Undefined 0 (2012) 1–0
IOS Press

The Understandability of OWL Statements in Controlled English

Editor(s): Mark Gahegan, The University of Auckland, New Zealand
Solicited review(s): Glen Hart, Ordnance Survey, UK; Beryl Plimmer, The University of Auckland, New Zealand; Paul Smart, University of Southampton, UK

Tobias Kuhn *
Department of Informatics & Institute of Computational Linguistics, University of Zurich, Switzerland
Department of Intelligent Computer Systems, University of Malta, Msida, Malta
`kuhntobias@gmail.com — http://www.tkuhn.ch`

Abstract. Different kinds of controlled natural language (CNL) have been proposed as a front-end for Semantic Web systems, in order to make them more accessible to users with no background in formal notations and methods. This paper investigates whether OWL statements in CNL are indeed easier to understand than in other notations. To this aim, an experiment with 64 participants was conducted that compares a controlled natural language to a classical OWL notation. Concretely, Attempto Controlled English was compared to a simplified version of the Manchester OWL Syntax. For a reliable and tool-independent evaluation of understandability, the experiment is based on a novel evaluation framework making use of simple and intuitive diagrams. The results show that CNL is easier to understand, needs less learning time, and is more accepted by its users.

Keywords: Understandability, User Experiment, Controlled Natural Language, Attempto Controlled English (ACE), Web Ontology Language (OWL)

1. Introduction

Different controlled natural languages (CNL) have been proposed for the area of knowledge representation, and specifically for the Semantic Web, in order to overcome the problem that common formal languages are often hard to understand for people unfamiliar with formal notations [24,22]. It has often been taken for granted that CNL is easier to understand, but so far there is no solid evidence for this. Here, an experiment is presented that was designed and performed to test this claim. Apart from that, the goal is to get measures for the learning time needed and the user acceptance.

This paper is an extended and revised version of a workshop paper [16], with a slight shift of focus: here, the focus is more on the actual experiment and less on the evaluation framework. For more information on the evaluation framework, the paper cited above should be consulted.

In what follows, the background on controlled natural language in the area of the Semantic Web is given in Section 2, related approaches on evaluating CNL are discussed in Section 3, the approach of this work is introduced in Section 4, the design of the experiment is described in Section 5, the results thereof are discussed in Section 6, and finally the conclusions are drawn in Section 7.

*This work was funded by the research grant (Forschungskredit) programs 2006 and 2008 of the University of Zurich. I would like to thank Alain Cohn, Norbert E. Fuchs, Kaarel Kaljurand, Marc Lutz, Cerstin Mahlow, Philipp Müller and Michael Piotrowski for their suggestions and corrections. Furthermore, I would like to acknowledge the Institute for Empirical Research in Economics of the University of Zurich for organizing the recruitment of the participants for the experiment.

0000-0000/12/$00.00 © 2012 – IOS Press and the authors. All rights reserved

2. Background

Controlled natural languages [29,20] are artificially defined subsets of natural languages with restricted vocabulary, syntax or semantics. The goal is to create languages that look natural but without the ambiguity and complexity of full natural language. Some CNLs (those that are relevant for this paper) are completely formal and can be automatically mapped to logic. Examples of such languages include ACE [7], CELT [19], CLCE [26], CLP [4], Formalized-English [18] and PENG [25].

Several controlled natural languages have been specifically designed for the ontology language OWL to be used in the context of the Semantic Web. Rabbit [11] is one of them, designed to enhance communication between domain experts and ontology engineers in order to create ontologies for specific domains. The Sydney OWL Syntax [5] is another example, based on the language PENG. It provides a bidirectional mapping to OWL, so that statements in the Sydney OWL Syntax can be translated into OWL expressions and vice versa. CLOnE (Controlled Language for Ontology Editing) [27,9] is a very simple language defined by only eleven sentence patterns which roughly correspond to eleven OWL axiom patterns. Due to its simple design, only a small subset of OWL is covered. Lite Natural Language [1] is a fourth example, which maps to DL-Lite, a logical formalism optimized for good computational properties and which is equivalent to a subset of OWL. Metalog Pseudo Natural Language [17] is yet another example, building upon RDF and Prolog. The language Attempto Controlled English (ACE) has a more general focus and is more expressive than OWL, but a mapping to OWL has been defined for a subset of it [13]. Schwitter et al. [22] show a comparison of three controlled languages in the area of the Semantic Web (Sydney OWL Syntax, Rabbit, and ACE).

All these languages are designed to make the interaction with formal ontological statements (e.g. in OWL) easier and faster for users unfamiliar with formal methods and notations. It is not as obvious as it seems at first sight, however, that CNLs are indeed easier to understand than other notations. As a concrete example, is "every article is a publication" easier to understand than "article SubTypeOf publication"? How can we be sure about this? Rigorous user studies are required to test such claims.

3. Related Work

Existing approaches to evaluate CNLs can be subdivided into two categories: task-based and paraphrase-based approaches.

In task-based experiments, as presented by Bernstein and Kaufmann [2] and Funk et al. [9,8], participants receive tasks to add certain knowledge to the knowledge base using a tool that is based on a CNL. Such experiments test the ability of users to write CNL statements or test a certain system as a whole. The capacity of users to write statements, however, does not directly imply that they understand them. Another problem is that it is hard to determine how much the CNL contributes to the general usability and understandability, and how much is due to other aspects of the tool. It is also hard to compare CNLs to other formal languages, because different languages often require different tools.

Paraphrase-based approaches are a way how CNLs can be tested in a tool-independent manner. In contrast to task-based approaches, they aim at evaluating the comprehensibility of a CNL rather than the usability of tools based on CNL. Examples of such experiments are described by Hart et al. [11] and Hallett et al. [10]. In paraphrase-based experiments, participants are given CNL statements and they have to choose from a number of paraphrases in natural language, only one of which is correct. Such experiments have the advantage that they do not depend on a particular tool, but there are other problems. Since natural language is inherently vague and ambiguous, it has to be ensured somehow that the participants understand the natural language paraphrases in the way they are intended, which just takes the same problem to the next level. It is hard to determine whether understanding is necessary to fulfill the task. The participants might do the right thing without understanding the sentences (e.g. just by following some syntactic patterns), or by misunderstanding both — statement and paraphrase — in the same way.

Thus, it is difficult to get reliable results concerning understandability of languages with either of the existing approaches.

4. Approach

In order to overcome the problems of existing approaches identified above, a novel, diagram-based approach is applied to test and compare the understand-

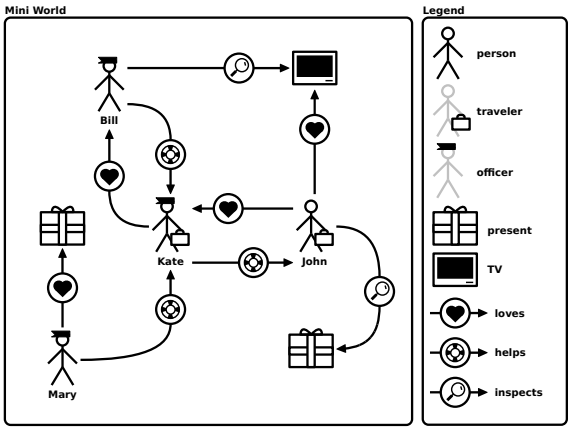

Fig. 1. This is an example of an ontograph. The legend to the right defines the types and relations. The mini world to the left shows the actual individuals with their actual types and relation.

ability of different languages. This approach is based on the ontograph framework [16,15], relying on a very simple and intuitive graphical notation called *ontographs*. The basic idea is to describe simple situations in graphical diagrams, so that these situation descriptions can be used in human subject experiments as a common basis to test the understandability of different formal languages.

Figure 1 shows an example of an ontograph diagram. It consists of a legend that introduces types and relations and of a mini world that describes the actual individuals, their types, and their relations. The legend shown to the right introduces types (e.g. "officer" and "present") and relations (e.g. "helps") by linking their name to their graphical representation. The mini world shown to the left describes actual situations consisting of individuals with their types and relations.

The ontograph notation has some important characteristic properties. First of all, it does not allow for expressing incomplete knowledge. Nothing can be left unspecified: every statement about the mini world is either necessarily true or necessarily false. Another important property is the lack of generalization capabilities: there is no support for any kind of quantification. Every individual and every relation instance has to be represented individually.

Ontograph diagrams can be used in experiments as a neutral and simple language for comparison. In order to test the understandability of a language, an ontograph and several statements (written in the language to be tested) can be shown to the participants of an experiment, who have to decide which of the statements are true and which are false with respect to the situation depicted by the ontograph.

In contrast to other approaches, good performance of the participants of an experiment implies understanding, at least in a certain model-theoretic sense of the word. Ontographs can be considered a language to describe first-order models: the individuals shown in the mini world represent the domain elements; their icons represent the interpretation of the unary predicates; and the arrows represent the interpretation of the binary predicates. The statements shown to the participants of an experiment correspond to simple first-order theories (or theories in any other kind of logic based on model theory). From this point of view, the task of the participants is to decide whether or not certain theories have the shown ontograph as a model. Now, all semantic properties of first-order theories (consistency, entailment, equivalence, etc.) are solely defined on their models, and these definitions are very simple when the mapping between theories and models is taken for granted. For example, two statements are equivalent if and only if they have exactly the same models. It is thus sensible to say that an agent — computer program or human — understands a certain logic language if the agent is able to correctly map theories to models.[1] So, participants performing well in ontograph experiments (systematically, not by mere luck) prove that they understand the core aspect of the language, i.e. the mapping to models.

5. Experiment Design

The experiment to be presented compares the understandability of OWL statements in a common notation with their representations in a controlled natural language. The experiment was performed on 64 participants.

5.1. Languages

The most important design decision for the experiment is the choice of the languages to be compared against each other. On the CNL side, ACE has been chosen because it is a very mature CNL and has a clearly defined mapping to OWL. The Manchester OWL Syntax, a usability-oriented syntax of OWL, has been chosen as the second language. It "was developed

[1] Based on the same idea, Bos [3] suggests to use "Textual Model Checking" as a way to evaluate NLP systems.

in response to a demand from a wide range of users, who do not have a Description Logic background, for a 'less logician like' syntax" and its inventors claim that "it is quick and easy to read and write" [12]. Thus, it seems to be an appropriate language for such an understandability experiment.

However, the Manchester OWL Syntax requires the statements to be grouped by their main ontological entity (the one in subject position). This is a reasonable approach for the definition of complete ontologies, but it makes it impossible to state short and independent statements that could be used for a direct comparison to ACE in an experimental setting. For this reason, a simplified version of the Manchester OWL Syntax has been defined specifically for this experiment. The resulting language, to be called "Manchester-like language" or "MLL", uses the same or very similar keywords but enables short and independent statements.

Table 1 shows the simple grammar of MLL used in the experiment. Furthermore, the table shows how four different subsets are defined, each of which was tested independently in a separate series (see below).

MLL adopts the color codes of the Manchester OWL Syntax for improved readability. Turquoise is used for the boolean operators on types and for the inverse operator on relations. More complex operators used for type restrictions involving relations are displayed in magenta.

Table 1

This table shows the grammar rules of MLL (in Backus-Naur style) and the subsets thereof used for the individual series. Each series uses exactly seven out of the 22 grammar rules.

		Series 1	2	3	4
S ::=	I **HasType** T	×	×	×	
	I R I		×		
	I **not** R I		×		
	T **SubTypeOf** T	×	×	×	
	R **SubRelationOf** R				×
	T **EquivalentTo** T	×			
	R **EquivalentTo** R				×
	T **DisjointWith** T	×			
	R **DisjointWith** R				×
	R **HasDomain** T			×	
	R **HasRange** T			×	
	R **IsSymmetric**				×
	R **IsAsymmetric**				×
	R **IsTransitive**				×
T ::=	**not** T	×	×		
	T **or** T	×		×	
	T **and** T	×			
	R **some** T		×		
	R **only** T		×		
	R **min** N T			×	
	R **max** N T			×	
R ::=	**inverse** R				×

5.2. Learning Time

Obviously, the understanding of a language highly depends on the amount of time spent for learning the language. This means that one has to define a certain time frame when evaluating the understandability of languages. Some languages might be the best choice if there is only little learning time; other languages might be less understandable in this situation but are more suitable in the long run.

So far, little is known about how the understandability of CNLs compares to the understandability of common formal languages. CNLs are typically designed to be understandable with no learning at all. Since this does not hold for other formal languages like the Manchester OWL Syntax, it would not be appropriate to compare ACE to such a language in a zero learning time scenario.

For this reason, learning times of about 20 minutes have been chosen. This seems to be a reasonable first step away from the zero learning time scenario. The effect of longer learning times remains open to be studied in the future.

5.3. Ontographs and Statements

Four series of ontographs were created that cover certain types of statements. The first series only contains individuals and types, but no relations. The following example shows such a statement in ACE and its equivalent in MLL:

ACE: Mary is an officer or is a golfer.
MLL: Mary **HasType** officer **or** golfer

The statements of the second series contain relations with different kinds of simple universal quantifications, for example:

ACE: Every woman buys a picture.
MLL: woman **SubTypeOf** buys **some** picture

The third series contains domain, range, and number restrictions:

ACE: Every officer buys at least 2 presents.

MLL: officer **SubTypeOf** buys **min** 2 present

The fourth series consists in essence only of relations:

ACE: If X helps Y then Y admires X.
MLL: helps **SubRelationOf** **inverse** admires

For each of the four series, three ontographs were created. For each ontograph, twenty statement pairs were defined in a way that each pair consists of an ACE statement and a semantically equivalent MLL statement. Some of the statement pairs are true with respect to their ontograph and the others are false.

The concrete statements were built according to forty statement patterns, covering a very large part of the expressiveness of OWL. Table 2 shows these statement patterns and how they are distributed to the four series. Each pattern has an identifier like 1/2, where the first number denotes the series and the second one numbers patterns within each series. The concrete statements together with their ontograph diagrams are available in the author's doctoral thesis [15] and online[2].

5.4. Participants

Another important design decision is the choice of the participants. For this experiment, students of different fields of study have been chosen as participants. Apart from the fact that students are usually flexible and close to the research facilities of the university, there are more reasons why students are a good choice in this particular case. Students are used to think systematically and logically but they are usually not familiar with formal logical notations (unless this lies in their field of study). In this way, they resemble domain experts who have to formalize their knowledge and who should profit from languages like ACE.

The requirements for the participants were defined as follows: They had to be students or graduates with no higher education in computer science or logic. Furthermore, at least intermediate level skills in written German and English were required, because the experiment itself was explained and performed in German, and English was needed to understand the ACE sentences.

64 students were recruited who fulfill these requirements. They were on average 22 years old and 42% of them were female and 58% were male. These students exhibited a broad variety of fields of study:[3] 40% studied applied sciences (mostly engineering, business, pharmaceutics and law), 27% natural sciences (mostly biology and physics), 19% social sciences (mostly economy and psychology), 11% humanities (mostly languages), and 3% mathematics (with no focus on logic).

In order to enable a good comparison between the two languages, each participant was tested on ACE and on MLL. However, since participants cannot be expected to concentrate for much longer than one hour, only one of the four ontograph series could be tested per participant.

In order to rule out learning effects, half of the participants received the ACE task first and then the MLL task while the other half received the tasks in the reverse way.

5.5. Procedure

The experiment was conducted in a computer room with a computer for each participant. The main part of the experiment was performed on the computer screen. Additionally, the participants received different printed sheets during the experiment. The overall procedure consisted of six stages:

1. Instructions with control questions
2. Learning phase 1
3. Testing phase 1
4. Learning phase 2
5. Testing phase 2
6. Questionnaire

For the instruction phase, the participants received a printed instruction sheet that explained the experiment procedure, the payout, and the ontograph notation. The reverse side of the instruction sheet contained control questions for the participants to answer, which allowed us to check whether the participants understood the instructions.

For the first learning phase, the participants received a language description sheet of the first language (either ACE or MLL). This sheet only explained the subset of the language that is used for the respective series. For this reason, each series had its own instruction sheets for both languages. During the learning phase, an ontograph was shown on the screen together with ten true statements marked as "true" and ten false state-

[2] http://attempto.ifi.uzh.ch/site/docs/ontograph/

[3] The following numbers are the results of rough calculations, since not all fields of study can be clearly categorized.

Table 2

This table shows the statement patterns for ACE and MLL used for the experiment. They are divided into four series. The placeholders $\langle I \rangle$ stand for individuals (i.e. proper names), $\langle T \rangle$ for types (i.e. nouns), and $\langle R \rangle$ for relations (i.e. verbs)

PATTERN	ACE	MLL
1/1	$\langle I \rangle$ is a $\langle T \rangle$.	$\langle I \rangle$ **HasType** $\langle T \rangle$
1/2	$\langle I \rangle$ is not a $\langle T \rangle$.	$\langle I \rangle$ **HasType** **not** $\langle T \rangle$
1/3	$\langle I \rangle$ is a $\langle T_1 \rangle$ or is a $\langle T_2 \rangle$.	$\langle I \rangle$ **HasType** $\langle T_1 \rangle$ **or** $\langle T_2 \rangle$
1/4	$\langle I \rangle$ is a $\langle T_1 \rangle$ and is a $\langle T_2 \rangle$.	$\langle I \rangle$ **HasType** $\langle T_1 \rangle$ **and** $\langle T_2 \rangle$
1/5	Every $\langle T_1 \rangle$ is a $\langle T_2 \rangle$.	$\langle T_1 \rangle$ **SubTypeOf** $\langle T_2 \rangle$
1/6	No $\langle T_1 \rangle$ is a $\langle T_2 \rangle$.	$\langle T_1 \rangle$ **DisjointWith** $\langle T_2 \rangle$
1/7	Every $\langle T_1 \rangle$ is a $\langle T_2 \rangle$ and every $\langle T_2 \rangle$ is a $\langle T_1 \rangle$.	$\langle T_1 \rangle$ **EquivalentTo** $\langle T_2 \rangle$
1/8	Every $\langle T_1 \rangle$ who is not a $\langle T_2 \rangle$ is a $\langle T_3 \rangle$.	$\langle T_1 \rangle$ **and** (**not** $\langle T_2 \rangle$) **SubTypeOf** $\langle T_3 \rangle$
1/9	Every $\langle T_1 \rangle$ is a $\langle T_2 \rangle$ or is a $\langle T_3 \rangle$.	$\langle T_1 \rangle$ **SubTypeOf** $\langle T_2 \rangle$ **or** $\langle T_3 \rangle$
1/10	Nobody who is a $\langle T_1 \rangle$ or who is a $\langle T_2 \rangle$ is a $\langle T_3 \rangle$ and is a $\langle T_4 \rangle$.	$\langle T_1 \rangle$ **or** $\langle T_2 \rangle$ **SubTypeOf** **not** ($\langle T_3 \rangle$ **and** $\langle T_4 \rangle$)
2/1	$\langle I_1 \rangle$ $\langle R \rangle$ $\langle I_2 \rangle$.	$\langle I_1 \rangle$ $\langle R \rangle$ $\langle I_2 \rangle$
2/2	$\langle I_1 \rangle$ does not $\langle R \rangle$ $\langle I_2 \rangle$.	$\langle I_1 \rangle$ **not** $\langle R \rangle$ $\langle I_2 \rangle$
2/3	$\langle I \rangle$ $\langle R \rangle$ a $\langle T \rangle$.	$\langle I \rangle$ **HasType** $\langle R \rangle$ **some** $\langle T \rangle$
2/4	$\langle I \rangle$ $\langle R \rangle$ no $\langle T \rangle$.	$\langle I \rangle$ **HasType** **not** ($\langle R \rangle$ **some** $\langle T \rangle$)
2/5	$\langle I \rangle$ $\langle R \rangle$ something that is not a $\langle T \rangle$.	$\langle I \rangle$ **HasType** $\langle R \rangle$ **some** (**not** $\langle T \rangle$)
2/6	$\langle I \rangle$ $\langle R \rangle$ nothing but $\langle T \rangle$.	$\langle I \rangle$ **HasType** $\langle R \rangle$ **only** $\langle T \rangle$
2/7	Every $\langle T_1 \rangle$ $\langle R \rangle$ a $\langle T_2 \rangle$.	$\langle T_1 \rangle$ **SubTypeOf** $\langle R \rangle$ **some** $\langle T_2 \rangle$
2/8	Everything that $\langle R \rangle$ a $\langle T_1 \rangle$ is a $\langle T_2 \rangle$.	$\langle R \rangle$ **some** $\langle T_1 \rangle$ **SubTypeOf** $\langle T_2 \rangle$
2/9	Every $\langle T_1 \rangle$ $\langle R \rangle$ nothing but $\langle T_2 \rangle$.	$\langle T_1 \rangle$ **SubTypeOf** $\langle R \rangle$ **only** $\langle T_2 \rangle$
2/10	Everything that $\langle R \rangle$ nothing but $\langle T_1 \rangle$ is a $\langle T_2 \rangle$.	$\langle R \rangle$ **only** $\langle T_1 \rangle$ **SubTypeOf** $\langle T_2 \rangle$
3/1	Everything that $\langle R \rangle$ something is a $\langle T \rangle$.	$\langle R \rangle$ **HasDomain** $\langle T \rangle$
3/2	Everything that is $\langle R \rangle$ by something is a $\langle T \rangle$.	$\langle R \rangle$ **HasRange** $\langle T \rangle$
3/3	Everything that $\langle R \rangle$ something is a $\langle T_1 \rangle$ or is a $\langle T_2 \rangle$.	$\langle R \rangle$ **HasDomain** $\langle T_1 \rangle$ **or** $\langle T_2 \rangle$
3/4	Everything that is $\langle R \rangle$ by something is a $\langle T_1 \rangle$ or is a $\langle T_2 \rangle$.	$\langle R \rangle$ **HasRange** $\langle T_1 \rangle$ **or** $\langle T_2 \rangle$
3/5	$\langle I \rangle$ $\langle R \rangle$ at least 2 $\langle T \rangle$.	$\langle I \rangle$ **HasType** $\langle R \rangle$ **min** 2 $\langle T \rangle$
3/6	$\langle I \rangle$ $\langle R \rangle$ at most 1 $\langle T \rangle$.	$\langle I \rangle$ **HasType** $\langle R \rangle$ **max** 1 $\langle T \rangle$
3/7	Every $\langle T_1 \rangle$ $\langle R \rangle$ at least 2 $\langle T_2 \rangle$.	$\langle T_1 \rangle$ **SubTypeOf** $\langle R \rangle$ **min** 2 $\langle T_2 \rangle$
3/8	Everything that $\langle R \rangle$ at least 2 $\langle T_1 \rangle$ is a $\langle T_2 \rangle$.	$\langle R \rangle$ **min** 2 $\langle T_1 \rangle$ **SubTypeOf** $\langle T_2 \rangle$
3/9	Every $\langle T_1 \rangle$ $\langle R \rangle$ at most 1 $\langle T_2 \rangle$.	$\langle T_1 \rangle$ **SubTypeOf** $\langle R \rangle$ **max** 1 $\langle T_2 \rangle$
3/10	Everything that is a $\langle T_1 \rangle$ or that is a $\langle T_2 \rangle$ $\langle R \rangle$ at most 1 $\langle T_3 \rangle$.	$\langle T_1 \rangle$ **or** $\langle T_2 \rangle$ **SubTypeOf** $\langle R \rangle$ **max** 1 $\langle T_3 \rangle$
4/1	If X $\langle R \rangle$ Y then Y $\langle R \rangle$ X.	$\langle R \rangle$ **IsSymmetric**
4/2	If X $\langle R \rangle$ Y then Y does not $\langle R \rangle$ X.	$\langle R \rangle$ **IsAsymmetric**
4/3	If X $\langle R \rangle$ somebody who $\langle R \rangle$ Y then X $\langle R \rangle$ Y.	$\langle R \rangle$ **IsTransitive**
4/4	If X $\langle R_1 \rangle$ Y then X $\langle R_2 \rangle$ Y.	$\langle R_1 \rangle$ **SubRelationOf** $\langle R_2 \rangle$
4/5	If X $\langle R_1 \rangle$ Y then X $\langle R_2 \rangle$ Y.	$\langle R_1 \rangle$ **SubRelationOf** $\langle R_2 \rangle$
4/6	If X $\langle R_1 \rangle$ Y then Y $\langle R_2 \rangle$ X.	$\langle R_1 \rangle$ **SubRelationOf** **inverse** $\langle R_2 \rangle$
4/7	If X $\langle R_1 \rangle$ Y then X does not $\langle R_2 \rangle$ Y.	$\langle R_1 \rangle$ **DisjointWith** $\langle R_2 \rangle$
4/8	If X $\langle R_1 \rangle$ Y then Y does not $\langle R_2 \rangle$ X.	$\langle R_1 \rangle$ **DisjointWith** **inverse** $\langle R_2 \rangle$
4/9	If X $\langle R_1 \rangle$ Y then X $\langle R_2 \rangle$ Y. If X $\langle R_2 \rangle$ Y then X $\langle R_1 \rangle$ Y.	$\langle R_1 \rangle$ **EquivalentTo** $\langle R_2 \rangle$
4/10	If X $\langle R_1 \rangle$ Y then Y $\langle R_2 \rangle$ X. If Y $\langle R_2 \rangle$ X then X $\langle R_1 \rangle$ Y.	$\langle R_1 \rangle$ **EquivalentTo** **inverse** $\langle R_2 \rangle$

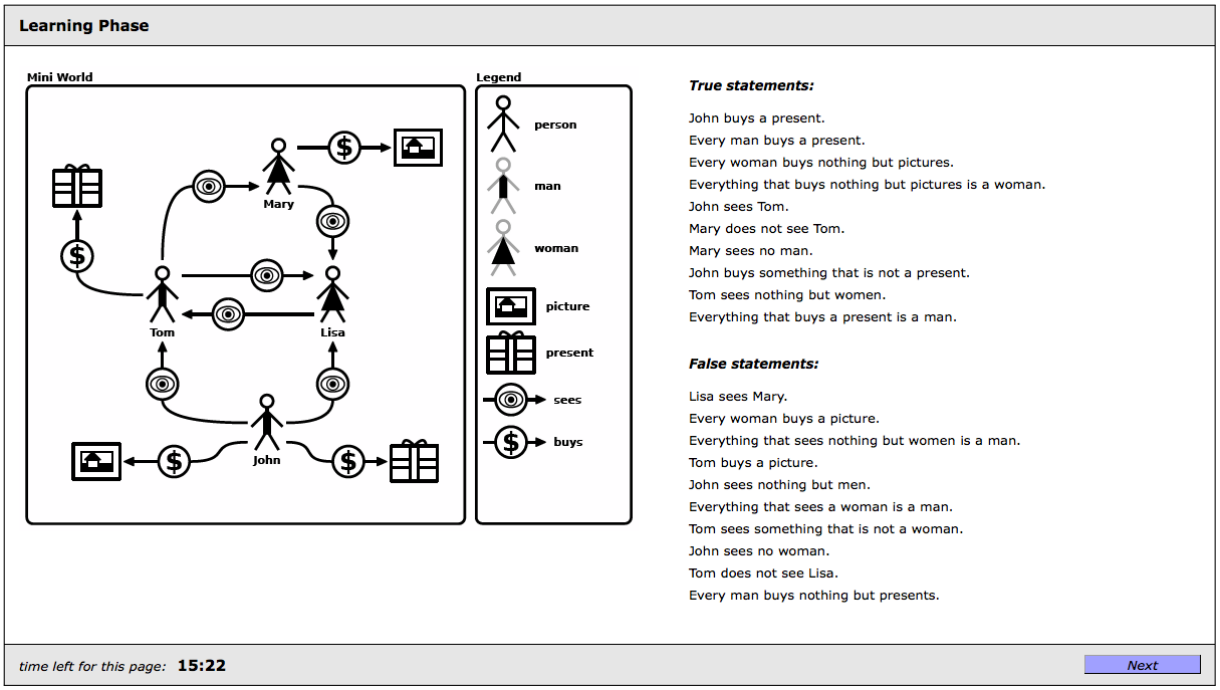

Fig. 2. This is the screen shown during the learning phase of the experiment (for the language ACE).

ments marked as "false" in the respective language. Figure 2 shows a screenshot of the experiment screen during the learning phase.

During the testing phase, a different ontograph was shown on the screen. Furthermore, ten statements in the respective language were shown together with radio buttons that let the participants choose between "true", "false" and "don't know" (they could keep the language description sheet that they got for the learning phase). Figure 3 shows how the screen of the testing phase looked like.

For the steps 4 and 5, the procedure of the steps 2 and 3 was repeated for the second language (i.e. ACE if the first language was MLL and vice versa) with the same ontograph for the learning phase but a new one for the testing.

Finally, the participants received a questionnaire form inquiring about their background and their experiences during the experiment.

The learning phases had a time limit of 16 minutes each, and the time limit for the testing phases was 6 minutes. The participants were forced to proceed when the time limit ran out but they could proceed earlier. In this way, it can not only be investigated how understandable the languages are but also how much time the participants needed to learn them.

5.6. Language Description Sheets

The proper design of the language description sheets is crucial for this experiment. If the participants perform better in one language than in the other, it might be that the respective language was merely described better than the other. Thus, the language description sheets have to be written very carefully to be sure that they are not misunderstood and are optimal for learning the respective language under the given time restrictions. Especially the description sheets for MLL are critical. In contrast to ACE, MLL is not designed to be understood without training. For this reason, a special effort was made to ensure the quality of the MLL description sheets, involving several steps.

First of all, the four series were designed in a way that at most seven MLL keywords are used per series. Since each series has its own language description sheets, not more than seven keywords are described on each sheet.

In a second step, the different MLL description sheets were given to three persons, who comply with the restrictions of the experiment but who did not participate in it. These three persons gave feedback about what they did not understand and what could be improved.

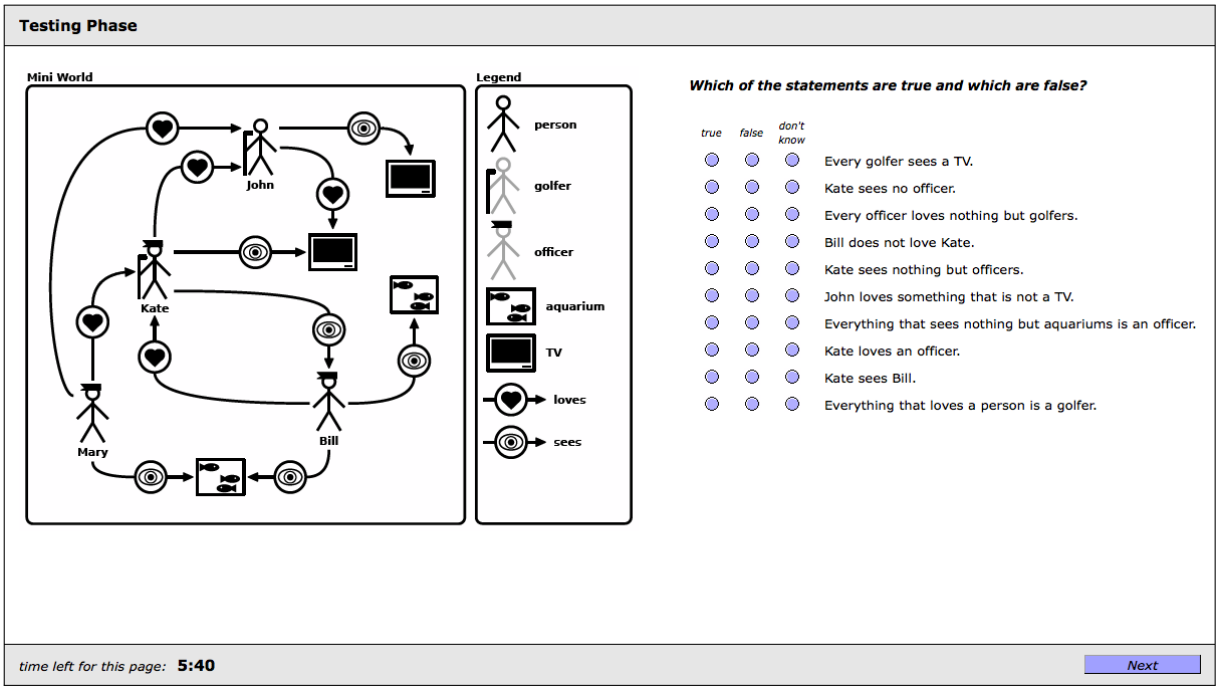

Fig. 3. This is the screen shown during the testing phase of the experiment (for the language ACE).

As a third step, the test run with nine participants (see below) was used to receive more feedback about the understandability and usefulness of the language description sheets. After the test run, the participants were told to highlight everything that was difficult to understand. Only very few things were highlighted and only a couple of small changes had to be made for the main experiment. Altogether, the language description sheets were compiled very carefully.

5.7. Payout

In order to be able to recruit a large number of participants, it was necessary to financially compensate them. The experiment was designed to last approximately one hour: 22 minutes for each of the two tasks plus some additional time for reading the instructions, for answering the control questions, and for filling out the questionnaire. The participants received a fixed amount of 20 Swiss francs as their minimal reward.

In order to provide incentives for good performance, the participants additionally received a variable amount of money depending on the number of statements they managed to classify correctly. Every correct classification was worth 0.60 Swiss francs. Every "don't know" answer and statements for which the time limit ran out gave 0.30 Swiss francs. Thus, the participants could earn between 20 and 32 Swiss francs, and they could get 26 Swiss francs for sure by always choosing "don't know".

While there were monetary incentives for good classification scores, the time needed for the learning and testing phases had no influence on the payout. The reason for this was that quickly choosing "don't know" for all statements (or just guessing quickly and randomly) would become a profitable strategy if short time values gave high rewards. In order to prevent from such behavior, no monetary incentives were given for the amount of time needed. It was assumed that no explicit incentives are necessary for the participants to continue with the procedure when they think that they accomplished the task as good as they could (and this was indeed the case, as the results will show).

5.8. Test Run

In order to test the design of the experiment, a test run was performed before the main experiment with nine participants. They could spend up to 20 minutes for each learning phase and up to 10 minutes for each testing phase.

The participants performed very well. On average, they classified 8.92 out of 10 statements correctly. This is an indication that the experiment in general worked out well. ACE was understood slightly better (0.28 points). Interestingly, four of the nine participants had a perfect score of 10 for both languages. On the one hand, this is good because it shows that the task was feasible and that the instructions were clear. On the other hand, it is bad because it does not return any indication which of the languages was better. For this reason, the task was made slightly harder for the main experiment. This should reduce the number of perfect scores and thus give better indication on which language is better.

The easiest and cleanest way to make the task harder was to reduce the time limits. The maximum learning time was reduced from 20 to 16 minutes and the maximum testing time from 10 to 6 minutes.

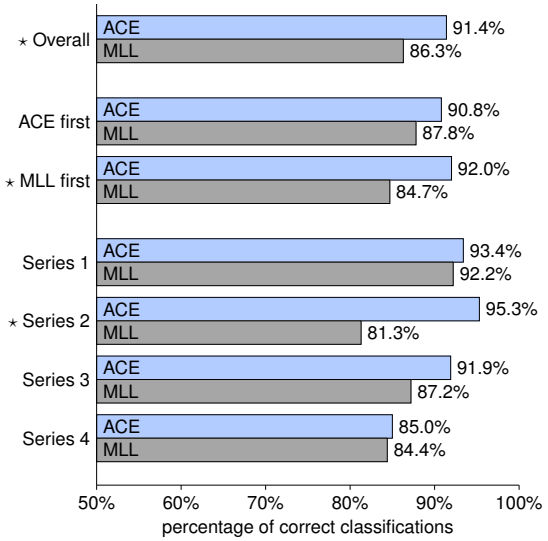

Fig. 4. This chart shows the percentage of correct classifications. Significant differences are marked with "⋆" (see Table 3 for details).

6. Results

The results of the experiment allow for comparing ACE and MLL on the basis of the classification results, the time required by the participants, and the answers they gave in the questionnaire.

6.1. Classification Scores

Figure 4 shows the average percentages of correct classifications per testing phase. "Don't know" answers and the cases where the time limit ran out are counted as 0.5 correct classifications. 50% is the baseline because an average of five correct classifications out of ten can be achieved by mere guessing (or by always choosing "don't know" or by letting the time limit run out).

91.4% of the statements were classified correctly in the case of ACE and 86.3% in the case of MLL. Thus, out of the ten statements of a testing phase, ACE was on average 0.5 points better than MLL. This is a considerable and statistically significant difference (the details of the used statistical test to compare the two samples are explained later on). One has to consider that these values are already close to the ceiling in the form of the perfect score of 10, which might have reduced the actual effect.

The results of the participants who received ACE first and then MLL can now be compared with the ones who received MLL first. As expected, both languages were understood better when they were the second language. This can be explained by the fact that the participants were more familiar with the procedure, the task, and the ontograph notation. However, even in the case when ACE was the first language and MLL the second one, ACE was understood better (but in this case not within statistical significance).

Looking at the results from the perspective of the different series, one can see that ACE was better in all cases but only the series 2 and 3 exhibit a clear dominance of ACE (and this dominance is significant only for series 2). According to these results, one could say that languages like MLL are equally easy to understand for very simple statements as the ones in series 1 and for statements about relations as they appear in series 4. In the case of series 1, the reason might be that these statements are so simple that they can be understood even in a rather complicated language. In the case of series 4, the reason is probably that Description Logic based languages like MLL can express these statements without variables whereas ACE needs variables, which are borderline cases in terms of naturalness.

In summary, the results show that — while both languages are understood reasonably well — ACE is easier to understand than MLL.

6.2. Time

As a next step, we can look at the time values. For simplicity reasons and since the learning process was

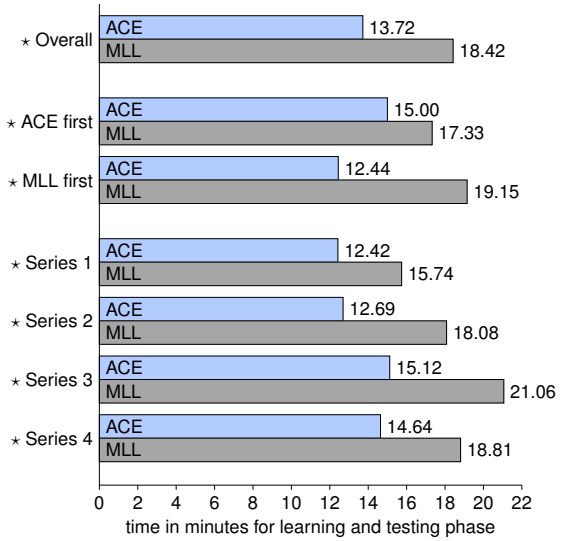

Fig. 5. This chart shows the average time needed for learning and testing phase. Significant differences are marked with "⋆" (see Table 3 for details).

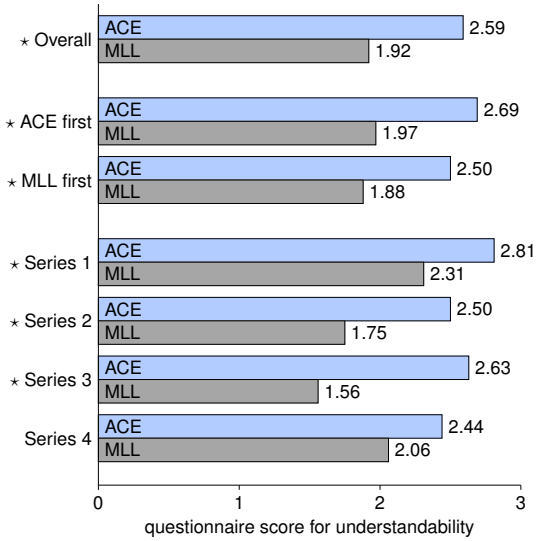

Fig. 6. This chart shows the average scores for perceived understandability derived from the questionnaire. 0 means "very hard to understand", 1 means "hard to understand", 2 means "easy to understand", and 3 means "very easy to understand". Significant differences are marked with "⋆" (see Table 3 for details).

presumably not restricted to the learning phase but continued during the testing phase, the time needed for both phases together is here called *learning time*.

Figure 5 shows these learning times. The participants could spend at most 22 minutes: 16 minutes for the learning phase and 6 minutes for the testing phase. The results show that they needed much less time for ACE than for MLL. In the case of ACE less than 14 minutes were needed, whereas in the case of MLL the participants needed more than 18 minutes. Thus, compared to ACE, MLL required 29% more time to be learned (with a better degree of understanding for ACE, as we have seen above).

Again, the group can be split into the participants who received ACE first and those who received MLL first. Doing so shows the expected effect: ACE and MLL required less time as second language. However, ACE required less time than MLL no matter if it was the first language or the second.

Looking at the different series, we can see that this effect spreads over all four of them. MLL required on average 3 to 6 minutes more learning time than ACE.

The better time values of ACE compared to MLL are statistically significant for the whole sample and also for all presented subsamples.

Analyzing the time values for the two individual phases (learning and testing) reveals that the learning phase contributes much more to the observed difference than the testing phase: 96% of the time difference originates from the learning phase, and only 4% from the testing phase. The testing phase required about the same amount of time for both languages.

6.3. Perceived Understandability

As a third dimension, we can look at the "perceived understandability". The questionnaire that the participants filled out after the experiment contained two questions asking how understandable they found ACE and MLL, respectively. They could choose from four options: "very hard to understand" (value 0), "hard to understand" (1), "easy to understand" (2) and "very easy to understand" (3). The perceived understandability does not necessarily have to coincide with the actual understandability and can be a very valuable measure for the acceptance of a language and the confidence of its users.

Figure 6 shows the scores for perceived understandability. Overall, ACE got much better scores than MLL. MLL was close but below "easy to understand" scoring 1.92, whereas ACE was closer to "very easy to understand" than to "easy to understand" scoring 2.59.

There is no obvious explanation for the fact that both languages scored better when ACE was the first language. It might just be a statistical artifact.

For all four series, ACE received clearly better scores. It is interesting that this also holds for the series 1 and 4 where ACE was not much better than MLL in terms of actual understanding. Thus, even though the actual understanding of these statements does not show a clear difference, the acceptance and confidence of the participants seem to be higher in the case of ACE.

6.4. Significance

Wilcoxon signed-rank tests [28] are applied for checking the statistical significance of the observed differences between ACE and MLL. Table 3 shows the obtained p-values for the three dimensions (i.e. classification score, time, and questionnaire score). For the complete sample, the values are well within the 95% confidence level for all three dimensions (in fact, they are even within the 99% level).[4]

6.5. Regression

In a next step, a regression analysis can be performed to find out which factors were relevant for a good understandability of the two languages. The data set consists of the classification scores of the 128 test phases (two test phases for each of the 64 participants). Apart from the factors that originate from the experiment design (i.e. the two different languages in different order and the different ontograph series), it can be speculated that the gender of the participants, their age, and their English skills could have an influence on the results.

Table 4 shows the result of the regression test. *sc_norm* is the dependent variable and stands for the classification score normalized to 5 (i.e. the normalized score is obtained by subtracting 5 from the number of correctly classified statements). The normalization has the effect that 0 stands for what can be achieved, on average, without understanding.

There are eight independent variables: *ace*, *first_lang*, *series_2*, *series_3*, *series_4*, *female*, *age_above_18* and *very_good_engl*. *ace* stands for the language that is tested, with 0 meaning MLL and 1 meaning ACE. *first_lang* is 1 if the language was the first language, and 0 if it was the second one. *series_2* is 1 if the test was performed on series 2, and otherwise it is 0. In the same way, *series_3* and *series_4* represent series 3 and 4, respectively. Series 1 is encoded by setting all these three variables to 0. *female* determines the gender of the participant: 0 means male, 1 means female. *age_above_18* is an integer denoting the age in years of the participant after subtracting 18 years. *very_good_engl*, finally, stands for the degree of English understanding of the participants as they stated it in the questionnaire. 0 means that the participant has good English skills but not very good ones; 1 stands for very good English skills.

Thus, the baseline of the regression (i.e. the case where all independent variables are zero) is about testing MLL as the second language using series 1 where the participant is a 18 year old male person with good (but not very good) English skills. This baseline situation is represented by the constant coefficient *cons*. The value of the constant coefficient can be interpreted as the average normalized score in the baseline situation. Thus, an 18 year old male person with good English skills who was tested on series 1 in MLL as the second language scored on average 9.302 ("denormalized" by adding 5 to 4.302) points out of 10. The values of the other coefficients can be interpreted as the difference from the baseline when changing the respective factor.

If ACE is tested instead of MLL (but everything else is left unchanged) then the score was on average 0.52 points higher (which is in line with the result described above). This effect is statistically significant.

The regression test confirms that performance for second languages was better than for first ones. Being the first language decreased the score on average by 0.22 points. However, this is not significantly different from 0.

As expected, series 1 led to the best scores. Using series 2, 3 or 4 decreased the score on average by 0.28 to 0.88 points compared to series 1, but this is again not statistically significant.

The possible speculation that the gender of the participants has an influence on their performance could not be confirmed. Women on average performed better than men, but not significantly.

A significant effect could be found, however, with respect to the age of the participants. The age of 18 years was chosen as the baseline because this was the minimal age encountered. The older the participants were the worse they performed. Every year of additional age led to a significant decrease of the score by on average 0.07 points.

Finally, we can look at the degree of English skills of the participants. The participants were mostly native German speakers. In order to ensure that they had suf-

[4] Since all these tests were pre-planned and are very straightforward, no adjustments are applied to handle type I error inflation.

Table 3

This table shows the p-values of Wilcoxon signed-rank tests. The null hypothesis is that the given values are not different for ACE and for MLL. This null hypothesis can be rejected in 16 of the 21 cases on a 95% confidence level and these cases are marked with "⋆"

	classification score		time		questionnaire score	
complete sample	0.003421	⋆	1.493×10^{-10}	⋆	3.240×10^{-7}	⋆
ACE first	0.2140		0.006640	⋆	7.343×10^{-5}	⋆
MLL first	0.005893	⋆	3.260×10^{-9}	⋆	0.001850	⋆
Series 1	0.5859		0.01309	⋆	0.02148	⋆
Series 2	0.003052	⋆	0.002624	⋆	0.02197	⋆
Series 3	0.1250		9.155×10^{-5}	⋆	0.0004883	⋆
Series 4	0.6335		0.002686	⋆	0.1855	

Table 4

This table shows the result of the regression analysis with *sc_norm* being the dependent variable. To the left, a small descriptive analysis of the data is shown. "⋆" indicates the coefficients of the regression analysis that are statistically significant on a 95% confidence level

	descriptive		*regression*				
	Avg.	Range	Coef.	Std. Err.	t	$P > \|t\|$	
sc_norm	3.883	−2 to 5					
ace	0.500	0 or 1	0.516	0.180	2.86	0.006	⋆
first_lang	0.500	0 or 1	−0.219	0.180	−1.22	0.229	
series_2	0.250	0 or 1	−0.480	0.337	−1.42	0.159	
series_3	0.250	0 or 1	−0.278	0.349	−0.80	0.429	
series_4	0.250	0 or 1	−0.880	0.522	−1.69	0.097	
female	0.422	0 or 1	0.141	0.298	0.47	0.637	
age_above_18	4.109	0 to 28	−0.072	0.030	−2.44	0.018	⋆
very_good_engl	0.391	0 or 1	0.203	0.297	0.68	0.496	
cons			4.302	0.325	13.23	0.000	⋆

ficient knowledge of English to understand the ACE sentences, they were asked in the questionnaire about their English skills. The four choices were "(almost) no skills", "only little skills", "good skills" and "very good skills". The first two choices were only for control reasons because such people do not meet the requirements of the experiment. Thus, the data set only contains participants with good or very good English skills. As one could expect, people with very good skills performed better than those who had only good skills, but not significantly.

6.6. Individual Statements

Finally, we can look at the individual statement patterns and how they were classified by the participants. Figure 7 visualizes these results. With the exception of one ACE statement pattern and four MLL statement patterns, the predominance of correct classifications over wrong ones is statistically significant on a 95% confidence level using simple binomial tests.

In the case of ACE, the statements of each pattern were correctly classified by at least 69% of the participants. In contrast, the statements of some MLL patterns were correctly classified in only 50% of the cases. While there is no pattern for which MLL scored more than 2 points better than ACE, there are six patterns for which ACE scored 3 or more points better than MLL. It is interesting to have a closer look at these six statement patterns: 1/3, 1/5, 2/5, 2/8, 3/8 and 4/9. Table 5 shows examples in ACE and MLL for these six patterns to give an impression of the better understandability of ACE.

The statements of 1/3 are simple statements using "or" like "Mary is an officer or is a golfer", which seems to be easier to understand than the respective sentence in MLL "Mary HasType officer or golfer".

1/5 is a very interesting case: this pattern represents plain subtype statements. Such statements are very simple and essential for the creation of ontologies. Many existing ontologies consist to a large extent of such statements. In MLL, they are represented

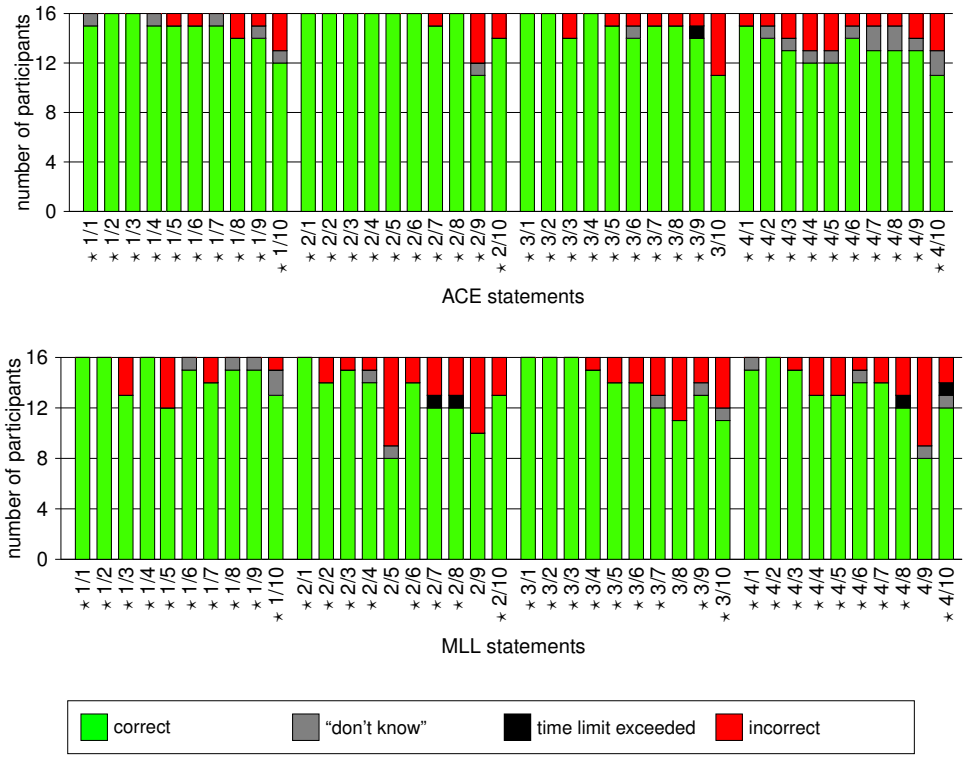

Fig. 7. These two charts show the classification results for each statement pattern and for each of the two languages. The predominance of the correct classifications over the incorrect ones is statistically significant for all patterns marked with "⋆".

Table 5

This table shows the sentence patterns where ACE was clearly better understood than MLL

PATTERN	ACE EXAMPLE	MLL EXAMPLE
1/3	Mary is an officer or is a golfer.	Mary **HasType** officer **or** golfer
1/5	Every golfer is a man.	golfer **SubTypeOf** man
2/5	John buys something that is not a present.	John **HasType** buys **some** (**not** present)
2/8	Everything that buys a present is a man.	buys **some** present **SubTypeOf** man
3/8	Everything that inspects at least 2 letters is an officer.	inspects **min** 2 letter **SubTypeOf** officer
4/9	If X loves Y then X helps Y. If X helps Y then X loves Y.	loves **EquivalentTo** helps

in the form "golfer SubTypeOf man". The results show that the ACE version "every golfer is a man" is easier to understand.

The biggest difference between ACE and MLL is manifested by the statements of 2/5. While everyone scored perfectly in the case of ACE, only half of the participants were able to do so in the case of MLL. These statements are relatively complex and are represented in ACE for example as "John buys something that is not a present" and in MLL as "John HasType buys some (not present)". Arguably, the reason for this difference is that the combination of the operators "some" and "not" is hard to understand in the case of MLL whereas it comes completely natural in the case of ACE.

The statements of 2/8 and 3/8 are similar in the sense that both are subtype statements with a complex left hand side. In MLL, they are represented for example by "buys some present SubTypeOf man" and "inspects min 2 letter SubTypeOf officer", which seems to be hard to understand. The ACE versions of these examples are "everything that buys a present is a man" and "everything that inspects at least 2 letters is an officer", respectively. Thus, the keyword "SubTypeOf" in

general seems to be rather hard to understand, whereas the natural quantifier "every" is understood very well.

4/9, finally, was also understood much better in the case of ACE. The respective statements denote the equivalence of two relations. This case is interesting because the respective statements look very different in ACE and in MLL. MLL expresses this with short statements like "loves EquivalentTo helps", whereas in ACE the same statement is expressed in a much more verbose manner: "If X loves Y then X helps Y. If X helps Y then X loves Y.". Even though the meaning of "EquivalentTo" was clearly explained in the language description sheets for MLL, the more verbose ACE version was understood much better by the participants. Variables as ACE uses them (which are rather rare in natural language) seem to be easier to understand than the more abstract keywords of MLL that do not depend on variables.

7. Conclusions

The results show that ACE is understood significantly better than MLL, which is a simplified version of the Manchester OWL Syntax. Furthermore, ACE required much less time to be learned and was perceived as more understandable by the participants. These results suggest that CNLs should be used instead of languages like the Manchester OWL Syntax, at least in situations where users have to deal with knowledge representations after little or no training, or if they are unfamiliar with formal notations.

Certain patterns could be identified, for which the positive effect of controlled natural language compared to classical ontology languages seems to be particularly strong. Interestingly, plain subtype statements belong to this category.

What remains open to be studied are the shapes of the learning curves for both types of languages. It could be that common logic notations catch up or get ahead of CNL in terms of understandability when people have learned and used them over longer periods of time. However, considering that statements in languages like ACE are about as concise as statements in notations like MLL, and above all that the natural language structures that make up a CNL have been shaped by the evolution of man over countless generations, this all makes it seem unlikely that artificial language structures at some point become easier to understand for human beings than natural ones.

Of course, controlled natural languages also have their drawbacks compared to common formal languages: They are slightly harder to parse by a computer, and they can easily be confused with full natural language. However, these are just minor problems. A real problem is the fact that CNL statements are much harder to *write* than to read and understand. Different solutions to this problem have been proposed and evaluated elsewhere [21,23,2,14,6].

Altogether, the presented work shows that controlled natural language can potentially lead to user interfaces that are easier and faster to understand than interfaces based on common ontology languages. The results justify the past efforts on designing CNLs for the Semantic Web and should encourage future initiatives.

References

[1] Raffaella Bernardi, Diego Calvanese, and Camilo Thorne. Lite natural language. In Jeroen Geertzen, Elias Thijsse, Harry Bunt, and Amanda Schiffrin, editors, *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*, Tilburg, Netherlands, 2007. Tilburg University.

[2] Abraham Bernstein and Esther Kaufmann. GINO — a guided input natural language ontology editor. In Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora Aroyo, editors, *The Semantic Web - ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 144–157. Springer, Berlin / Heidelberg, November 2006.

[3] Johan Bos. Let's not Argue about Semantics. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2835–2840. European Language Resources Association (ELRA), Paris, 2008.

[4] Peter Clark, Phil Harrison, Thomas Jenkins, John Thompson, and Richard H. Wojcik. Acquiring and using world knowledge using a restricted subset of English. In Ingrid Russell and Zdravko Markov, editors, *Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2005)*, pages 506–511. AAAI Press, Menlo Park, CA, USA, 2005.

[5] Anne Cregan, Rolf Schwitter, and Thomas Meyer. Sydney OWL Syntax - towards a controlled natural language syntax for OWL 1.1. In Christine Golbreich, Aditya Kalyanpur, and Bijan Parsia, editors, *Proceedings of the OWLED 2007 Workshop on OWL: Experiences and Directions*, volume 258 of *CEUR Workshop Proceedings*. CEUR-WS, 2007.

[6] Enrico Franconi, Paolo Guagliardo, and Marco Trevisan. An intelligent query interface based on ontology navigation. In Siegfried Handschuh, Tom Heath, VinhTuan Thai, Ian Dickinson, Lora Aroyo, and Valentina Presutti, editors, *Proceedings of the Workshop on Visual Interfaces to the Social and Semantic Web (VISSW 2010)*, volume 565 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[7] Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn. Attempto Controlled English for knowledge representation. In Cristina Baroglio, Piero A. Bonatti, Jan Małuszyński, Massimo Marchiori, Axel Polleres, and Sebastian Schaffert, editors, *Reasoning Web — 4th International Summer School 2008*, volume 5224 of *Lecture Notes in Computer Science*, pages 104–124. Springer, Berlin / Heidelberg, 2008.

[8] Adam Funk, Brian Davis, Valentin Tablan, Kalina Bontcheva, and Hamish Cunningham. Controlled language IE components version 2. SEKT Project Deliverable D2.2.2, University of Sheffield, UK, 2007.

[9] Adam Funk, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham, Brian Davis, and Siegfried Handschuh. CLOnE: Controlled language for ontology editing. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference (ISWC 2007 + ASWC 2007)*, volume 4825 of *Lecture Notes in Computer Science*, pages 142–155. Springer, Berlin / Heidelberg, 2007.

[10] Catalina Hallett, Donia Scott, and Richard Power. Composing questions through conceptual authoring. *Computational Linguistics*, 33(1):105–133, 2007.

[11] Glen Hart, Martina Johnson, and Catherine Dolbear. Rabbit: Developing a controlled natural language for authoring ontologies. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *The Semantic Web: Research and Applications — 5th European Semantic Web Conference (ESWC 2008)*, volume 5021 of *Lecture Notes in Computer Science*, pages 348–360. Springer, Berlin / Heidelberg, 2008.

[12] Matthew Horridge, Nick Drummond, John Goodwin, Alan L. Rector, Robert Stevens, and Hai Wang. The Manchester OWL syntax. In Bernardo Cuenca Grau, Pascal Hitzler, Conor Shankey, and Evan Wallace, editors, *Proceedings of the OWLED '06 Workshop on OWL: Experiences and Directions*, volume 216 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006.

[13] Kaarel Kaljurand. *Attempto Controlled English as a Semantic Web Language*. PhD thesis, Faculty of Mathematics and Computer Science, University of Tartu, Estonia, December 2007.

[14] Tobias Kuhn. How controlled English can improve semantic wikis. In Christoph Lange, Sebastian Schaffert, Hala Skaf-Molli, and Max Völkel, editors, *Proceedings of the Fourth Semantic Wiki Workshop (SemWiki 2009)*, volume 464 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[15] Tobias Kuhn. *Controlled English for Knowledge Representation*. PhD thesis, Faculty of Economics, Business Administration and Information Technology of the University of Zurich, Switzerland, 2010.

[16] Tobias Kuhn. An evaluation framework for controlled natural languages. In Norbert E. Fuchs, editor, *Proceedings of the Workshop on Controlled Natural Language (CNL 2009)*, volume 5972 of *Lecture Notes in Computer Science*, pages 1–20. Springer, Berlin / Heidelberg, 2010.

[17] Massimo Marchiori. Towards a people's web: Metalog. In *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 320–326. IEEE Computer Society, Washington, DC, USA, 2004.

[18] Philippe Martin. Knowledge representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English. In Uta Priss, Dan Corbett, and Galia Angelova, editors, *Conceptual Structures: Integration and Interfaces — Proceedings of the 10th International Conference on Conceptual Structures (ICCS 2002)*, volume 2393 of *Lecture Notes in Artificial Intelligence*, pages 77–91. Springer, Berlin / Heidelberg, 2002.

[19] Adam Pease and John Li. Controlled english to logic translation. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, pages 245–258. Springer Netherlands, 2010.

[20] Jonathan Pool. Can controlled languages scale to the Web? In *Proceedings of the 5th International Workshop on Controlled Language Applications (CLAW 2006)*. Machine Translation Archive, www.mt-archive.info, 2006.

[21] Richard Power, Donia Scott, and Roger Evans. What you see is what you meant: direct knowledge editing with natural language feedback. In Henri Prade, editor, *Proceeding of the 13th European Conference on Artificial Intelligence*, pages 677–681. John Wiley & Sons, Chichester, UK, 1998.

[22] Rolf Schwitter, Kaarel Kaljurand, Anne Cregan, Catherine Dolbear, and Glen Hart. A comparison of three controlled natural languages for OWL 1.1. In *Proceedings of the Fourth OWLED Workshop on OWL: Experiences and Directions*, volume 496 of *CEUR Workshop Proceedings*. CEUR-WS, 2008.

[23] Rolf Schwitter, Anna Ljungberg, and David Hood. ECOLE — a look-ahead editor for a controlled language. In *Controlled Translation — Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Application Workshop (EAMT-CLAW03)*, pages 141–150, Ireland, 2003. Dublin City University.

[24] Rolf Schwitter and Marc Tilbrook. Controlled Natural Language meets the Semantic Web. In Ash Asudeh, Cecile Paris, and Stephen Wan, editors, *Proceedings of the Australasian Language Technology Workshop 2004*, volume 2 of *ALTA Electronic Proceedings*, pages 55–62. Australasian Language Technology Association, 2004.

[25] Rolf Schwitter and Marc Tilbrook. Let's talk in Description Logic via controlled natural language. In Eric McCready, editor, *Proceedings of the Third International Workshop on Logic and Engineering of Natural Language Semantics (LENLS2006)*, pages 193–207, Tokyo, 2006. The Japanese Society for Artificial Intelligence.

[26] John F. Sowa. Common logic controlled English (draft). `http://www.jfsowa.com/clce/specs.htm`, February 2004.

[27] Valentin Tablan, Tamara Polajnar, Hamish Cunningham, and Kalina Bontcheva. User-friendly ontology authoring using a controlled language. In *Proceedings of LREC 2006 — 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[28] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.

[29] Adam Wyner, Krasimir Angelov, Guntis Barzdins, Danica Damljanovic, Brian Davis, Norbert Fuchs, Stefan Hoefler, Ken Jones, Kaarel Kaljurand, Tobias Kuhn, Martin Luts, Jonathan Pool, Mike Rosner, Rolf Schwitter, and John Sowa. On controlled natural languages: Properties and prospects. In Norbert E. Fuchs, editor, *Proceedings of the Workshop on Controlled Natural Language (CNL 2009)*, volume 5972 of *Lecture Notes in Computer Science*, pages 281–289. Springer, Berlin / Heidelberg, 2010.