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

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

Abstract. Although logical entailments are often considered "explainable", experience with justifications in DLs has shown that explaining why a logical consequence holds still requires some effort. However, a full formal proof of a DL entailment may be considered too long-winded, and a textual representation of a proof may be preferred. It may also depend on the user's experience and individual preferences which representation of a proof constitutes a good explanation for them. Building on previous work on explaining DL consequences to users, we ran an experiment to compare 4 different forms of proofs: formal proofs and textual proofs, which are either very detailed or condensed. A multiple linear regression with contrast coding revealed that the participants rated short proofs as being easier than long proofs, independent of their representation. On the other hand, we could not verify any influence of prior experience with logic on how easy or difficult the different kinds of proofs were considered.

1 Introduction

Explanations of automated decisions are currently an important topic of research. However, apart from the popular discussion about how "explainable" different AI methods are, the main task of explanations is *understanding*, i.e., that the information transmitted is actually received by the human user, and hence is a topic for social science research [21]. This large existing body of research needs to be taken into account when designing explainable AI systems. Particular lessons are that the *context* of an explanation is important, i.e., an explanation depends on the environment as well as the user, and that an effective explanation should be a *dialog* between the explainer and the explainee, allowing for a cycle of questions and clarifications.

In the DL community, research on explanations first focused on proofs for explaining entailments [7,19], but it was quickly realized that often it is enough to point out the responsible axioms from the ontology, i.e., so-called *justifications* [5,10,28]. There have been some experiments on determining factors that make a justification hard to understand [11]. However, recent research has been going back towards providing (partial) proofs [12,15]. Explanations using proofs with intermediate steps have been investigated through several approaches, for example modified sequent calculi [7], an empirical study using around 500 OWL ontologies to find new "understandable" inference rules [22], a probabilistic model for measuring the understandability of proofs consisting of a few inference rules [23], and an approach for designing inference rules keeping the bounded cognitive resources of the user in mind [8]. In parallel, several approaches for converting description logic axioms and proofs into textual representations have been developed and evaluated [2,18,24,25].

In an effort to understand which of these approaches are most promising for improving explainability, our experiment continues in this line of research. We investigated which representations of DL proofs are preferred by users (who had some prior experience in logic). The main formats we used are full formal proofs in a tree-shaped representation, e.g. based on consequence-based reasoning procedures [16,30], and linearized translations of these proofs into text, e.g. as produced by various verbalization techniques [2,18,25]. 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. We did not consider plain justifications as explanations in our experiment, since we chose examples that were identified as more challenging in the DL literature, and thus require more in-depth explanation. Differently from previous studies, we directly compared textual and formal proof formats.

Our main finding is that shorter proofs were rated as being easier than ones that contained detailed reasoning chains. In particular, the long textual proof representation was often considered to be too long-winded and repetitive. However, the experience level of our participants did not seem to influence their preferences. Nevertheless, there clearly are differences in how different individuals were working with the different proof representations.

2 Proofs

We assume a basic familiarity with DLs, in particular $\mathcal{ALC}$ [4]. Let $\mathcal{O}$ be an ontology and α be an axiom that follows from $\mathcal{O}$ ($\mathcal{O} \models \alpha$), which represents a surprising or unintuitive consequence for the user. A *justification* is a minimal subset $\mathcal{J} \subseteq \mathcal{O}$ such that $\mathcal{J} \models \alpha$, which already points to some of the axioms from $\mathcal{O}$ that are responsible for α. However, actually understanding why α follows may require a more detailed proof. Informally, a *proof* is a tree of connected inference steps $\frac{\alpha_1 \dots \alpha_n}{\alpha}$, where each step is sound, i.e., $\{\alpha_1, \dots \alpha_n\} \models \alpha$ holds (see Fig. 1). Usually, one imagines such a proof to be built using *inference rules* from an appropriate calculus [3, 30]. However, there also exist approaches to generate DL proofs that are not based on proof calculi. They usually start with a justification, and extend it with intermediate axioms (*lemmas*) using heuristics [10,11], concept interpolation [27], or forgetting [1].

$$\dfrac{\dfrac{\mathsf{A} \sqsubseteq \exists \mathsf{r}.\top \qquad \mathsf{A} \sqsubseteq \forall \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})}{\mathsf{A} \sqsubseteq \exists \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})} \qquad \mathsf{C} \sqcap \mathsf{B} \sqsubseteq \bot}{\mathsf{A} \sqsubseteq \bot} \qquad \begin{aligned} \mathcal{O} = \{\ & \mathsf{A} \sqsubseteq \exists \mathsf{r}.\top, \\ & \mathsf{C} \sqcap \mathsf{B} \sqsubseteq \bot, \\ & \mathsf{A} \sqsubseteq \forall \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C})\ \} \end{aligned}$$

Fig. 1. A proof for the unsatisfiability of A w.r.t. $\mathcal{O}$, i.e., that $\mathcal{O} \models \mathsf{A} \sqsubseteq \bot$.

Since the only requirement is soundness of each proof step, adjacent steps can be merged to obtain shorter proofs, e.g. for the proof in Fig. 1 we could merge the two steps:

$$\dfrac{\mathsf{A} \sqsubseteq \exists \mathsf{r}.\top \qquad \mathsf{A} \sqsubseteq \forall \mathsf{r}.(\mathsf{B} \sqcap \mathsf{C}) \qquad \mathsf{C} \sqcap \mathsf{B} \sqsubseteq \bot}{\mathsf{A} \sqsubseteq \bot}$$

This directly corresponds to the justification $\mathcal{J} = \mathcal{O}$. Thus, a justification can also be seen as a very short proof. However, it can be useful to highlight intermediate inference steps, for example to pinpoint the precise cause of an unintended entailment α: even if all axioms from the ontology seem reasonable to the user, at some point in the proof they will start to have unintuitive consequences, which can be used to track down the problem.

A textual representation of a proof is necessarily a *linearization*, where the inference steps are explained in a sequence, for example in a top-down left-right order. A text corresponding to the formal proof in Fig. 1 could be the following:

> Since every A has an r-successor and every A has only r-successors that are B and C, every A has an r-successor that is B and C. Since every A has an r-successor that is B and C and there is no object which is C and B at the same time, there is no A.

3 Experiment

We conducted an experiment where we assessed participants' understanding of these different proof representations. The full details of the experiment are available online[1].

3.1 Description of the experiment

Introduction and Goals. With the current experiment, we attempt to find which proof representation is most understandable, also considering experience with logic. The objective is to find a difference in the difficulty rating of longer or shorter proofs with either textual or a formal representation.

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

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.

In Table 1, the descriptive statistics of the participants' self-rated experience with logic can be found.

Table 1. Descriptive statistics for participants' self-rated experience with logic. 1 = *no knowledge at all / no experience*, 5 = *expert / a lot of experience*. n = 16, *SD* = standard deviation.

	Mean	*SD*	*Range*
Propositional Logic	3.25	1.00	1-5
First-Order Logic	2.94	0.68	2-4
Description Logic	2.31	0.79	1-4

Conditions and Design. We used two different conditions (factors) 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* (see Table 2). We used a 2 × 2 within-subjects design, which means that each participant saw all four combinations.

Table 2. The distribution of the conditions and topics into four participant groups.

Topic	*Group 1*	*Group 2*	*Group 3*	*Group 4*
"Cell Culture"	Short Text	Long Formal	Long Text	Short Formal
"DNA"	Short Formal	Short Text	Long Formal	Long Text
"T-Cells"	Long Text	Short Formal	Short Text	Long Formal
"Amputation"	Long Formal	Long Text	Short Formal	Short Text

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 [6]. We chose to employ real-domain examples rather than versions with abstract concept and role names to increase the readability and motivation, as was also done in [25]. Additionally, the domain for the experiment phase must not be too familiar to participants: with background knowledge they can easily spot controversial axiom(s) in an ontology and do not require a proof for the unintuitive entailment. Therefore, all four examples were chosen from the medical domain and were adapted from examples in the literature on DL explanations, in particular [6, 14, 20, 29]. For each of them, four different proof representations (see Table 2) were manually created, not automatically generated, to make them comparable in difficulty. To keep consistency among the long formal proofs, we used a fix set of basic inference rules based on the one for ELK [16]. Short formal proofs were obtained by omitting tautologies and merging two transitivity or transitivity and additive rule instances into one. Textual proofs were obtained by (naively) applying a set of phrase templates to every inference step in corresponding short or long formal proofs. For example, for a rule instance with the premises *prem1*, *prem2* and the conclusion *concl*, verbalisation of this instance can be as follows: “From the facts that VERB(*prem1*) and VERB(*prem2*), it follows that VERB(*concl*)”, where VERB is a unified template verbalization of DL axioms according to their logical constructors.

In Figure 2, we depict a short formal and a short textual representation for one of the examples; the longer variants can be found in the appendix. Each proof was shown below a list of the involved ontology axioms on the left ($\mathsf{Cell} \sqsubseteq \mathsf{Compound}$ etc.), a textual translation of these axioms on the right (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.”).

Procedure and Tasks. Due to the Corona pandemic, we could not conduct the experiment in person as planned. Instead we used video conferences with Zoom[2] and GoToMeeting[3] to communicate with our participants.

The experiment was conducted in English. Every participant got an individual appointment. They were asked via e-mail whether they had a printer. If this was not the case, then the experiment sheets were send to them via post. If they had a printer, then they received the sheets as a PDF via e-mail with specific instructions on how to print them. One day before their appointment, the participants received an e-mail with a reminder and the link to the video conference meeting. To start the experiment, the participant was welcomed by the experimenter. The purpose and the procedure of the experiment was explained. They were then instructed to put the printed or sent papers in front of them and a link to the questionnaire and a participant number were sent to them. They went through the first six pages of the questionnaire asking them about their demographic data and experience with logic. To refresh participant’s memory and to make sure everyone started with at least the same basic knowledge about

[2] `https://zoom.us`

[3] `https://www.gotomeeting.com`

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

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. 2. 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 in the same font size as the textual one.

description logics, they were shown a 10-minute video with a crash course on description logics.

Afterwards, they were asked to arrange the camera setup in such a way that the papers in front of them and their hands were visible. Then the experimenter started recording the participant. To demonstrate both proof representations (textual and formal), a training phase was conducted. The participant saw a textual proof and a formal proof consecutively. The experimenter explained that they would see an ontology which implies an unintuitive statement. We used the think-aloud technique, in which the participant was asked to read the ontology and the proof carefully and simultaneously summarize the proof out loud. They were encouraged to take notes on the paper, point or draw connections between the axioms and the proof and to explain their line of reasoning. After each proof they were asked to answer the following two questions and to mark their answers on the sheets: "Which part of the proof was the most difficult to understand and why?" and "Which axioms in the ontology do you think are most responsible for the unintuitive consequence and why?" During the training phase the participant was allowed to ask questions. After the two training examples, the experiment phase began. One after another, they saw four ontologies, each of which implied an unintuitive statement. The instructions were the same as in the training phase except that no questions from the participant were answered by the experimenter. They were reminded to think aloud while working through the proofs. To minimize the impact of the order in which they saw the different proof representations, we used a Latin square, so every combination of the two conditions was seen first by one fourth of the participants (see Table 2).

After each proof, the participant filled out one question in the questionnaire about the perceived difficulty of the proof and received the same two questions as in the training phase. Subsequently, the final question in the questionnaire asked the participant to rank the proofs based on their comprehensibility (first rank = very easy, fourth rank = very difficult). It was possible to give several proofs the same rank. They were asked to comment on the ranking, what they liked and disliked about the proofs and afterwards also if they noticed any differences or similarities between the proofs. They were then told that there were two different representational forms of proofs and asked which they preferred and why.

Finally, the experimenter made sure that the participant submitted the questionnaire, stopped the recording and told them to fill in their participant number on each sheet, scan it or take a photo and send us the final sheets.

To assure a stable video conference connection, the experimenter's video was off the whole time. To match the video and the questionnaire after the experiment, each participant was given the participant number, i.e., an individual code which they had to use in the questionnaire and on the sheets and under which we saved the video. Overall the experiment lasted around one and a half hours. After receiving all six pages (two for the training and four for the experiment phase), we sent them a code for an Amazon voucher worth 20 €.

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' logic experience we asked them how they would rate their experience with propositional, description, first order logic on a Likert-like rating scale from 1 (no knowledge) to 5 (expert). We further evaluated how they rated the difficulty of each proof on a Likert-like rating scale from 1 (very easy) to 5 (very difficult). To compare the different proof representations, we asked the participants to rank the proofs at the end of the experiment based on their comprehensibility.

Hypotheses. We stated three hypotheses concerning the participants' self-rating of the difficulty of the proofs.

Hypothesis 1: It is easier to understand a short, concise explanation than a longer version (in the same representation format). This will be shown by a lower difficulty rating for the short proofs than for the long proofs.

Hypothesis 2: Users with less experience in logic can understand the longer text better than a short formal proof. This will be shown by a lower difficulty rating of the long textual proof.

Hypothesis 3: Users with more experience in logic can understand a long formal proof better than a long text. This will be shown by a lower difficulty rating of the long formal proof.

Problems and Limitations. Originally, this experiment was planned as in-person: all the technical equipment as well as a suitable room would have been arranged by us to minimize possible distractions. However, due to the Corona pandemic, we had to re-design the experiment to an online setting. As immediate drawbacks of a less controlled environment, we encountered i) technical problems with the video conference due to unstable Internet connections, ii) poor quality of video and audio recordings, e.g. a low-resolution phone camera, iii) compatibility issues with different browsers, e.g. for video playback, iv) concerns about privacy for receiving and storing the video and audio recordings of participants and their homes, v) various interruptions related to the environment, i.e., family members or domestic animals.

Additionally, there was a more conceptual problem concerning prior knowledge about the example domains. For example, if a participant sees a (for them) contradictory axiom in the ontology (e.g. "every vegetarian is a person," because they would consider herbivorous animals to also be vegetarians; or "every artist is a professional," since there exist also amateur artists),[4] they would not focus anymore on the actual task of understanding why this ontology implies the goal axiom.

3.2 Results

Quantitative Results. To compute the quantitative analyses IBM SPSS Statistics (Version 26) predictive analytics software for Windows [13] and the Macro PROCESS [9] 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 want to assess if the strength of the relationship between the independent variables (condition combinations) and the dependent variable (rating after each proof) changes with the third variable (experience), i.e., whether there is an interaction between the conditions and the experience. 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. To put the conditions into the moderator analysis for *Hypothesis 2*, the *Long Text* was coded with +1 and *Short Formal* with -1 (and similarly for *Hypothesis 3* with *Long Text* and *Long Formal*). For each self-rated measure of experience (propositional logic, first-order logic, description logic), one moderation analysis was computed.

[4] These occurred in the training examples.

Table 3. Values of the moderator analyses of experience with propositional and first-order logic for *Hypothesis 2*; df = 28.

		β	t	p
Propositional	Main effect of condition	0.56	2.93	<.01
	Main effect of experience	-0.52	-2.60	<.05
	Interaction condition × experience	-0.15	-0.76	0.456
First-order	Main effect of condition	0.56	2.82	<.01
	Main effect of experience	-0.66	-2.17	<.05
	Interaction condition × experience	0.01	0.03	0.977

For *Hypothesis 2*, the moderation analyses for the self-rated experience with propositional and first-order logic revealed a significant model, $R^2 = .25$, $F(3,\ 28) = 3.07$, $p < .05$ and $R^2 = .31$, $F(3,\ 28) = 4.21$, $p < .05$, respectively, but only the two main effects were shown to be significant and not the interaction (see Table 3). Thus, there was no significant moderation. The main effect of the condition combinations showed that the *Short Formal* representation was rated as being easier than the *Long Text.* The main effect of experience means that, the more experience the participants had, the easier they rated the proof. However, the two main effects were independent from each other and showed no interaction, which means that neither the experience with propositional logic nor the experience with first-order logic influenced the relationship between the condition and the rating. For the self-rated experience with description logic, the moderation model was not found to be significant. Thus, *Hypothesis 2* could not be supported by our data. Experience with logic did not make a difference on the understanding of the different proof representations.

For *Hypothesis 3*, from the three moderation models, only the model with the self-rated experience with propositional logic turned out to be significant, $R^2 = .25$, $F(3,\ 28) = 3.07$, $p < .05$. Only the main effect of experience was significant here, $\beta = -.43$, $t(28) = -2.10$, $p < .05$. Thus, also *Hypothesis 3* could not be supported by our data. Experience with logic did not make a difference on the understanding of the different long conditions.

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 3. 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. Their statement can be found in the next subsection. 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.

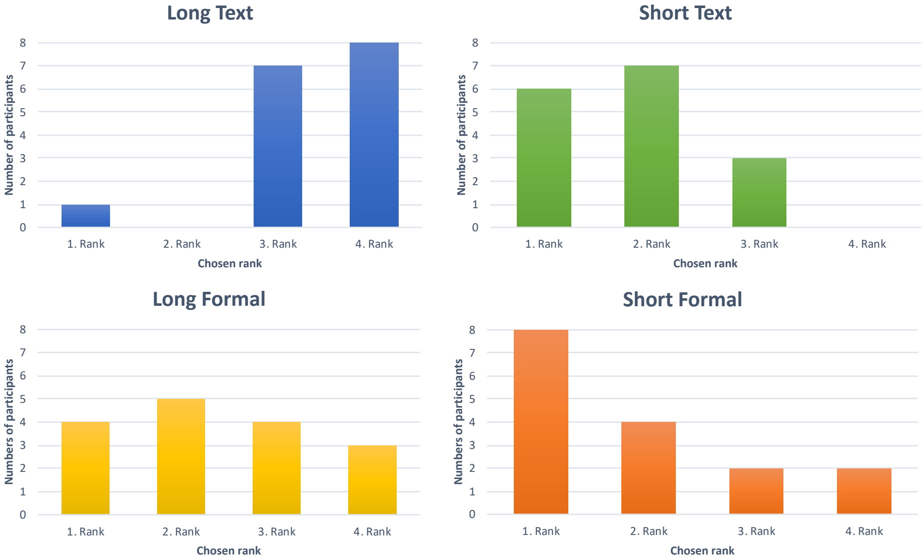

Fig. 3. The participants' ranking of conditions with 1 = very easy and 4 = very difficult

Qualitative Results. For the qualitative evaluation, we transcribed the audio recordings into text and arranged the texts according to the condition or condition combination. Then, we compared the participants' statements on each condition based on similarity and differences according to the hypotheses and grouped them contentwise.

In general, the participants' statements also supported the view that the formal proofs were preferred over the textual proofs. Participants described the formal proofs as "easier to understand" and also as "clearer". One participant mentioned that for them it was "easier to find certain parts" of the proof. Another one that that their "orientation is better" within the proof, in order to go through it step-by-step, and that it is "easy to follow the proof". The textual proofs were often characterized as inconvenient, "less understandable" or "hard to understand" or even "annoying".

On the other hand, one participant stated that "for short proofs text is ok." Only one participant said that they "preferred a proof explained through words rather than through a schema." The participant commented on it as follows: "Texts add more redundancy and this redundancy helps with connecting. My intuition for language can spot inconsistencies better. So, if a domain is abstract or unknown and a proof is complex (like in Mathematics) and requires to put together several pieces and arguing about them, wording might help a lot."

Many participants were struggling with inferences involving both ∀- and ∃-quantifiers in formal proofs. However, its textual equivalent or a second encounter of ∀+∃ were perceived much better (this combination appeared already in the training examples). Several participants mentioned that a graphical presentation is better when there are parallel threads: "I could clearly see what and where they come from. And then forget about what I just learned so I could go to another part. And once I again need this part, I know exactly where it came from later." Also, some participants found it very difficult to deal with longer domain-specific terms, since it requires more time to understand and to keep track of them. These opinions confirm the motivation behind [8] for restricting the length of axioms and developing a proof system that takes into account bounded cognitive resources.

4 Discussion and Conclusions

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 clearly showed individual user preferences between formal and textual representations. For further research in this direction, the questions from [26] could be used to get another quantitative approach. One possibility to further assess the difference between formal and textual representation 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 [2, 18, 25], 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 [15]. 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 [17] 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.

Another take-away message we learned from our experiment is that preparing such experiments to take place online from the beginning also has an upside since then they are more resilient against unforeseen events.

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 experiment to their students, and of course to the participants.

References

1. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding small proofs for description logic entailments: Theory and practice. In: Albert, E., Kovacs, L. (eds.) LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning. EPiC Series in Computing, vol. 73, pp. 32–67. EasyChair (2020). `https://doi.org/10.29007/nhpp`
2. 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`
3. Baader, F., Brandt, S., Lutz, C.: Pushing the $\mathcal{EL}$ envelope. In: Kaelbling, L.P., Saffiotti, A. (eds.) Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI'05). pp. 364–369. Professional Book Center (2005), `http://ijcai.org/Proceedings/09/Papers/053.pdf`
4. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017). `https://doi.org/10.1017/9781139025355`
5. Baader, F., Peñaloza, R., Suntisrivaraporn, B.: Pinpointing in the description logic $\mathcal{EL}^+$. In: Proc. of the 30th German Annual Conf. on Artificial Intelligence (KI'07). Lecture Notes in Computer Science, vol. 4667, pp. 52–67. Springer, Osnabrück, Germany (2007). `https://doi.org/10.1007/978-3-540-74565-5_7`
6. 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`
7. Borgida, A., Franconi, E., Horrocks, I.: Explaining $\mathcal{ALC}$ subsumption. In: ECAI 2000, Proceedings of the 14th European Conference on Artificial Intelligence, Berlin, Germany, August 20-25, 2000. pp. 209–213 (2000), `http://www.frontiersinai.com/ecai/ecai2000/pdf/p0209.pdf`

8. Engström, F., Nizamani, A.R., Strannegård, C.: Generating comprehensible explanations in description logic. In: Informal Proceedings of the 27th International Workshop on Description Logics, Vienna, Austria, July 17-20, 2014. pp. 530–542 (2014), http://ceur-ws.org/Vol-1193/paper_17.pdf
9. Hayes, A.F.: Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Publications (2017)
10. Horridge, M.: Justification Based Explanation in Ontologies. Ph.D. thesis, University of Manchester, UK (2011), https://www.research.manchester.ac.uk/portal/files/54511395/FULL_TEXT.PDF
11. Horridge, M., Bail, S., Parsia, B., Sattler, U.: Toward cognitive support for OWL justifications. Knowledge-Based Systems **53**, 66–79 (2013). https://doi.org/10.1016/j.knosys.2013.08.021
12. Horridge, M., Parsia, B., Sattler, U.: Justification oriented proofs in OWL. In: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. pp. 354–369 (2010). https://doi.org/10.1007/978-3-642-17746-0_23
13. IBM Corp.: IBM SPSS Statistics for Windows [computer software], https://www.ibm.com/products/spss-statistics
14. Kalyanpur, A.: Debugging and Repair of OWL Ontologies. Ph.D. thesis, University of Maryland, College Park, USA (2006), http://hdl.handle.net/1903/3820
15. 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
16. 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
17. Kobsa, A., Wahlster, W.: User models in dialog systems. Springer (1989)
18. 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
19. McGuinness, D.L.: Explaining Reasoning in Description Logics. Ph.D. thesis, Rutgers University, NJ, USA (1996). https://doi.org/10.7282/t3-q0c6-5305
20. 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
21. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
22. 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
23. 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
24. 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

25. 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
26. Schiller, M.R., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of el subsumptions. Description Logics **1879** (2017)
27. Schlobach, S.: Explaining subsumption by optimal interpolation. In: Alferes, J.J., Leite, J.A. (eds.) Proc. of the 9th Eur. Conf. on Logics in Artificial Intelligence (JELIA'04). Lecture Notes in Computer Science, vol. 3229, pp. 413–425. Springer-Verlag (2004). https://doi.org/10.1007/978-3-540-30227-8_35
28. Schlobach, S., Cornet, R.: Non-standard reasoning services for the debugging of description logic terminologies. In: Gottlob, G., Walsh, T. (eds.) Proc. of the 18th Int. Joint Conf. on Artificial Intelligence (IJCAI 2003). pp. 355–362. Morgan Kaufmann, Acapulco, Mexico (2003), http://ijcai.org/Proceedings/03/Papers/053.pdf
29. 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
30. 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

A Long Proofs

On the following two pages we depict the long versions of the proofs in Figure 2:

> Every cell culture contains a cell and contains only cells. Thus, every cell culture contains only cells. 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 only cells, it follows that every cell culture contains something which is both an atom and a cell.
>
> Trivially, every object which is both an atom and a cell is an atom and a cell. Thus, every object which is both an atom and a cell is a cell. Every cell is a compound. Therefore, every object which is both an atom and a cell is a compound. Trivially, every object which is both an atom and a cell is an atom and a cell. Thus, every object which is both an atom and a cell is an atom. From the facts that every object which is both an atom and a cell is an atom and that every object which is both an atom and a cell is a compound, it follows that every object which is both an atom and a cell is both an atom and a compound. There is no object which is both a compound and an atom 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.

$$\dfrac{\dfrac{\dfrac{\mathsf{CellCulture} \sqsubseteq \exists \mathsf{contains}.\mathsf{Cell} \sqcap \forall \mathsf{contains}.\mathsf{Cell}}{\mathsf{CellCulture} \sqsubseteq \forall \mathsf{contains}.\mathsf{Cell}} \quad \dfrac{\mathsf{CellCulture} \sqsubseteq \mathsf{MaterialObject} \quad \mathsf{MaterialObject} \sqsubseteq \exists \mathsf{contains}.\mathsf{Atom}}{\mathsf{CellCulture} \sqsubseteq \exists \mathsf{contains}.\mathsf{Atom}}}{\mathsf{CellCulture} \sqsubseteq \exists \mathsf{contains}.(\mathsf{Atom} \sqcap \mathsf{Cell})} \quad \dfrac{\dfrac{\dfrac{\dfrac{}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \mathsf{Atom} \sqcap \mathsf{Cell}}}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \mathsf{Atom}} \quad \dfrac{\dfrac{\dfrac{}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \mathsf{Atom} \sqcap \mathsf{Cell}}}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \mathsf{Cell}} \quad \mathsf{Cell} \sqsubseteq \mathsf{Compound}}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \mathsf{Compound}}}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \mathsf{Atom} \sqcap \mathsf{Compound}} \quad \mathsf{Atom} \sqcap \mathsf{Compound} \sqsubseteq \bot}{\mathsf{Atom} \sqcap \mathsf{Cell} \sqsubseteq \bot}}{\mathsf{CellCulture} \sqsubseteq \bot}$$