologies. J. Autom. Reasoning **53**(1), 1–61 (2014). https://doi.org/10.1007/s10817-013-9296-3
10. Kobsa, A., Wahlster, W.: User models in dialog systems. Springer (1989)
11. Kuhn, T.: The understandability of OWL statements in controlled english. Semantic Web **4**(1), 101–115 (2013). https://doi.org/10.3233/SW-2012-0063
12. Meehan, T.F., Masci, A.M., Abdulla, A., Cowell, L.G., Blake, J.A., Mungall, C.J., Diehl, A.D.: Logical development of the cell ontology. BMC Bioinformatics **12**(1), 6 (2011). https://doi.org/10.1186/1471-2105-12-6
13. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Measuring the understandability of deduction rules for OWL. In: Proceedings of the First International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2012, Galway, Ireland, October 8, 2012. pp. 1–12 (2012), http://www.ida.liu.se/~patla/conferences/WoDOOM12/papers/paper4.pdf
14. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Predicting the understandability of OWL inferences. In: Cimiano, P., Corcho, Ó., Presutti, V., Hollink, L.,

Rudolph, S. (eds.) The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7882, pp. 109–123. Springer (2013). https://doi.org/10.1007/978-3-642-38288-8_8

15. Schiller, M.R.G., Glimm, B.: Towards explicative inference for OWL. In: Informal Proceedings of the 26th International Workshop on Description Logics, Ulm, Germany, July 23 - 26, 2013. pp. 930–941 (2013), http://ceur-ws.org/Vol-1014/paper_36.pdf
16. Schiller, M.R.G., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of EL subsumptions. In: Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. (2017), http://ceur-ws.org/Vol-1879/paper43.pdf
17. Schulz, S.: The role of foundational ontologies for preventing bad ontology design. In: Proc. of the 1st Int. Workshop on BadOntoloGy (BOG'18), part of The Joint Ontology Workshops (JOWO'18). CEUR Workshop Proceedings, vol. 2205. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2205/paper22_bog1.pdf
18. Simancik, F., Kazakov, Y., Horrocks, I.: Consequence-based reasoning beyond horn ontologies. In: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. pp. 1093–1098 (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-187