Applying a Model of Text Comprehension to Automated Verbalizations of $\mathcal{EL}$ Derivations

Tanja Perleth, Marvin Schiller, and Birte Glimm

University of Ulm, Germany
{tanja.perleth, marvin.schiller, birte.glimm}@uni-ulm.de

Abstract. Ontology verbalization techniques have been introduced to generate natural-language texts from ontology axioms and deduction steps. This allows users without knowledge of formal languages (e.g. OWL) to follow deductive inferences derived in ontologies. Since these explanations can be generated from different ontologies (not necessarily developed with verbalization in mind) and from potentially long, complex derivations, the question is to what extent these explanations are readable, understandable and useful for human readers. We apply the cognitive-psychology-based model of Kintsch and van Dijk to explanations generated automatically using a verbalization system for derivations in the $\mathcal{EL}$ fragment of description logic to model the reading process of a human reader. This allows for checking whether the generated explanations have a coherent text base (according to Kintsch and van Dijk's model) and for re-ordering the presented steps accordingly.

1 Introduction

To facilitate the work with ontologies, verbalization techniques and tools have been developed (e.g. [15, 9, 11, 1] to mention just a few). They serve to generate natural-language texts for ontology axioms and inferences, which is helpful both for ontology debugging (to explain why an inference holds) and for users not familiar with formal languages such as OWL. Systems for verbalizing inferences (e.g. [11, 14]) typically use consequence-based reasoning (i.e. inference rules) and text patterns to generate natural-language explanations, which, depending on the complexity of the required inference steps, can become long and hard to read. Therefore, it needs to be established in how far these explanations are understandable and usable for human readers and – if necessary – how they can be improved. We explore the application of a cognitive text processing model to automatically generated explanations to assess the complexity involved in reading and understanding them. This complements previous work aiming at characterizing the cognitive complexity of individual inference rules and the cognitive complexity of so-called *justifications* (cf. Sect. 2).

The contribution of this work consists in an implementation of the cognitive text processing model proposed by Kintsch and van Dijk [7], and its application to automatically generated explanations for subsumptions in the $\mathcal{EL}$ fragment of description logic (DL) by a verbalization tool, as detailed in Sect. 3. This includes

the transformation of generated texts to an abstract semantic representation of its surface structure, a so-called *text base* that is formed out of *propositions* (as defined by Kintsch and van Dijk). We demonstrate how the model helps to identify explanations that are deemed to be difficult to understand, namely those that lack a coherent text base, and those that require long-term memory search. The main idea is that the complexity of understanding an explanation depends not only on the employed inference steps, but is also affected by text construction. In Sect. 4 we report on a first study that compares generated explanations with a coherent text base with a corresponding version without coherent text base.

2 Related Work

Related work includes verbalization techniques for ontologies and measures for the cognitive complexity of inferences (and explanations generated for them).

2.1 Verbalization

Ontology verbalization techniques have been proposed to present ontological axioms and derivations in the form of natural-language texts. Approaches that focus on the verbalization of axioms (and descriptions of classes based on sets of axioms) include a tool developed by the SWAT project [15], the OntoVerbal verbalizer [9] and NaturalOWL [1]. Approaches that address the verbalization of derivations include the CLASSIC system [10] and the work of Borgida et al. [2] and Nguyen et al. [11]. In this work, we use our own verbalization tool (henceforth referred to as "verbalizer") [14, 13]. Similarly to the above-mentioned approaches (and similarly to the "tracing" facility of the ELK reasoner [5]), it uses consequence-based reasoning to construct a proof tree for a derivation. These proofs are then translated to natural-language texts using patterns (a comprehensive list is found in [14, Fig. 1]). During the translation, the proof tree is traversed using post-order traversal; i.e. for an inference step to be explained, the derivations of the premises are explained before the conclusion is stated. In this paper, we consider an entailment from (a smaller version of) the Galen ontology[1] as a running example: The entailment $\mathit{Bursa} \sqsubseteq \mathit{HollowBodyStructure}$ ("a bursa is a hollow body structure") can be derived from the following axioms.

$$\begin{aligned}
\mathrm{Bursa} &\sqsubseteq \mathrm{GenericInternalStructure}\\
\mathrm{GenericInternalStructure} &\sqsubseteq \mathrm{GenericBodyStructure}\\
\mathrm{GenericBodyStructure} &\sqsubseteq \mathrm{BodyStructure}\\
\mathrm{Bursa} &\sqsubseteq \exists \mathrm{hasTopology}.(\mathrm{Topology} \sqcap \exists \mathrm{hasState}.\mathrm{hollow})\\
\mathrm{HollowBodyStructure} &\equiv (\mathrm{BodyStructure} \sqcap \exists \mathrm{hasTopology}.(\mathrm{Topology} \sqcap \exists \mathrm{hasState}.\mathrm{hollow})
\end{aligned}$$

To establish how the entailment follows from the axioms, a proof tree is constructed (concept names are abbreviated, e.g. "GIS" for "GenericInternalStructure"):

[1] `http://www.cs.man.ac.uk/~horrocks/OWL/Ontologies/galen.owl`

$$\dfrac{\dfrac{\dfrac{\mathrm{B} \sqsubseteq \mathrm{GIS} \quad \mathrm{GIS} \sqsubseteq \mathrm{GBS} \quad \mathrm{GBS} \sqsubseteq \mathrm{BS}}{\mathrm{B} \sqsubseteq \mathrm{BS}} \quad \mathrm{B} \sqsubseteq \exists \mathrm{hT}.(\mathrm{T} \sqcap \exists \mathrm{hS}.\mathrm{hollow})}{\mathrm{B} \sqsubseteq (\mathrm{BS} \sqcap \exists \mathrm{hT}.(\mathrm{T} \sqcap \exists \mathrm{hS}.\mathrm{hollow})) \quad \mathrm{HBS} \equiv (\mathrm{BS} \sqcap \exists \mathrm{hT}.(\mathrm{T} \sqcap \exists \mathrm{hS}.\mathrm{hollow}))}}{\mathrm{B} \sqsubseteq \mathrm{HBS}}$$

This derivation is then translated to text (numbers are inserted for reference):

"(1) Since a bursa is a generic internal structure, which is a generic body structure, which is a body structure, a bursa is a body structure.
(2) Furthermore, since a bursa is something that has a topology that has a hollow state, a bursa is a body structure that has a topology that has a hollow state.
(3) A hollow body structure is a body structure that has a topology that has a hollow state.
(4) Thus, a bursa is a hollow body structure."

In our previous work [14], we experimentally obtain first indications of the understandability of explanations generated from derivations up to a length of seven inference steps. Furthermore, techniques to shorten the generated explanations are discussed and evaluated (e.g. omitting inference steps considered "trivial" from the explanations and introducing "shortcut" inference rules). In this work, we consider a further extension to the inference rule set by an additional "shortcut" rule. The new rule $\boldsymbol{R}_{\sqsubseteq *}$ (which is used in step (1) of the above example) represents an n-fold application of $\boldsymbol{R}_{\sqsubseteq}$:

$$\boldsymbol{R}_{\sqsubseteq *}\ \frac{(P_1)\ C_1 \sqsubseteq C_2 \quad \ldots \quad (P_{n+1})\ C_{n+1} \sqsubseteq C_{n+2}}{(C)\ C_1 \sqsubseteq C_{n+2}} \qquad \boldsymbol{R}_{\sqsubseteq}\ \frac{(P_1)\ C_1 \sqsubseteq C_2 \qquad (P_2)\ C_2 \sqsubseteq C_3}{(C)\ C_1 \sqsubseteq C_3}$$

The verbalization pattern for $\boldsymbol{R}_{\sqsubseteq *}$ is: "Since $verb(P_1)$, which is $verb(C_3)$, ... , which is $verb(C_{n+2})$, $verb(C)$.", where $verb()$ represents the application of verbalization patterns to basic OWL formulae (cf. [14]).

2.2 Cognitive Complexity

To measure the cognitive complexity of OWL inferences, several models have been proposed. A first step in understanding why an entailment holds consists of finding minimal sets of axioms from which a consequence can be derived, so-called *justifications* [3]. Horridge et al. [3] describe how the cognitive difficulty of a justification can be determined. Their measure is based on twelve dimensions, which were established through an exploratory study and the authors' intuitions. The resulting complexity score takes into account the structure and semantics of a justification and its entailed conclusion. Justifications are classed as hard if they exceed a threshold value, otherwise they are considered as easy.

Nguyen et al. [11] tested single description logic inference rules and established a so-called *facility index* for each rule. The facility indices were obtained in a study where the subjects had to judge whether a given inference is correct or not. The facility index for a certain rule represents the ratio of correct answers to the total number of answers. The model of Nguyen et al. [12] assumes that the difficulties of rule applications are multiplicative. This hypothesis was tested

(and to some degree confirmed) using derivations made up of two inference steps. However, longer derivations were not tested.

Whereas the above mentioned approaches deal with the cognitive complexity of inference rules and justifications, the readability of verbalizations as such, and the modeling of the reading process by a human reader, are not taken into consideration. Our work thus provides an additional perspective by addressing the text comprehension aspect.

3 Modeling

We first present a brief summary of the text comprehension model by Kintsch and van Dijk, focusing only on those parts of the theory relevant for our work. Then we apply this approach to generated verbalizations.

3.1 Theory

The text comprehension model of Kintsch and van Dijk specifies the construction of a semantic representation of a text, called the *text base*. This representation is based on *propositions* and relations between them. The elements of a proposition are defined by Kintsch [6] as *word concepts*, each being a lexical unit in its base form. Propositions are represented as follows.

$$(\text{PREDICATE}, \text{ARGUMENT}_1, ..., \text{ARGUMENT}_n)$$

Predicates are often realized on the surface structure as verbs, adjectives, adverbs or conjunctions. Arguments are mostly nouns, prepositions and embedded propositions which fulfill different semantic functions such as subject, object or goal. The order in which the predicates appear in the text determines the sequence of the propositions in the text base. So propositions are numbered accordingly.

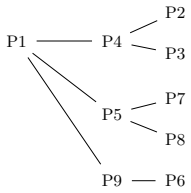
We introduce a further kind of proposition to combine several propositions into one to constitute more complex expressions. The original idea to expand the model by introducing the notion of *facts* came from the authors themselves [6, p. 390]. In this work, facts are defined as an n-tuple of propositions.

$$(\text{PROPOSITION}_1, ..., \text{PROPOSITION}_n)$$

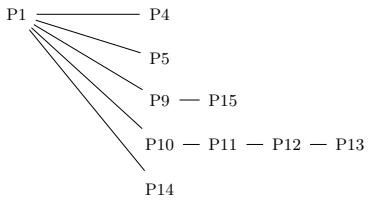
The text processing model assumes a text base to be *coherent*. That is, each proposition must have at least one referentially cohesive relationship to another preceding proposition. If the text base is not coherent, then the model cannot be used. Referential coherence is established by the overlap of arguments between two statements. For instance, a proposition (P1, A1, A2) is defined as referentially coherent with (P2, A2, A3) due to sharing the argument A2. If a proposition is embedded in another, e.g. (P3, A4, (P1, A1, A2)), these two are also considered referentially coherent. This establishes the coherent tree structure of all propositions of a cohesive text base connected in a so-called *coherence graph*.

No.	Proposition/Fact
P1	(IS, BURSA, P4)
P2	(GENERIC, STRUCTURE)
P3	(INTERNAL, STRUCTURE)
P4	(P2, P3)
P5	(IS, P4, P7)
P6	(BODY, STRUCTURE)
P7	(P2, P6)
P8	(IS, P7, P6)
P9	(IS, BURSA, P6)
P10	(IS, BURSA, P11)
P11	(THAT HAS, SOMETHING, P12)
P12	(THAT HAS, TOPOLOGY, P13)
P13	(HOLLOW, STATE)
P14	(IS, BURSA, P15)
P15	(THAT HAS, P6, P12)

(a) Propositions and facts



(b) Coherence graph for cycle 1



(c) Coherence graph for cycle 2

Fig. 1: Proposition list and coherence graphs for the first two sentences in the example

When processing a text, Kintsch and van Dijk assume that the propositions are processed sentence by sentence, in so-called *cycles*. This stepwise processing serves to model the capacity limitation of human short-term memory which is a part of working memory. While reading a sentence, n propositions are processed in working memory of which only s propositions can be stored in the short-term memory *buffer* to be carried over to the next processing cycle. The capacity of the short-term memory depends on the individual characteristics of the reader. In the following, we restrict the capacity to four propositions (cf. [7]). In each cycle, all n propositions of the current sentence are processed by connecting them to referentially coherent propositions that were stored in the buffer during the prior cycle. If no referential coherence is found, the search for a connection includes all previously processed propositions. This procedure is called *long-term memory search*. Each cycle creates a sub-graph, all of which are finally combined into a cohesion graph of the complete text base.

We illustrate this process in Fig. 1 using the first two sentences of the running example (from Sect. 2.1). Initially, the first proposition for initiating the construction of the coherence graph must be selected. The first proposition of the list is selected if it is not embedded in any directly following fact. Otherwise, the latter fact is selected. In our example, P1 is taken as the root and the propositions P2 to P9 are incorporated into the coherence graph. This process begins at the first level of the graph by connecting all propositions having a referential coherence to P1. This applies to P9, P4 and P5 since P1 and P9 share the common argument BURSA, P4 is embedded in P1 and P1 shares with P5 the

proposition P4 as an argument. These propositions form the second level. Now each proposition in this level is checked (in ascending order of their number) for a connection to the remaining propositions. Once all current propositions have been connected into the coherence graph by repeating this procedure, four of them are stored in the short-term memory buffer using the so-called *leading-edge strategy* [8]. This strategy models the aspects of *frequency* and *recency* in human short term memory. Starting from the subgraph of the first cycle, all propositions from Fig. 1 (b) along the path with the highest number including the top proposition are selected, as long as each number is higher than the previous one. Next, the propositions are selected level by level in ascending order of their number starting with the highest level possible. If the storage capacity is reached meanwhile or all available propositions are stored, the process terminates. The selected propositions in the first cycle are P1, P9, P5 and P4. These connected propositions form the initial coherence graph for the next cycle where the propositions P10 to P15 are included in the graph.

3.2 Application to $\mathcal{EL}$ Explanations

For the applicability of Kintsch and van Dijk's model, explanations with a coherent text base are assumed. The model does not specify whether texts without a coherent text base are understandable, (much) more difficult, or not at all understandable. In this work, we hypothesize that explanations without a coherent text base are more difficult to understand than explanations with a coherent text base. The reason is that humans have to resort to their long-term memory to connect the current sentence to the previous text to form a coherence graph. The process of long-term memory search is further described as resource-consuming. As a possible measure of complexity (in the case of texts with a coherent text base), it is therefore appropriate to determine whether, and how often, long-term memory search takes place to obtain an estimate of the cognitive difficulty of explanations. Based on these assumptions, the cognitive complexity of explanations can be divided into three levels.

Complexity level 1: explanations without a coherent text base
(the most difficult to understand)
Complexity level 2: explanations with a coherent text base but also with long-term memory search (difficulty depends on the number of instances of long-term memory search)
Complexity level 3: explanations with a coherent text base and without long-term memory search (easiest to understand)

As a result, explanations without a coherent text base should be restructured in such a way that they have a coherent text base and require no long-term memory search.

The input for the model of Kintsch and van Dijk is a list of propositions that are generated from the explanations produced by the verbalizer. These explanations are not unconstrained texts, for which building an all-encompassing

translation to propositions would be laborious. Rather, the explanations follow a fixed set verbalization patterns based on the available inference rules and the structure of OWL formulae (cf. [14]), so only these text patterns need to be taken into account (instantiated with concept and role names). Fillwords in these patterns such as "since" or "thus" as well as the phrase "according to its definition" which improve reading fluency, are ignored. During verbalization, labels for concept and role names are used where available in an ontology, otherwise camel-cased names are simply split into words.

Predicates. Predicates are mostly determined by the respective logical constructor. This results in the following representation.

$$P_{\sqsubseteq} = (\text{IS, ARGUMENT}_1, \text{ARGUMENT}_2)$$
$$P_{\sqcap} = (\text{AND, ARGUMENT}_1, ..., \text{ARGUMENT}_n)$$

$P_{\exists}$ is formed differently since the predicate is determined from the role name of the existential restriction (usually a verb). For better readability of the propositions "THAT" is prefixed in the predicate. Depending on which word types are included in the role name, the predicate may have two or three arguments.

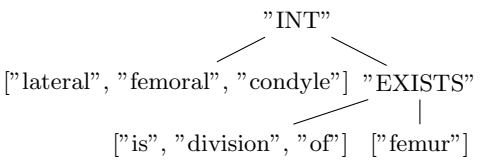
$$P_{\exists} = \ (\text{THAT} + \textit{verb}, \text{ARGUMENT}_1, ..., \text{ARGUMENT}_n) \text{ for } n \ \leq 3$$

Arguments. For $P_{\sqsubseteq}$ and $P_{\sqcap}$, the children of the corresponding constructor determine the arguments of the respective proposition. If the children of a $\sqcap$-node are a concept name and an $\exists$-node, proposition generation is delegated to the $\exists$-node, and the concept name is passed on to become ARGUMENT_1 in the proposition generated for the $\exists$-node (as in P12 in Fig. 1, generated from *Topology* $\sqcap \exists$*hasState.hollow*. If an $\exists$-node has no parent or a $\sqsubseteq$-node as a parent, the word concept SOMETHING is used for ARGUMENT_1 (cf. P11 in Fig. 1).
Facts. The semantic representation of some concept and role names requires the construction of several propositions. This is the case when class and role names consist of several words (e.g. *GenericBodyStructure*). To construct (possibly nested) arguments from these, the nouns, verbs, adjectives, etc. contained in such composite names need to be distinguished. For this purpose, WordNet[2] is employed together with an additional list of prepositions and conjunctions. The more complex expression is obtained by combining the references to the corresponding propositions.

$$P_{FACT} = \ (\text{PROPOSITION}_1, ..., \text{PROPOSITION}_n)$$

For better illustration consider the logical expression *LateralFemoralCondyle* $\sqcap$ $\exists$*isDivisionOf.Femur*. A tree structure (Fig. 2 (a)) is created from the initial logical structure, with nodes representing constructors and leaves representing concept or role names. The propositions (Fig. 2 (b)) are formed by going through this structure recursively. As described above, proposition generation at the first node ($\sqcap$-node) is skipped and begins at the $\exists$-node where the left-hand leaf of

[2] `https://wordnet.princeton.edu/`



(a) Tree structure

$P1 =$ (LATERAL, CONDYLE)
$P2 =$ (FEMORAL, CONDYLE)
$P3 =$ (P1, P2)
$P4 =$ (THAT IS, P3, P5)
$P5 =$ (OF, DIVISION, FEMUR)

(b) Proposition list

Fig. 2: Tree structure and generated proposition list

the ∃-node's parent node is processed first. Since the array consists of two adjectives and one noun at the end (as determined by WordNet), P1 and P2 are constructed. In line with [7, 16], attributes belonging to the same concept are represented as separate individual propositions (e.g. P1 and P2) of the same rank. To join these two propositions, the P_{FACT} proposition P3 is generated, which becomes ARGUMENT_1 of the ∃-node. The propositions for ARGUMENT_2 are determined by part-of-speech analysis of the ∃-node's children (role name and filler), with IS (the verb) becoming part of P4's predicate, and OF (a preposition) relating the noun DIVISION with FEMUR (P5).

Implementing the described cyclic text processing model helped us to check if the explanations generated by the verbalizer have a coherent text base. We discovered situations (labeled 1 and 3) where this was not the case:

$$1:\ \frac{\dfrac{\vdots}{A \sqsubseteq B} \quad B \sqsubseteq C \quad \dfrac{\vdots}{C \sqsubseteq D}}{A \sqsubseteq D} \qquad 2:\ \frac{\dfrac{\dfrac{\vdots}{A \sqsubseteq B} \quad B \sqsubseteq C}{A \sqsubseteq C} \quad \dfrac{\vdots}{C \sqsubseteq D}}{A \sqsubseteq D} \qquad 3:\ \frac{A \sqsubseteq C \quad \dfrac{\vdots}{C \sqsubseteq D}}{A \sqsubseteq D}$$

When verbalizing an inference step that includes a premise that is provided as an axiom (e.g. $B \sqsubseteq C$ in situation 1 and $A \sqsubseteq C$ in situation 3), and a later premise that itself requires some derivation ($C \sqsubseteq D$), our procedure explains how $C \sqsubseteq D$ is derived before stating that the premises together yield the conclusion. However, when explaining the derivation of $C \sqsubseteq D$, C and D appear in the explanation without necessarily being mentioned in the previous text, in which case they cannot serve to establish referential coherence. In case of situation 1, a re-structuring (situation 2) of the derivation (by introducing an auxiliary inference step) solves this problem (since the statement that $A \sqsubseteq C$ is derived provides referential coherence for explaining the derivation of $C \sqsubseteq D$). In case of situation 3, this re-structuring is not possible. Here, the solution is to mention $A \sqsubseteq C$ before explaining $C \sqsubseteq D$ (with a shared C for referential coherence), before stating that together they yield the conclusion. While the above illustrations show the inference rules for transitivity of $\sqsubseteq$, the above observations also apply to some other inference rules used by the verbalizer with at least two premises, e.g. situation 3 also applies to:

$$\frac{A \sqsubseteq \exists r.C \quad \dfrac{\vdots}{C \sqsubseteq D}}{A \sqsubseteq \exists r.D}$$

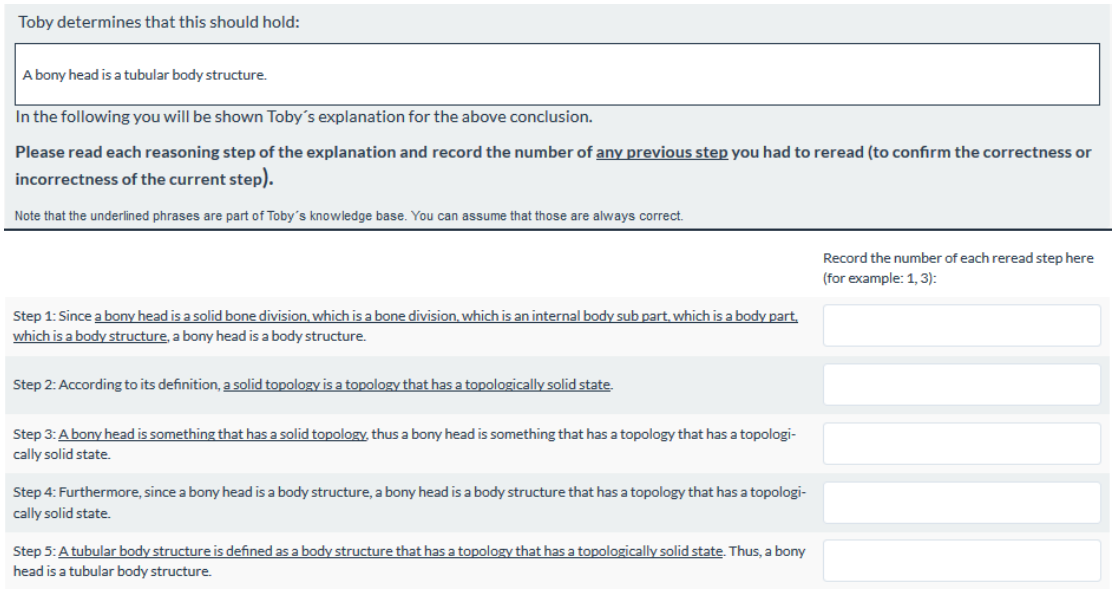


Fig. 3: Explanation shown in the study (with incoherence-inducing step 2)

4 Experimental Study

An online study was carried out to assess how a coherent text base affects the understandability of generated explanations. We compared generated explanations without coherent text base with a corresponding explanation for which a coherent text base was established. To find candidate explanations among those that can be generated using the "verbalizer" tool, we employed our implementation of the text understanding model described above. Since we are not interested in "toy example" ontologies, we selected four explanations from the (aforementioned version of the) Galen ontology that were not too long (5–7 inference steps) and with an incoherent text base (cf. e.g. Fig. 3).

As an objective measure for participants' understanding, the explanations were manipulated to contain errors. Participants were asked to indicate if "the presented reasoning is logically correct (i.e. each step is a consequence of the available knowledge)". Thus, for each explanation, four different versions were created (with/without error and with/without coherent text base). Unfortunately, a mistake was made during experiment preparation that affected one of the presented explanations. The results therefore only refer to three explanations (named E1, E2 and E3 in the following).

Participants. English-speaking participants were recruited through advertisements at Ulm University, social media and personal contact, with a raffle for 25€ gift vouchers as an incentive. Eighteen participants (15 males, three females)

Table 1: Variants of explanations E1-E4 presented to the four experimental groups. V1: incoherent/correct, V2: incoherent/incorrect, V3: coherent/correct, V4: coherent/incorrect

Group 1	Group 2	Group 3	Group 4
E1, V4	E1, V2	E1, V3	E1, V1
E2, V1	E2, V3	E2, V2	E2, V4
E3, V3	E3, V1	E3, V4	E3, V2
E4, V2	E4, V4	E4, V1	E4, V3

Table 2: Participants' classification of explanations

	No. of responses			
Explanation	correct	incorrect	excluded	total
E1 incoher.	3	5	1	9
E1 coherent	5	2	1	8
E2 incoher.	6	2	0	8
E2 coherent	5	3	1	9
E3 incoher.	6	2	1	9
E3 coherent	6	2	0	8

completed the study (among 42 who started the survey). One participant was excluded due to suspected "straight-lining" (cf. [4]), thus 17 remained.

Procedure. Participants were split into experimental groups according to Table 1. After a short introduction to the study's requirements, participants were introduced to the task using an example explanation. They were informed that the explanations to be judged are generated from a knowledge database, and that they may contain errors. After completing a pre-test questionnaire, each participant was shown four explanations (e.g. Fig. 3). They were asked to read the explanations step by step. For each of them, the participants should indicate whether the explanation is correct, and they had to provide subjective ratings concerning understandability and the adequacy of the order in which the reasoning steps are presented. Further input fields were provided for the participants to indicate which steps they considered to be erroneous or hard to understand.

Results. The classification accuracy is shown in Table 2 for the three different explanations E1–3 in their original version (incoherent text base) and their improved version with coherent text base. As mentioned, respondents were asked in which step they suspected the error. If this indication did not match the "actual" error, the response was excluded (cf. Table 2).

Regarding the question whether the coherence of the text base may have affected the classification accuracy, only the data for the explanation E1 hint at a potential effect (slightly better classification performance when the text base is coherent). A binomial test yields $p = 0.074$, and a Chi-Square goodness-of-fit test yields $\text{Chi}^2(1)=4.5125$, $p < 0.05$, though both these indications should be taken with great caution due to small sample size. When considering all responses for explanations E1-3, the experiment did not yield significant improvements in classification accuracy for a coherent vs. an incoherent text base.

Figure 4 shows the mean scores of participants' answers (for correctly classified explanations only) to our questions regarding readability. Participants had a neutral to slightly positive tendency with regards to question Q1 "[The] explanation is easy to understand", both with and without coherent text base. Responses to Q2 "The order in which the reasoning steps are presented is appropriate" and Q3 "The order in which the reasoning steps are presented should be changed" indicated that participants mostly agreed with the presented order. They had a tendency to agree to Q4 "Each step by itself is understandable", though answers

were mixed. Similarly, answers to Q5 "Some sentences are difficult to read" were mixed. Overall, the answers suggest that participants in most cases did not consider the explanations hard to understand. However, coherence of the text base did not lead to a more positive assessment by the participants.

5 Conclusions & Discussion

The application of the text understanding model by Kintsch and van Dijk revealed that some explanations generated by our verbalization tool lacked a coherent text base (already in the inexpressive $\mathcal{EL}$ fragment of DL). This allowed us to take the issue into account and to explore in a first study whether this helped to improve understandability. Our first results are inconclusive in this respect. Improving the coherence of the text base did not yield a clear effect: Classification accuracy had a tendency to improve for one of the explanations employed as material, but remained the same for two others. Understandability as reported by the participants was also not found to improve.

Several factors may have played a role: the overall difficulty of the task (as evidenced by the number of participants who quit) and the resulting small sample size, the set of "naturalistic" explanations taken from the Galen ontology, and other inadequacies of the generated explanations (length, repetitiveness, ambiguities) that might in addition to incoherence affect readability. Furthermore, most participants were not English native speakers. Nevertheless, the reported observations are useful for setting up further, more targeted studies with a larger number of participants. The examination of other predictions made by Kintsch and van Dijk's model, e.g. whether long-term memory search affects the readability of the explanations, remains for future work, as well as applying our methodology to more expressive fragments of DL (since the employed verbalization tool is only gradually extended to more expressive logics than $\mathcal{EL}$).

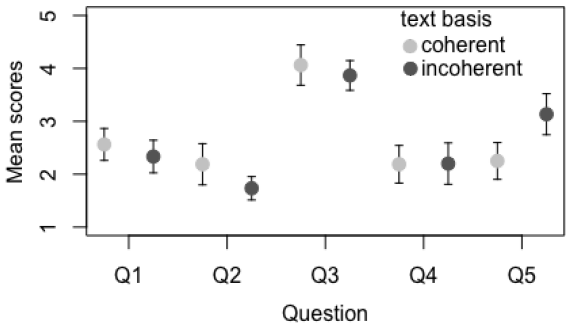


Fig. 4: Mean responses to questions Q1-Q5 for explanations E1-E3 on a 5-point Likert scale (1: agree – 5: disagree), together with standard errors of the mean.

Acknowledgments. We acknowledge the support of the German Research Foundation (DFG) within the project "Live Ontologies" (KA 3470/2-2) and the technology transfer project "Do it yourself, but not alone: Companion Technology for Home Improvement" of the Transregional Collaborative Research Center SFB/TRR 62 with the industrial project partner Robert Bosch GmbH. We thank four reviewers and F. S. for helpful comments 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. Borgida, A., Franconi, E., Horrocks, I.: Explaining $\mathcal{ALC}$ subsumption. In: Horn, W. (ed.) Proceedings of the 14th European Conference on Artificial Intelligence, pp. 209–213. IOS Press (2000)
3. Horridge, M., Bail, S., Parsia, B., Sattler, U.: The cognitive complexity of OWL justifications. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) The Semantic Web – ISWC 2011. LNCS, vol. 7031, pp. 241–256. Springer (2011)
4. Jones, M.S., House, L.A., Gao, Z.: Respondent screening and revealed preference axioms: Testing quarantining methods for enhanced data quality in web panel surveys. Public Opinion Quarterly 79(3), 687–709 (2015)
5. 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)
6. Kintsch, W.: The representation of meaning in memory. Oxford, England: Lawrence Erlbaum (1974)
7. Kintsch, W., van Dijk, T.A.: Toward a model of text comprehension and production. Psychological review 85(5), 363–394 (1978)
8. Kintsch, W., Vipond, D.: Reading comprehension and readability in educational practice and psychological theory. In: Nilsson, L.G. (ed.) Memory: Processes and problems. Hillsdale, N.J.: Erlbaum (1978)
9. 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(6), 227–239 (2013)
10. McGuinness, D.L.: Explaining Reasoning in Description Logics. Ph.D. thesis, Rutgers University (1996)
11. 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)
12. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Predicting the understandability of OWL inferences. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) The Semantic Web: Semantics and Big Data, pp. 109–123. Springer (2013)
13. Schiller, M.R.G., Glimm, B.: Towards explicative inference for OWL. In: Eiter, T., Glimm, B., Kazakov, Y., Krötzsch, M. (eds.) Proceedings of the 28th International Workshop on Description Logics, CEUR Workshop Proceedings, vol. 1014 (2013)
14. Schiller, M.R.G., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of EL subsumptions. In: Artale, A., Glimm, B., Kontchakov, R. (eds.) Proceedings of the 30th International Workshop on Description Logics (DL), CEUR Workshop Proceedings, vol. 1879 (2017)
15. Stevens, R., Malone, J., Williams, S., Power, R., Third, A.: Automating generation of textual class definitions from OWL to English. Journal of Biomedical Semantics 2(Suppl. 2), S5 (2011)
16. van Dijk, T.A., Kintsch, W.: Strategies of discourse comprehension. New York: Academic Press (1983)