Towards a More Natural Multilingual Controlled Language Interface to OWL

Normunds Gruzitis and Guntis Barzdins
Institute of Mathematics and Computer Science, University of Latvia
normundsg@ailab.lv, guntis@latnet.lv

Abstract

The paper presents an ongoing research that aims at OWL ontology authoring and verbalization using a deterministic controlled natural language (CNL) that would be as natural and intuitive as possible. Moreover, we focus on a multilingual CNL interface to OWL by considering both highly analytical and highly synthetic languages (namely, English and Latvian). We propose a flexible two-level translation approach that is enabled by the Grammatical Framework and that has allowed us to develop a more natural, but still predictable multilingual CNL on top of the widely used Attempto Controlled English (its subset for OWL, ACE-OWL). This has also allowed us to exploit the readily available ACE parser and verbalizer not only for the modified and extended version of ACE-OWL, but also for the corresponding controlled Latvian.

1 Introduction

Several notations are widely used to make the formal OWL ontologies more intelligible for both domain experts and knowledge engineers. They can be divided in several groups: graphical notations, like UML and its profiles (Barzdins et al., 2010), controlled natural languages (CNL), like Attempto Controlled English or ACE (Kaljurand and Fuchs, 2007), and human-readable formal syntaxes, like the Manchester OWL Syntax (Horridge et al., 2006). The latter kind of notation explicitly follows the underlying formalism and therefore requires substantial training to obtain acceptable reading and writing skills. CNL, in contrast, provides the most informal and intuitive means for knowledge representation and has been successfully used in ontology authoring, where involvement of domain experts is crucial (Dimitrova et al., 2008). Graphical notations are in between and provide a complementary view, unveiling the high-level structure of the ontology in a more comprehensible way. In this paper we focus on untrained domain experts and end-users, and, thus, on CNL that has to be as natural and grammatical as possible. Moreover, we focus on multilingual ontology verbalization to facilitate ontology localization and reuse.

Note that CNL has to ensure deterministic interpretation of its statements, and bidirectional mapping to OWL, so that the CNL user could easily predict or grasp the precise meaning of the specification that is being written or read, and so that the roundtrip from OWL to CNL and back would not introduce any semantic changes in the ontology (if the user has not made changes in the verbalization). In addition to the highly restricted syntactic subset of full natural language, this is typically achieved by a small set of interpretation rules and a monosemous (domain-specific) lexicon.

The state of the art CNLs for OWL (Schwitter et al., 2008) are based on English — a highly analytical language (strict word order, simple morphology, systematic use of determiners) that facilitates the rather straightforward translation of CNL sentences into their semantic representation (axioms in description logic). Regardless of the chosen notation, English is often used also as a meta-language for naming the logical symbols (class and property names) at the ontology level.

Angelov and Ranta (2010) have recently shown that the Grammatical Framework (GF), a formalism and a resource grammar library that provide means for developing parallel grammars, is a convenient framework for rapid implementation of multilingual CNLs. Such seamless cross-translation capability allows easy reuse of the tools developed for existing CNLs — in this way we will reuse the ACE to OWL and OWL to ACE translators.

However, in the case of highly synthetic languages (like Slavic and Baltic) that have rich morphology and relatively free word order, the bidirectional translation to English (i.e., ACE or some other CNL) is not straightforward, especially if we are dealing with statements that represent not only axioms[1] but also rules. For rules (such as SWRL), anaphoric noun phrases (NP) are frequently used: in English they are marked by the definite article, while in Baltic and in most of the Slavic languages such markers are generally not explicitly used and are not encoded even in noun endings. Thus, one of the central problems during the semantically precise translation is how to distinguish between axioms and rules, and how to convey, which information is new (potential antecedents) and which is already given (anaphors).

In this paper we primarily consider Latvian — a member of the Baltic language group. In Section 2 we briefly describe its design and coverage. In Section 3 we illustrate the proposed two-level approach that is used to translate controlled Latvian to (and from) OWL via ACE as an interlingua[2]. We show that this approach allows also for flexible and independent development of an extended and/or modified (adjusted) controlled English interface at the end-user side, if compared to ACE, especially its subset for OWL (ACE-OWL). We conclude the paper with a brief discussion on the current results and future tasks.

2 Grammar

The information structure of a sentence indicates what we are talking about (the topic) and what we are saying about it (the focus) (Hajicova, 2008). In (controlled) English, changes in the information structure typically are reflected by the use of different syntactic constructions, for instance, by using the passive voice instead of the active voice. In Latvian, this is typically reflected by a different word order, for instance, by changing a subject-verb-object (SVO) sentence into OVS or SOV sentence. Thus, in languages like Latvian the word order is syntactically (rather) free, but semantically bound.

Although the topic and focus parts of a sentence, in general, are not reflected by systematic (deterministic) changes in the word order, it has been shown (Gruzitis, 2010) that, in the case of controlled Latvian, the information structure of a sentence can be systematically and reliably conveyed by relying on simplified analysis of the topic-focus articulation (TFA), i.e., on simple word order patterns: if the object comes after the verb (the neutral word order) it belongs to the focus part of the sentence (new information), but if it precedes the verb — to the topic part (given information). As the initial evaluation shows (Gruzitis et al., 2010), the "correct" word order is both intuitively satisfiable by a native speaker and enables the automatic detection of anaphoric NPs in controlled Baltic languages (Latvian and Lithuanian). The simplified TFA method can be adjusted also to controlled Slavic languages.

It should be noted that in Latvian it could be theoretically possible to impose the mandatory use of artificial determiners, by using, for example, indefinite and demonstrative pronouns, however, such "articles" would be unnatural in most cases. Lack of articles is even more apparent in Lithuanian, which, in contrast to Latvian, has no historic influence from the comparatively analytical German.

The survey by Gruzitis et al. (2010) confirmed other important aspects as well that should be addressed, in order to make controlled Latvian more natural and intuitive:

- Due to the rich morphology, there are various alternatives and certain reductions possible in the syntactic realization of a sentence, while preserving both the information structure and the abstract syntax tree (in terms of GF), e.g., making of complex attributes instead of relative clauses may lead to more concise and intelligible sentences[3].
- Explicit determiners ("articles") in certain cases are preferred: an indefinite pronoun ("a") improves the reading of a singular SVO sentence, if the object is not restricted by a relative clause, but a demonstrative pronoun ("the") helps in complex rule statements (in addition to the word order).
- Sentences in the plural are often preferred over their counterparts in the singular.

[1] In this paper we consider only TBox axioms.

[2] Note that any other CNL could be used instead of ACE. We have chosen ACE because of its easily available infrastructure (open source tools and web services) and the active developer community (see `http://attempto.ifi.uzh.ch`).

[3] Such transformations can be applied to a limited extent also in English (e.g., *"animal that eats an animal"* can be expressed as *"animal-eating animal"*).

- Limitations of the OWL expressivity (SVO triples only, no time dimension etc.) to some extent can be lessened on the surface level of the CNL (while preserving the deterministic interpretation), e.g., by using (where appropriate) non-SVO constructions, like adverbial modifiers of place instead of direct objects, and nouns (roles) instead of verbs (actions), and by using the present perfect tense instead of the simple tense (to express a past event that has present consequences).

Therefore, in addition to a grammar that generates the best possible (default) verbalization patterns (taking into account the information structure), we have developed a parallel grammar that allows for completely optional use of determiners and accepts the various syntactic alternatives and extensions[4]. We have also developed a parallel prototype grammar for controlled English that is based on the full ACE[5] with some improvements: we have extended support for the present perfect tense (e.g., by allowing phrases like *"has done something"*), and we have taken a pattern from the Sydney OWL Syntax (Schwitter et al., 2008) to provide an alternative way for expressing inverse nominalized properties (e.g., *"everything has something as a part"* instead of *"everything has-part something"* or *"for everything its part is something"*). It should be mentioned that in the highly inflective controlled Latvian both direct and inverse nominalized properties are verbalized in a more flexible and uniform way.

To achieve a full compliance with the Latvian counterpart, the controlled English grammar has to be further extended with respect to non-SVO sentences (clauses): although adverbial modifiers of place (prepositional constructions) are allowed in the full ACE (e.g., *"someone lives in something"*), there is no support for inverse use of a property in such cases, i.e., it is neither allowed to start a relative clause with the relative pronoun "where", nor to change the fixed word order (like in *"something is a place where someone lives in"*). Again, in controlled Latvian the support for the various relative clauses is ensured in a uniform way.

3 Implementation

The possible steps of our approach that can be performed during the roundtrip from CNL to OWL and vice versa are illustrated in Figure 1. `LavDefSg` is a grammar that defines the default verbalization patterns using Latvian singular sentences, `LavDefPl` is its counterpart for plural sentences, and `LavVar` is an extended combination of both, extensively allowing for free variations (at both the syntactic and lexical level). `LavVar` is used for robust, still predictable parsing (in the ontology authoring direction), while one of the default grammars (depending on the choice of the end-user) — for paraphrasing `LavVar` sentences and for verbalizing existing ontologies. `EngDef` implements the ACE-based English grammar, and `EngVar` provides few lexical and syntactic alternatives. Finally, `AceOwl` implements the chosen interlingua, i.e., accepts/generates sentences that are generated/accepted by the ACE-OWL verbalizer/parser. All these grammars are implemented in GF and are related by a common abstract syntax. Note that translation (reduction) to/from `AceOwl` is an internal step of which the end-user is not aware.

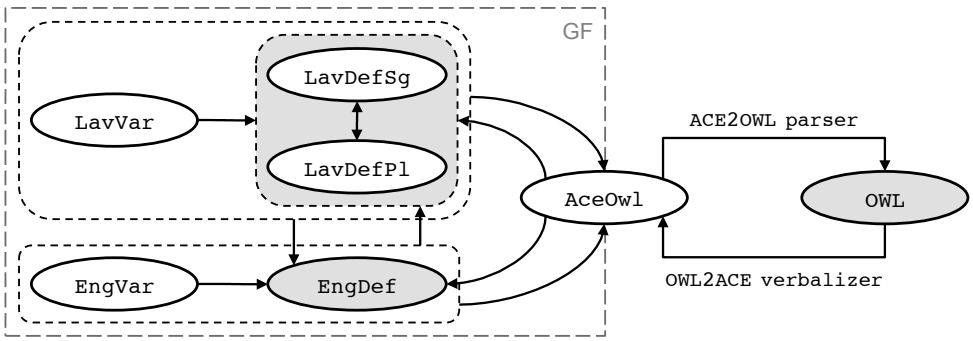

Figure 1: The overall data flow of the automatic translation process among controlled Latvian, English, and OWL. Existing tools are exploited for the transition to/from OWL, using ACE-OWL as an interchange format (covered by the `AceOwl` grammar). Other transitions are ensured by the parallel GF grammars.

[4] An online demo is available at `http://eksperimenti.ailab.lv/cnl/`. Support for plural sentences is being developed.

[5] The full ACE supports prepositional phrases, adjectives and other constructions that are not allowed in ACE-OWL.

Table 1: A sample wildlife ontology, automatically verbalized in controlled English (by `EngDef`) and Latvian (by `LavDefSg`). Underlined are properties that are expressed by nouns (roles) instead of verbs (explicit predicates).

1	Everything that *eats* something is **an** *animal*.	Tas, kas kaut ko *ēd*, ir *dzīvnieks*.
2	Every *carnivore* is **an** *animal* that *eats* **an** *animal*. Every *animal* that *eats* **an** *animal* is **a** *carnivore*.	Ikviens *plēsējs* ir *dzīvnieks*, kas *ēd* **kādu** *dzīvnieku*. Ikviens *dzīvnieks*, kas *ēd* **kādu** *dzīvnieku* ir *plēsējs*.
3	Every *herbivore* is **an** *animal* that *eats* nothing but things that are **a** *plant* or that are **a** *part* of nothing but *plants*.	Ikviens *zālēdājs* ir *dzīvnieks*, kas *ēd* tikai kaut ko, kas ir *augs* vai kas ir tikai *auga daļa*.
4	Every *giraffe* is **a** *herbivore*.	Ikviena *žirafe* ir *zālēdājs*.
5	Everything that is *eaten* by **a** *giraffe* is **a** *leaf*.	Tas, ko *ēd* **kāda** *žirafe*, ir *lapa*.
6	Everything that has **a** *leaf* as **a** *part* is **a** *branch*.	Tas, kura *daļa* ir **kāda** *lapa*, ir *zars*.
7	Every *tasty plant* is **a** *nourishment* of **a** *carnivore*.	Ikviens *garšīgs augs* ir **kāda** *plēsēja barība*.
8	No *animal* is **a** *plant*.	Neviens *dzīvnieks* nav *augs*.
9	If X *eats* Y then Y is **a** *nourishment* of X.	Ja X-s *ēd* Y-u, tad Y-s ir X-a *barība*.

For a demonstration we use a sample African wildlife ontology that is verbalized in Table 1.

During the translation from Table 1 to ACE-OWL (Table 2), all non-SVO statements are reduced to artificial SVO statements (e.g., *"lives in something"* to *"lives-in something"*, *"part of something"* to *"part-of something"*), and all terms are normalized into fixed forms that are conveyed as is to the ontology[6]. The result, in general, is ungrammatical (from the linguistic perspective), but we do not try to make it more grammatical where possible (e.g., the past participle form could be used in the 5th statement) — we use it only as a technical interchange format that normally is not visible to the end-user. However, it is a good illustration that explicitly unveils the nature and limitations of OWL.

Note that certain conversions are done at the end-user level (while paraphrasing from `Var` to `Def`) and are further reflected in OWL. For instance, the present perfect tense can be converted to the simple tense (e.g., *"has done something"* to *"does something"*) or vice versa, if such alternatives are listed in the domain lexicon (individually for each language and property).

Table 2: An automatically generated ACE-OWL text, translated from Table 1 (by the `AceOwl` grammar), or verbalized from the original OWL ontology (by the ACE verbalizer). The prefixes that indicate the POS categories, although accepted by the ACE parser, are used here only for the sake of clarity. The semantic interpretation is acquired by the ACE parser and is given in parallel (in the Manchester notation).

1	Everything that **v:eats** something is an n:animal.	`ObjectProperty: eats Domain: animal`
2	Every n:carnivore is an n:animal that **v:eats** an n:animal. Every n:animal that **v:eats** an n:animal is a n:carnivore.	`Class: carnivore EquivalentTo: animal and (eats some animal)`
3	Every n:herbivore is an n:animal that **v:eats** nothing but things that are a n:plant or that **v:part-of** nothing but n:plant.	`Class: herbivore SubClassOf: animal and (eats only (plant or (part-of only plant)))`
4	Every n:giraffe is a n:herbivore.	`Class: giraffe SubClassOf: herbivore`
5	Everything that is **v:eats** by a n:giraffe is a n:leaf.	`Class: inverse (eats) some giraffe SubClassOf: leaf`
6	Everything that is **v:part-of** by a n:leaf is a n:branch.	`Class: inverse (part-of) some leaf SubClassOf: branch`
7	Every n:tasty-plant **v:nourishment-of** a n:carnivore.	`Class: tasty-plant SubClassOf: nourishment-of some carnivore`
8	No n:animal is a n:plant.	`Class: animal DisjointWith: plant`
9	If X **v:eats** Y then Y **v:nourishment-of** X.	`ObjectProperty: eats InverseOf: nourishment-of`

[6]This is achieved by passing an auto-generated user lexicon to the ACE parser, where all wordforms of each lexical entry are equivalent to that used for the logical symbol.

4 Discussion

The two-level translation approach has allowed us to develop a rather sophisticated multilingual CNL on top of the rather restricted ACE-OWL (in terms of naturalness). Of course, ACE-OWL itself can be developed to be equally natural, but the benefit of our approach is that it allows for more flexible, rapid[7] and independent extensions and adjustments to what users consider the most natural verbalization. The proposed approach enables not only a multilingual, but also a multi-dialect interface to OWL: different CNLs can be mixed together or used in parallel, and the interlingua can be relatively easily changed. It should be reminded that our goal is to ensure a predictable interpretation, therefore we could change the interlingua to CPL-Lite, for instance, but not to CPL, which is non-deterministic (Clark et al., 2010). Also note that GF not only enables the precise cross-grammar translation[8], but also facilitates the application of more flexible and linguistically less restrictive naming conventions at the OWL level.

One might ask why we use an interlingua at all, rather than proceed by translation to and from OWL directly in GF (by providing yet another concrete grammar for the Functional-Style Syntax or some other formal notation of OWL). Indeed, verbalization of existing ontologies could be done in this way, but a problem arises in the reverse direction — form CNL to OWL: the current implementation of GF does not provide support for dealing with anaphors[9]. Thus, by solving the interpretation issues via an interlingua, we get the ontology verbalization functionality for free.

One might also argue that the dependence on a handcrafted domain lexicon is a significant disadvantage. This is the price for flexibility, multilinguality, naturalness and precision. Although it would be possible to generate the English lexicon from a linguistically motivated ontology, the problem is how to acquire the precise translation equivalents. In the case of ontology authoring, common word lexicons could be reused, but, again, the alignment issue arises and specific multi-word units are often used.

In this paper we have considered only terminological (TBox) axioms and rules. It would be interesting to see to what extent the deterministic TFA method can be adjusted for assertional (ABox) statements. However, for populating an ontology with facts (individuals), some other kind of an interface (e.g., GUI forms or tables) could be more appropriate.

References

Angelov, K. and A. Ranta (2010). Implementing controlled languages in GF. In N. E. Fuchs (Ed.), *Controlled Natural Language*, Volume 5972 of *LNAI*, pp. 82–101. Springer.

Barzdins, J., G. Barzdins, K. Cerans, R. Liepins, and A. Sprogis (2010). OWLGrEd: a UML style graphical notation and editor for OWL 2. In *7th International OWLED Workshop*, Volume 614. CEUR.

Clark, P., W. R. Murray, P. Harrison, and J. Thompson (2010). Naturalness vs. predictability: A key debate in controlled languages. In N. E. Fuchs (Ed.), *Controlled Natural Language*, Volume 5972 of *LNAI*, pp. 65–81.

Dimitrova, V., R. Denaux, G. Hart, C. Dolbear, I. Holt, and A. G. Cohn (2008). Involving domain experts in authoring OWL ontologies. In *7th International Conference on the Semantic Web*, Volume 5318 of *LNCS*.

Gruzitis, N. (2010). Word order based analysis of given and new information in controlled synthetic languages. In P. Buitelaar, P. Cimiano, and E. Montiel-Ponsoda (Eds.), *1st International Workshop on the Multilingual Semantic Web*, Volume 571, pp. 29–34. CEUR.

Gruzitis, N., G. Nespore, and B. Saulite (2010). Verbalizing ontologies in controlled Baltic languages. In I. Skadina and A. Vasiljevs (Eds.), *4th International Conference on Human Language Technologies — The Baltic Perspective*, Volume 219 of *Frontiers in Artificial Intelligence and Applications*, pp. 187–194. IOS Press.

Hajicova, E. (2008). What we are talking about and what we are saying about it. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Volume 4919 of *LNCS*, pp. 241–262. Springer.

Horridge, M., N. Drummond, J. Goodwin, A. Rector, R. Stevens, and H. Wang (2006). The Manchester OWL syntax. In *2nd International Workshop on OWL: Experiences and Directions (OWLED)*.

Kaljurand, K. and N. E. Fuchs (2007). Verbalizing OWL in Attempto Controlled English. In *3rd International Workshop on OWL: Experiences and Directions (OWLED)*.

Schwitter, R., K. Kaljurand, A. Cregan, C. Dolbear, and G. Hart (2008). A comparison of three controlled natural languages for OWL 1.1. In *4th International OWLED Workshop*, Volume 496. CEUR.

[7] Especially if the GF resource library is used instead of developing all the concrete grammars from scratch.

[8] In few cases, e.g., in ambiguous coordination of relative clauses, the ACE interpretation rules have to be applied afterwards.

[9] Anaphors (incl. explicit variables) may appear not only in rules, but also in statements that define property axioms.