Verbalizing Ontologies in Controlled Baltic Languages

Normunds Grūzītis, Gunta Nešpore, Baiba Saulīte
Institute of Mathematics and Computer Science
University of Latvia

HUMAN LANGUAGE TECHNOLOGIES — THE BALTIC PERSPECTIVE
Riga, October 7-8, 2010
Every professor is a teacher.
No assistant is a professor.
Everyone that teaches something is a teacher.
Everything that is taught by someone is a course.
If X teaches Y then X does not take Y.
If X includes Y then X is constituted by Y.

**Sample Ontology**

Class: owl:Thing and (teaches some MandatoryCourse)  
SubClassOf: Professor  

Everyone that teaches a mandatory course is a professor.

ObjectProperty: enrolls SubPropertyChain: includes o inverse (takes)  
If X includes something that is taken by Y then X enrolls Y.
Motivation


R. Denaux, V. Dimitrova, A. Cohn, C. Dolbear, G. Hart: *Rabbit to OWL: Ontology Authoring with a CNL-based Tool*. In: Workshop on Controlled Natural Language, 2009

“Logic is a wonderful thing, but doesn’t always beat actual thought.”

Terry Pratchet

“This Guide is definitive. Reality is frequently inaccurate.”

Douglas Adams
Type of CNL

• Naturalist approach
  – A simpler form of the full natural language (NL)
  – Ambiguity resides to a lesser extent
  – Search for a best parse and interpretation
    • Heuristics for PP-attachment, WordNet-based WSD, etc.
  – CPL

• Formalist approach
  – An NL-like formal language
  – Well-defined and predictable (deterministic)
    • Fixed interpretation rules (in terms of the underlying formalism)
  – A monosemous lexicon
  – ACE, PENG, Rabbit

Baltic Languages

• **Highly synthetic**: rich morphology, free word order
  
  – **Explicit** linguistic markers, indicating which information is already given (anaphors) and which is new (antecedents), in general, are not available
    
    • “Articles” are rarely used and are “compensated” by more **implicit** linguistic markers; typically, by changes in the **word order**
    
    • The **definiteness** feature is not encoded even in noun endings
    
    • Definiteness feature is **encoded** in adjective and participle endings, however, these markers are **non-reliable** even in controlled language

• Closest sibling to the **Slavic** language group
Information Structure

- **Synthetic** language
  - Syntactically **free** word order
  - Semantically **fixed** word order

- Inspiring from the Prague Linguistic School:
  - Exploitation of the concept of **topic-focus articulation** for controlled synthetic language
    - **TOPIC** – given information – **to the left** from the verb
    - **FOCUS** – new information – **to the right** from the verb
  - **Hypothesis**: in controlled synthetic language “articles” can be reliably “reconstructed” from the **word order**:
    - **Intuitively** satisfiable by a human user
    - Ensures the **deterministic** automatic parsing

![Diagram showing TOPIC and FOCUS]

- **TOPIC**
  - What are we talking about?
- **FOCUS**
  - What are we saying about it?
Survey

• The aim:
  – Test the hypothesis that TFA is a reliable method in the case of CNL
  – Find the most natural and intuitive syntactic patterns that preserve the predictive (unambiguous) interpretation in OWL

• Evaluation of 15–17 statements of various complexity
  – Each statement was verbalized in two or three slightly different ways
  – Alternatives were ranked being either good, acceptable or poor
  – Respondents were able to propose their own suggestions

• ~80 Latvian and ~40 Lithuanian respondents
  – ~75% evaluated all examples; others — at least one third
Suggestions

• Use of the indefinite and demonstrative pronouns in certain cases improves the reading (in Latvian)
  – Ikvienu kursu māca kāds pasniedzējs. (Every course is taught by a teacher.)
  – Ikvienu kursu māca pasniedzējs, kas .. (Every course is taught by a teacher that ..)

• Simple vs. present perfect tense
  – Ikviena akadēmiskā programma ir uzņēmusi/uzņem kādu studentu.
    • Every academic program has enrolled/enrolls a student.

• Direct object vs. adverbial modifier of place
  – Ikvienā studentis ir uzņemts kādā akadēmiskajā programmā.
    • Every student is enrolled in an academic program.

• Relative clause vs. attribute
  – Ikvienā kurss, kas ir iekļauts kādā akadēmiskajā programmā, ..
    • Every course that is included in an academic program ..
  – Ikvienā kādā akadēmiskajā programmā iekļautais kurss ..
    • Every academic-program-included course ..
Pseudo-SVO Statements

- At the OWL level – SVO triples only
- At the CNL level, it can be very hard or even impossible:
  - to come up with an appropriate verb
  - to use an object (accusative case), so that the statement remains natural

- Predicate nominals (roles)
  - Of-constructions in English
  - Genitive (possessive) constructions in Baltic languages

- Adverbial modifiers (of place)
  - Currently we are considering only such modifiers that do not require a preposition, but are expressed by the locative case
    - In English, the preposition “in” or “at” is used
Multilingual Grammar

InverseStatement : S

UniversalQuantification : AllC teaches : P

ExistentialQuantification : ExC

Every : All Class : CBar

Some : Ex Class : CBar

S

Course : C

Teacher : C

= translation

every course is taught by a teacher

powered by Grammatical Framework
ACE as Interlingua

ACE parser

DRS

SWRL

http://eksperimenti.ailab.lv/cnl/
Everyone that teaches something is a teacher.
Everything that is taught by someone is a course.
Every course is a part of an academic program.
Everything that has a course as a part is an academic program.
Conclusion

• In controlled Latvian, which is a highly synthetic CNL, where definite and indefinite articles are not used, the topic-focus articulation can be reflected by systematic changes in the neutral word order
  – A simple and reliable mechanism
  – Native speakers tend to follow such guidelines rather intuitively

• The two-level translation approach has allowed us to develop a rather sophisticated controlled Latvian on the top of the very restricted ACE subset for OWL

• No good solution for the problem of animate/inanimate things

• TODO:
  – Plural sentences: more intuitive in many cases, no indefinite pronouns
  – Prepositional phrases (other than -in and -of)
  – Assertional statements
  – Prototype implementation for Lithuanian language
Thank you!