Word order based analysis of given and new information in controlled synthetic languages

Normunds Grūzītis
Institute of Mathematics and Computer Science, University of Latvia
Raina bulv. 29, Riga, LV-1459, Latvia
normundsg@ailab.lv

ABSTRACT

When an OWL ontology, together with SWRL rules, is defined or verbalized in controlled natural language (CNL) it is important to ensure that the meaning of CNL statements will be unambiguously (predictably) interpreted by both human and machine. CNLs that are based on analytical languages (namely, English) impose a number of syntactic restrictions that enable the deterministic interpretation. Similar restrictions can be adapted to a large extent also for synthetic languages, however, a fundamental issue reveals in analysis of given (topic) and new (focus) information. In highly analytical CNLs, detection of which information is new and which has been already introduced is enabled by systematic use of definite and indefinite articles. In highly synthetic languages, articles are not typically used. In this paper we show that topic-focus articulation in synthetic CNLs can be reflected by systematic changes in word order that are both intuitive for a native speaker and formal for the automatic parsing.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *Natural language interfaces*; I.2.7 [**Natural Language Processing**]

General Terms

Design, Experimentation, Languages

Keywords

Ontology Verbalization, Controlled Natural Language, Synthetic Language, Word Order, Information Structure, Topic-Focus Articulation, Anaphoric References

1. INTRODUCTION

One of the fundamental requirements in definition and verbalization of ontology structure, restrictions, and implication rules is the unambiguous interpretation (in terms of the underlying formalism) of controlled natural language (CNL) statements, so that the CNL user could easily predict the precise meaning of the specification he/she is writing or reading; that also includes the resolving of anaphoric references. To enable deterministic construction of discourse representation structures (DRS), several widely accepted restrictions are used in CNLs (e.g., in Attempto Controlled English [2]): a set of interpretation rules for potentially ambiguous syntactic constructions (an issue that is still present even in a highly restricted syntactic subset of natural language), a monosemous lexicon (i.e., domain-specific terminology), an assumption that the antecedent of a definite noun phrase (NP) is the most recent and most specific accessible NP that agrees in gender and number, and some other limitations.

There are several sophisticated CNLs that provide seemingly informal means for bidirectional mapping between controlled English and OWL [10]. Experiments show that the underlying principles of English-based CNLs can be successfully adapted also for other rather analytical languages, for example, for Afrikaans [7]. Moreover, Ranta and Angelov [8] have shown that the Grammatical Framework (GF), a formalism for implementation of multilingual CNLs, provides convenient means for writing parallel grammars that simultaneously cover similar syntactic fragments of several natural languages. Thus, if the abstract and concrete grammars are carefully designed, GF provides a syntactically and semantically precise translation from one CNL to another (again, assuming that the domain-specific translation equivalents are monosemous). This potentially allows exploitation of powerful tools that are already developed for one or the other "dialect" of controlled English also for non-English CNLs. For instance, the ACE parser [2] could be used for DRS construction, paraphrasing and mapping to OWL, and ACE verbalizer [5] could be used in the reverse direction, facilitating cross-lingual ontology development, verbalization, and querying.

While it seems promising and straightforward for rather analytical CNLs that share common fundamental characteristics, allowing (apart from other) for explicit detection of anaphoric references, it raises issues in the case of highly synthetic languages, where explicit linguistic markers, indicating which information is already given (anaphors) and which is new (antecedents), in general, are not available (here we are talking about individuals that are referenced by NPs, not by anaphoric pronouns). In analytical CNLs, analysis of the information structure of a sentence is based on the strict word order (basically, the subject-verb-object or SVO pattern) and systematic use of definite and indefinite articles. In highly synthetic languages, articles are rarely used and are "compensated" by more implicit linguistic markers; typically, by changes in the neutral word order, which is enabled by rich inflectional paradigms and syntactic agreement.

As a case study, we have chosen Latvian [6], a member of the Baltic language group (together with Lithuanian). Baltic languages are among the oldest of the remaining Indo-European languages. Syntactically they both are very closely related and are highly synthetic with rich morphology; however, the definiteness feature is not encoded even in noun endings as it is in case of Bulgarian [1], for instance. Thus, we will describe the correspondence between the given/new information and word order patterns in terms of topic-focus articulation [3]. Although the topic (theme) and focus (rheme) parts of a sentence in general

WWW 2010, April 26-30, 2010, Raleigh, North Carolina.

Latvian are not always reflected by systematic changes in the word order [9], in this paper we demonstrate that changes in word order are reliable markers in the case of controlled Latvian, allowing for systematic reconstruction of "missing" articles.

Few other markers may be used in Latvian to indicate that an NP is an anaphoric reference, namely, definite and indefinite endings of adjectives and participles, if they are used as attributes. However, such markers are optional and non-reliable even in controlled language — attributes in domain-specific terms (multi-word units) often have definite endings by default. We might also impose the usage of artificial determiners, using indefinite and demonstrative pronouns, but then it would be Latvian-like controlled language, not a subset of actual Latvian. The problem is even more apparent in case of Lithuanian that has not been historically influenced by German. Therefore the only formal and general feature that indicates the status of an NP is its position in a sentence — whether it belongs to the topic or focus part. Our hypothesis is that the requirement for compliance with predefined word order patterns in a controlled synthetic language is not only reasonable, but also makes the CNL more natural and is intuitively satisfiable by a native speaker. As our experiments show, the proposed approach is directly applicable for Lithuanian and can be adapted also for majority of Slavic languages (e.g., Russian and Czech) — the closest siblings to Baltic languages.

2. TERMINOLOGICAL STATEMENTS

In this paper we focus on terminological (TBox) statements of OWL ontologies [15] that are supplemented with limited data integrity constraints in form of SWRL rules [18] and SPARQL queries [16, 17] (see Section 3). In this section we will consider different types of statements defining atomic and complex classes, properties, and property restrictions of a simplified university ontology. Statements are given in parallel in Manchester OWL Syntax [4], ACE [2], and ACE compliant controlled Latvian.

In Figure 1 the example ontology is visualized according to the UML profile for OWL [14] — a user-friendly notation that unveils the structure of the ontology in a highly comprehensible form, but it is not well suited to capture complex restrictions and integrity constraints. In the following two sections we will do both verbalize the UML-defined structure and use CNL to define additional restrictions and integrity constraints.

2.1 Classes

Statements defining class hierarchies consist of subject and subject complement. Subject (topic) is always universally quantified, predicate noun — always existentially quantified:

(1) `Class: Professor SubClassOf: Teacher`
Every professor is a teacher.
*Katrs profesors **ir** ~~kāds~~ **pasniedzējs**.*

For the universal quantifier there is a corresponding determiner/pronoun both in English and Latvian (as well as in other analytical and synthetic languages). As to the existential quantifier there is no counterpart for the indefinite determiner in Latvian. We could artificially use an indefinite pronoun instead, but such construction would be more than odd in this case. Besides the fact that the subject complement is always indefinite it also always appears in the focus part of a sentence (here and further — formatted in bold), i.e., it always is new information and, thus, the explicit linguistic marker (indefinite pronoun) can be omitted without introducing any ambiguities. Similarly, class equivalence can be defined by stating subclass axioms in both directions (in two separate statements), and class disjointness — by substituting the determiner "every" with its antonym "no":

(2) `DisjointClasses: Assistant, Professor`
No assistant is a professor.
*Neviens asistents **nav profesors**.*

In Latvian (and in many other synthetic languages), a negated pronoun is used for the negative universal quantifier, and the statement is negated twice by the copula, but these are minor syntactic differences; the information structure remains the same. This assumption can be directly extended to complex classes that are combined from atomic ones by applying logical constructors:

(3) `Class: Course SubClassOf: owl:Thing and (MandatoryCourse or OptionalCourse)`
Every course is something that is a mandatory course or that is an optional course.
*Katrs kurss **ir kaut kas**, kas **ir obligātais kurss** vai kas **ir izvēles kurss**.*

So far about cases when the verb phrase (VP) is a predicate nominal. Another type of constructors for complex classes are property restrictions — VPs consisting of a transitive verb complemented by a direct object. In the following statement such VP is used to implicitly specify an anonymous superclass.

(4) `Class: Teacher SubClassOf: teaches some Course`
Every teacher teaches a course.
*Katrs pasniedzējs **pasniedz** [kādu] **kursu**.*

Here the role of word order comes in. In Latvian, 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). In the case of the inverse use of a property, the word order is changed for the whole statement (in both languages), moving the agent to the focus:

(5) `Class: Course SubClassOf: inverse (teaches) some Teacher`
Every course is taught by a teacher.
*Katru kursu **pasniedz** [kāds] **pasniedzējs**.*

Combinations of the introduced syntactic phrases can be further used to explicitly specify complex superclasses (Statement 6) and general class axioms, where anonymous is either the subclass (Statement 7), or both the super- and the sub-class (Statement 8).

(6) `Class: Student SubClassOf: Person and (inverse (enrolls) some AcademicProgram)`
Every student is a person that is enrolled by an academic program.
*Katrs students **ir persona**, ko **uzņem** [kāda] **akadēmiskā programma**.*

(7) `Class: owl:Thing and (teaches some MandatoryCourse) SubClassOf: Professor`
Everything that teaches a mandatory course is a professor.
*Katrs, kas **pasniedz** [kādu] **obligāto kursu**, **ir profesors**.*

(8) `Class: owl:Thing and (inverse (includes) some AcademicProgram) SubClassOf: inverse (teaches) some Teacher`
Everything that is included by an academic program is taught by a teacher.
*Katru, ko **ietver** [kāda] **akadēmiskā programma**, **pasniedz** [kāds] **pasniedzējs**.*

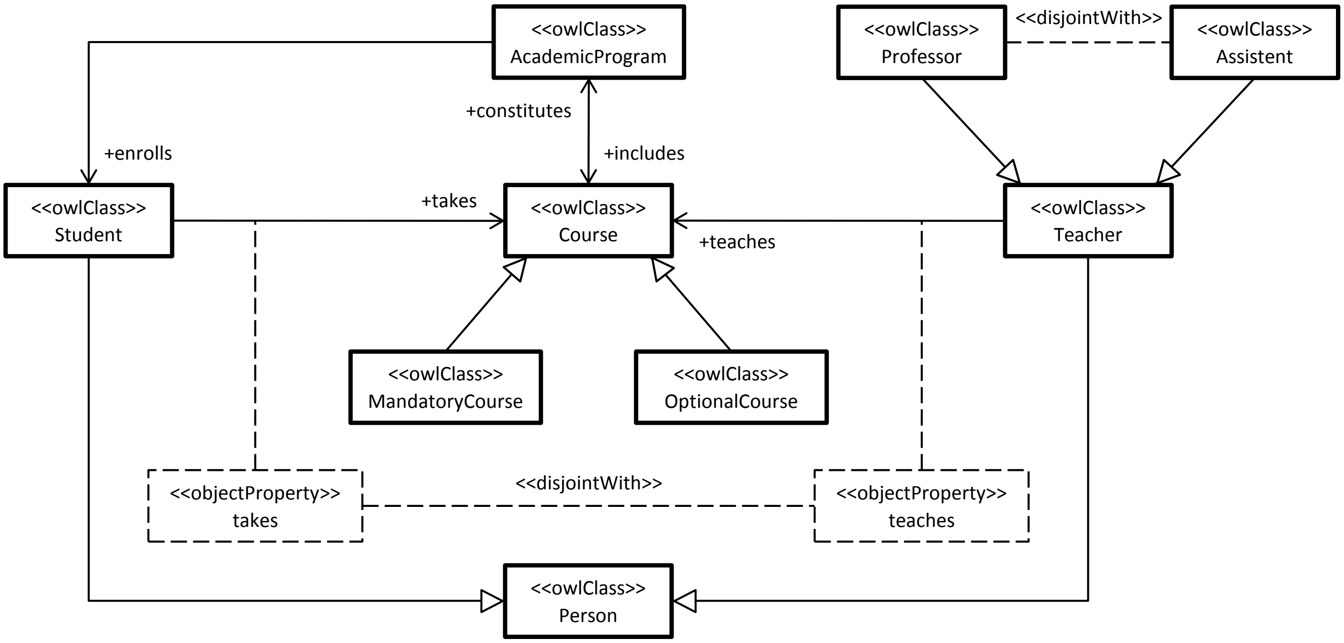

Figure 1. The structure of a simplified university ontology, visualized by the OWL2UML plug-in [19]. No complex class expressions or data integrity constraints are included. The automatically generated diagram is also slightly simplified for printing purposes.

In English, both active and passive voice sentences are still SVO sentences; the inverse direction of the property is indicated by the passive voice. In Latvian, the voice remains active, but the syntactic functions of both NPs are interchanged in such a case, making it an OVS sentence (semantic roles remain the same in both languages). Thus, it turns out that the object stands to the left from the verb — in the topic part, indicating that it should be given information. However, recall that in TBox statements there is no doubt which determiner has to be assigned with the topic — it is always universally quantified (unless it is an anaphoric pronoun that links a relative clause to its anchor).

Also, it should be mentioned that in the above provided SVO statements (4–8) the indefinite pronoun "kāds" is given in square brackets, which means that in these cases it might be optionally used as a counterpart for the indefinite article. This conforms to the language intuition (see Section 4) and emphasizes the indefiniteness of the NP — the explicit marker improves readability (interpretation) as there is no relative clause associated with the NP, which would then serve as an indicator of indefiniteness.

Another aspect that should be mentioned is that there has not been made any differentiation between animate and inanimate things — quantifiers "everything", "something", "nothing", and the relative pronoun "that" are used in all cases. This makes Statement 7 in English and Statement 8 in Latvian odd, which has been noticed by our respondents (see Section 4). However, we ignore this issue in this paper for the sake of simplicity and for compliance with the ACE verbalizer [5] (although ACE parser supports such differentiation, it is discarded at the ontological level).

2.2 Properties

The already introduced syntactic constructions can also be used to define properties and their restrictions. Thus, to specify the domain and range of a property, in the topic part of the statement one has to refer the universal class, which is then specified in the relative clause, referring the property of interest. The subject is complemented by a predicate nominal in the focus part (see the statements 9–10). Note that usage of definite and indefinite NPs (i.e., references to concrete classes) is not possible in property definitions (except when stating the domain and range), as this would go beyond the expressivity of OWL (see the next section).

(9) `ObjectProperty: teaches Domain: Teacher`
Everything that teaches something is a teacher.
*Katrs, kas **kaut ko pasniedz, ir pasniedzējs**.*

(10) `ObjectProperty: teaches Range: Course`
Everything that is taught by something is a course.
*Katrs, ko **kaut kas pasniedz, ir kurss**.*

Although property hierarchies, characteristics, and chains can be defined in the same manner, by additionally exploiting the anaphoric pronoun "it", in many cases usage of if-then constructions together with variables is at least more concise, if not more comprehensible as well (especially in property chaining). Exceptions are functional and inverse functional properties that are defined by cardinality restrictions and can be stated more naturally without anaphoric references, and reflexive and irreflexive properties — their verbalization in CNL is more natural if the reflexive pronoun "itself" is used.

For instance, the definition of a property chain given in Statement 11 could be paraphrased by using pronouns instead of variables — "*Everything that includes something that is taken by something enrolls it.*" — but such paraphrase would more likely confuse a human interpreter, especially when resolving the anaphoric reference.

(11) `ObjectProperty: enrolls SubPropertyChain: includes o inverse (takes)`
If [*something*] *X includes something that is taken by* [*something*] *Y then X enrolls Y.*
Ja [*kaut kas*] *X ietver kaut ko, ko ņem* [*kaut kas*] *Y, tad X uzņem Y.*

Note that the pronoun "something" may be omitted in the apposition phrases — this not only makes the statements more comprehensible, but also allows to reduce or even to hide the issue of discrimination between animate and inanimate things (accordingly, X and Y in Statement 11). For instance, when declaring that two properties are disjoint, we can avoid the use of the indefinite pronoun at all:

(12) `DisjointProperties: teaches, takes`
If X teaches Y then X does not take Y.
Ja X pasniedz Y, tad X neņem Y.

The concise form, however, has a drawback in the case of a highly synthetic CNL. In English, the strict word order (and the change of the voice) enables the unambiguous detection of which variable represents the agent in any of the SVO chunks. In Latvian, provided that all the properties involved are represented by transitive verbs (instead of comparative phrases, for instance, as in "*X is smaller than Y*"), the agent/subject can be recognized only due to the different ending if compared with the object; the verb itself does not change. Plain variables, of course, are not inflected. Although for a human interpreter it usually causes no ambiguities (due to the rich background knowledge, and knowledge of lexical semantics), suffixes have to be added to the variables to enable the automatic parsing. Nevertheless, this is still a more user-friendly solution (see Statement 13) than the use of the artificial apposition phrases. Moreover, even if indefinite apposition phrases are used, they are applicable only in the if-clauses; for the then-clauses definite apposition phrases should be introduced, making such statements even more unnatural.

(13) `ObjectProperty: includes InverseOf: constitutes`
If X includes Y then X is constituted by Y.
Ja X-s ietver Y-u, tad X-u veido Y-s.

Although property axioms can be seen as a special case, variables may be used in statements defining classes as well (e.g., Statement 7 in Section 2.1 can be paraphrased in ACE as "*If X teaches a mandatory course then X is a professor.*"). The formal nature of CNL then becomes explicit more widely, losing the seeming naturalness that, of course, is not a self-purpose; variable constructions should be allowed as an alternative to improve readability in certain cases (e.g., for tracking coreferences in complex rules). Allowing for such alternatives, however, introduces an issue in the verbalization direction — how to decide (encode in the grammar) in which cases variables are preferable over indefinite and definite NPs.

Nevertheless, variable constructions are partially out of the scope of this paper, as there is no need for information structure analysis to cope with utterances of anaphoric pronouns and variables. Note that we have already violated the word order guidelines in some of the previous examples — in Latvian, the indefinite pronoun "kaut kas" typically goes before the verb if it is not specified by a relative clause (see the statements 9–10). Thus, formally it belongs to the topic, although it is always new information. But, again, this causes no ambiguities.

In overall, if we are restricting our synthetic CNL to cover terminological statements (class and property definitions) only, information structure analysis is not necessary at all: since OWL axioms are variable-free, any noun phrase that is not explicitly universally quantified is existentially quantified. However, we have now laid the foundations to extend controlled Latvian for support of data integrity constraints.

3. DATA INTEGRITY CONSTRAINTS

In this section we will add some implication rules and data integrity queries to our example ontology, making an actual exploitation of the topic-focus articulation (TFA) — when specifying integrity constraints, one cannot avoid the usage of variables or definite/indefinite NPs.

3.1 SWRL Rules

In SWRL rules [18], variables are used, which cause at least one anaphoric reference, when a rule is verbalized in CNL. In terminological statements, changes in the word order are caused only due to an inverse use of a property, but in the case of rules (verbalized in controlled Latvian), the word order has to be changed also to indicate whether an NP (the subject or the object) introduces a new individual or is an anaphoric reference:

(14) `Rule: Student(?x1), MandatoryCourse(?x2), AcademicProgram(?x3), enrolls(?x3, ?x1), includes(?x3, ?x2) -> takes(?x1, ?x2)`
Every mandatory course that ***is included by an academic program is taken by every student*** *that* ***is enrolled by the academic program****.*
Katru obligāto kursu, ko ***ietver*** *[kāda]* ***akadēmiskā programma, ņem katrs students****, ko [šī] akadēmiskā programma* ***uzņem****.*

In the above statement, inverse properties are used in both relative clauses, causing the swapping of the subject and the object. In the first case, the subject ("akadēmiskā programma") stands to the right from the verb, indicating that it belongs to the focus part — new information. In the second case, the subject (again, "akadēmiskā programma") stands to the left from the verb — in the topic part of the clause, indicating that this is already given information (a reference to the individual introduced in the first relative clause). Thus, in the latter relative clause, the property (verb) is alone in the focus — the new information is the relationship between the two already given individuals (the student and the academic program).

The English and Latvian verbalizations of the above rule are more clearly aligned in Figure 2 (the optional "articles" are not used).

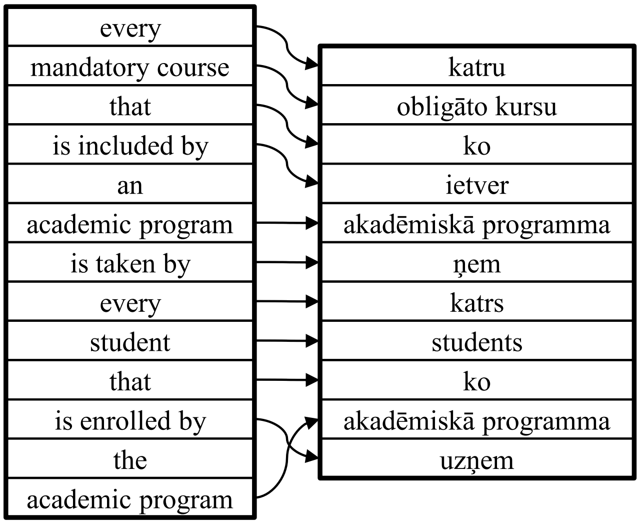

Figure 2. A word alignment graph (generated by the Grammatical Framework [8]), showing that given information in Latvian is reflected by changes in the neutral word order.

To emphasize that an NP is an anaphoric reference, we could optionally use the demonstrative pronoun "šis" ("this"), which usually sounds natural in Latvian. Theoretically, this would allow us to place the NP also to the right from the verb, but both stylistically and intuitively the preferable position is still to the left, which causes the correct intonation.

Let us consider one more rule (Statement 15) where the indefinite pronoun "kāds" is not anymore offered as an optional attribute. It is omitted due to the cascade of relative clauses that modify all the indefinite NPs.

(15)
```
Rule: Person(?x1), MandatoryCourse(?x2),
AcademicProgram(?x3), enrolls(?x3, ?x1),
includes(?x3, ?x2), takes(?x1, ?x2) ->
Student(?x1)
```

*Every person that **takes a mandatory course** that **is included by an academic program** that **enrolls the person is a student**.*

*Katra persona, kas **ņem obligāto kursu**, ko **ietver akadēmiskā programma**, kas [šo] personu **uzņem**, **ir students**.*

Due to the cases when the usage of the indefinite and demonstrative pronouns might improve the readability of a rule, an implementation of an additional concrete grammar of surrogate Latvian, where the usage of indefinite and demonstrative pronouns in the role of articles is always mandatory, would be a simple but naive trade-off. Such a coarse-grained grammar could be used to paraphrase (on demand) the purely TFA-based sentences, both user-provided and auto-generated (verbalized). Although reading of such surrogate statements perhaps is easier than writing, this would be confusing to the end-users anyway (see the next section). To provide more fine-grained paraphrases and to protect the users from a confusingly verbose look-ahead editor, the TFA-based grammar can be improved at least by distinguishing two types of NPs: those that are modified by relative clauses, and those that are not. In the latter case, usage of the indefinite or demonstrative pronoun is preferable.

However, if a statement is written in synthetic language, the best paraphrase could be its translation into analytical language (e.g., English), or alternatively (for an advanced user) — in a human-readable formal syntax like the Manchester OWL Syntax.

3.2 SPARQL Queries

When executing SWRL rules, potentially new facts are inferred and added to the ontology (ABox) whenever the body of a rule is satisfied. If this is the intention and if the rule (its verbalization) does not include negated atoms or disjunctions then it is the right choice; otherwise we would end up with unwanted entailments or would not be able to translate the statement into SWRL. For example, if we would try to redefine the property chaining defined in Statement 11 (in Section 2.2) as a more specific rule (by referring to the concrete classes) — "*Every student that takes a course that is included by an academic program is enrolled by the academic program.*" — the effect would be that a student, taking a course that is included by another academic program, is automatically enrolled by that program.

By asking SPARQL queries we can verify integrity constraints relying on the closed world assumption (negation as failure) [13], without introducing unintended entailments. Thus, Statement 11 can be alternatively (but not equally) specified in CNL as the following query:

(16)
```
ASK WHERE {
  ?x1 rdf:type Student.
  ?x1 takes ?x2.
  ?x2 rdf:type Course.
  ?x3 rdf:type AcademicProgram.
  ?x3 enrolls ?x1.
  NOT EXISTS {?x3 includes ?x2}}
```

Is there a student that takes a course that is not included by an academic program that enrolls the student?

*Vai **ir kāds students**, kas **ņem kursu**, ko **neietver akadēmiskā programma**, kas studentu **uzņem**?*

In the case of consistency checking, the ASK form of a query (yes/no question) is entirely appropriate and its verbalization syntax is not much different from that of rules. Note that in such queries the first NP is always indefinite (although appearing in the topic) and the corresponding indefinite pronoun is always explicitly attached to it.

The current SPARQL specification [16] does not directly provide an operator for negation as failure; it is possible by combining the OPTIONAL, FILTER and !BOUND operators. However, if the !BOUND operator is applied to a variable (in our example, x3) that is used also outside the OPTIONAL block then the result, of course, will not be what expected. Therefore, for the sake of simplicity, we have used the NOT EXISTS pattern that will be provided by the SPARQL 1.1 specification [17].

4. EVALUATION

To verify whether the proposed assumptions on which the proof-of-concept implementation [20] is based are linguistically motivated and universally applicable in highly synthetic CNLs, an initial evaluation was performed. About ten linguists (both Latvians and Lithuanians), specialized in Baltic languages, received nearly twenty examples, covering different types of statements and different levels of complexity. For each example several alternative translations were given in Latvian and Lithuanian in parallel (see Table 1). Among the alternative choices were the "literal" (surrogate) translation, the pure TFA-based translation, and a combination of the previous two, in order to seemingly improve the readability. The respondents were asked to sort the choices (in their native language) by priority from 1 to 3 (1 goes for the best translation of the original statement in ACE), or rejected at all (0). The respondents were introduced with the basic limitations of CNL, but they were also asked to follow their language intuition.

Table 1. Example statements, evaluated by a Lithuanian respondent. The English statement is the benchmark, the Latvian translations (in this case) — for comparison.

*Every student is **a** person that is enrolled by **an** academic program.*		
A_{LV}	*Katrs students ir **kāda** persona, ko uzņem **kāda** akadēmiskā programma.*	
A_{LT}	*Kiekvienas studentas yra **koks nors** asmuo, kurį priima **kokia nors** akademinė programa.*	**0**
B_{LV}	*Katrs students ir persona, ko uzņem akadēmiskā programma.*	
B_{LT}	*Kiekvienas studentas yra asmuo, kurį priima akademinė programa.*	**1**
C_{LV}	*Katrs students ir persona, ko uzņem **kāda** akadēmiskā programma.*	
C_{LT}	*Kiekvienas studentas yra asmuo, kurį priima **kokia nors** akademinė programa.*	**2**

The respondents were also invited to give an alternative translation for each example, if none of the proposed ones was enough satisfactory. This option was used rather frequently, resulting in some interesting suggestions. Those that can be systematized will be taken into account.

In overall, most of the literal translations, using the artificial “articles”, were rejected or assigned with the lowest priority. There was no consensus, however, whether the indefinite and demonstrative pronoun in certain cases should be used to improve readability or not; even the same respondent usually did not act consistently among different examples. However, in most cases, the usage of the pronoun is preferred, if the NP is not modified by a relative clause.

It should be mentioned that almost all the respondents were disappointed with the uniform approach to animate and inanimate things. Although this is not directly related to the topic of this paper, this issue has to be taken into account, which means that one more feature has to be incorporated in the domain-specific lexicons (in the noun and pronoun entries) and exploited in the grammars. However, the issue will still remain, if other tools are used in the workflow (e.g., the ACE verbalizer).

Based on these results, an improved grammar is being developed, which will be evaluated by a wider audience.

5. CONCLUSION

We have shown that 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. This provides a simple and reliable mechanism (guidelines) for deterministic (predictable) analysis of the information structure of a sentence, enabling automatic detection of anaphoric NPs. As the very initial evaluation confirms, native speakers tend to follow such guidelines rather intuitively. Moreover, in languages where the semantic and pragmatic aspects of the sentence are more studied [11], the general correlations between the word order and given/new information are being taught even in language learning courses for beginners [12].

At the time of writing, the proof-of-concept implementation of Latvian-English CNL covers most of the syntactic constructions that were introduced in the previously given examples. The aim for the near future is to extend the TFA-based grammar to cover the full expressivity of terminological statements and rules, while remaining compliant with ACE. Future work is (a) to introduce support for assertional statements — the problem is how to determine, whether the subject noun (in the case of the neutral word order) represents given or new information, (b) to make a more detailed investigation on data integrity queries that are important in practical applications and will make a rather extensive use of anaphoric references, and (c) to consider pros and cons for using the GF Resource Library [8].

6. ACKNOWLEDGMENTS

This research has been supported by the European Social Fund. The author would like to thank Gunta Nešpore, Baiba Saulīte, Guntis Bārzdiņš and Kārlis Čerāns for the valuable discussions on linguistic and semantic aspects, as well as the reviewers for the detailed comments and suggestions, and the respondents for their help in the initial evaluation of the proposed approach.

7. REFERENCES

[1] Angelov, K. Type-Theoretical Bulgarian Grammar. In *6th International Conference on Natural Language Processing* (Gothenburg, Sweden, 2008), LNCS/LNAI 5221, Springer.

[2] Fuchs, N.E., Kaljurand, K., and Schneider, G. Attempto Controlled English Meets the Challenges of Knowledge Representation, Reasoning, Interoperability and User Interfaces. In *19th International FLAIRS Conference* (Melbourne Beach, Florida, 2006), AAAI Press, 664--669.

[3] Hajičová, E. *Issues of Sentence Structure and Discourse Patterns*. Charles University, Prague, 1993.

[4] Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., and Wang, H. The Manchester OWL syntax. In *2nd International OWLED Workshop* (Athens, Georgia, 2006).

[5] Kaljurand, K., and Fuchs, N.E. Verbalizing OWL in Attempto Controlled English. In *3rd International OWLED Workshop* (Innsbruck, Austria, 2007).

[6] Nau, N. *Latvian*. (Languages of the World / Materials 217). Lincom, Munich, 1998.

[7] Pretorius, L., and Schwitter, R. Towards Processable Afrikaans. In *Workshop on Controlled Natural Language* (Marettimo Island, Italy, 2009), CEUR, vol. 448.

[8] Ranta, A., and Angelov, K. Implementing Controlled Languages in GF. In *CNL 2009 Workshop* (Marettimo Island, Italy, 2009), LNCS/LNAI 5972, Springer (to appear).

[9] Saulīte, B. Linguistic Markers of Information Structure in Latvian. In *18th International Congress of Linguists* (Seoul, Korea, 2008), The Linguistic Society of Korea, 3067--3076.

[10] Schwitter, R., Kaljurand, K., Cregan, A., Dolbear, C., and Hart, G. A Comparison of three Controlled Natural Languages for OWL 1.1. In *4th International OWLED Workshop* (Washington, DC, 2008).

[11] Sgall, P., Hajičová, E., and Panevová, J. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

[12] Short, D. *Teach Yourself Czech*. Hodder Education, London, 2003.

[13] Sirin, E., and Tao, J. Towards Integrity Constraints in OWL. In *6th International OWLED Workshop* (Chantilly, Virginia, 2009), CEUR Workshop Proceedings, vol. 529.

[14] Ontology Definition Metamodel. OMG Adopted Specification. http://www.omg.org/spec/ODM/1.0/ (2009)

[15] OWL 2 Web Ontology Language. W3C Recommendation. http://www.w3.org/TR/owl2-primer/ (2009).

[16] SPARQL Query Language for RDF. W3C Recommendation. http://www.w3.org/TR/rdf-sparql-query/ (2008)

[17] SPARQL Query Language 1.1. W3C Working Draft. http://www.w3.org/TR/sparql11-query/ (2010)

[18] SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission. http://www.w3.org/Submission/SWRL/ (2004).

[19] http://protegewiki.stanford.edu/index.php/OWL2UML

[20] http://eksperimenti.ailab.lv/cnl/