Modeling Grammars with Knowledge Representation Methods: Subcategorization as a Test Case

Raúl Aranovich[1][0000-0001-6438-8815]

[1] University of California Davis, One Shields Avenue, Davis, CA 95616
raranovich@ucdavis.edu

Abstract. An OWL ontology is used to model a grammar that accounts for subcategorization, showing that ontologies are able to generate (mildly) context-sensitive languages. Semantic Web knowledge representation methods offer a useful way to model the implicit knowledge that defines human linguistic abilities. When a grammar is modeled as a set of ontological constraints (i.e. classes with restrictions on their properties), ungrammatical sentences are defined as facts that lead to inconsistencies which can be discovered by a reasoner. Property chains are used to "pass on" the category of a syntactic complement as the value of a head's subcategorization feature, modeling the concept of structure sharing that is central to constraint-based theories of syntax like HPSG. By treating utterances as instances and syntactic constraints as axioms, this approach offers points of contact with efforts to model grammars as Linguistic Linked Open Data in the Semantic Web.

Keywords: Syntax, Subcategorization, Property Chain.

1 Introduction

In this paper I will argue that knowledge graphs built with RDF/OWL offer sufficient resources to model a grammar whose strong generative power goes beyond that of context-free grammars. I will focus on the problem of subcategorization in natural languages [1]. This approach points to the usefulness of formal knowledge representation methods for modeling the implicit human knowledge about natural language grammars, and as a testbed for theories of syntax.

Ontologies represent knowledge as a hierarchy of concepts and instances interconnected by relations. Declarative languages like RDF and OWL [2] allow for consistency checks on ontologies by modeling complex logical aspects of knowledge representation, and for the extraction of inferred knowledge. Applications of OWL ontologies to linguistics exists mostly for practical purposes (e.g. domain-specific terminologies, automatic population of ontologies from text), but they can also serve as a tool for theoretical research [3] [4]. More specifically, I am interested in showing that the declarative approach to knowledge representation behind RDF graphs and OWL ontologies provides a fruitful framework to formalize constraint-based approaches to syntax, and to discuss the formal complexity of such grammars. The main insight of this paper is that, when a grammar is modeled as a set of ontological restrictions on admissible structures, ungrammatical sentences can be formalized as sets of syntactic assertions that are in contradiction with the rest of the ontology. By modeling sentence structures as instances, then, syntactic theories can be tested by reasoning over them, since only grammatical sentences will be consistent with the rest of the ontology.

2 Knowledge representation and constraint-based syntax

Following work in constraint-based theories of syntax [5][6][7], I model syntactic categories as classes, and the immediate constituency relations that build phrase structure as relations. A `:headDtr` relation with domain `:Phrase` and range `:Word`, for instance, describes the relation between a phrase and its head, as follows:

```
:Phrase a owl:Class .
:Word a owl:Class ;
   owl:disjointWith :Phrase .
:headDtr a owl:ObjectProperty ;
   rdfs:domain :Phrase ;
   rdfs:range :Word .
```

Likewise, complements are modeled with the object property `:compDtr`. The grammar includes subclasses of words (e.g. `:Noun`, `:Verb`) as well as the corresponding subclasses of phrases (`:NounPhrase`, `:VerbPhrase`, etc). Local constraints on constituency, like the fact that a (transitive) verb phrase is headed by a verb, and has a noun phrase for a complement, are modeled as restrictions.

```
:Verb a owl:Class ;
  rdfs:subClassOf :Word .
:VerbPhrase a owl:Class ;
  rdfs:subClassOf :Phrase ;
  owl:equivalentClass [
    owl:intersectionOf ( [
      a owl:Restriction ;
      owl:onProperty :headDtr ;
      owl:someValuesFrom :Verb ] [
        a owl:Restriction ;
        owl:onProperty :compDtr ;
        owl:someValuesFrom :NounPhrase ] ) ] .
```

A grammar, then, is made of classes representing categories (phrasal or lexical). Well-formedness condition on syntactic structure are represented as restrictions. The actual syntactic structures generated by the grammar are instances of its classes and object properties. Take a sentence like (1), for example:

(1) Pan mocked Hook.

The verb *mocked* and its object form a constituent, which is an instance of the class `:VerbPhrase`. *Mocked* is also an instance, in a `:headDtr` relation with the mother node, while `:Hook` is its `:compDtr`.[1]

```
:Mocked a owl:NamedIndividual , :Verb .
:Hook a owl:NamedIndividual , :NounPhrase .
:Mocked_Hook a owl:NamedIndividual , :VerbPhrase ;
  :headDtr :Mocked ;
  :compDtr :Hook .
```

3 Subcategorization and Structure Sharing in OWL

Adding these assertions to the ontology will not lead to contradiction, since they follow from the class axioms. But the ontology is not yet powerful enough to rule out an ungrammatical sentence like (2) where a verb like *listen* is followed by an NP, not a PP:

(2) Pan listened *(to) Wendy.

A first step is to subcategorize verbs according to the class of their complements, with a property `:complement` with domain `:Word` and range `:Phrase`. The `:Verb` subclass `:TransitiveVerb` would be restricted so that the value of its `:complement` relation had to be an NP.[2] Likewise, `:PrepositionalVerb` subcategorizes for a PP. The second step is to come up with an implementation that will define the value of`:compDtr` to match the restriction on the verb. One way to achieve that is to define the relation `:complement` as a **property chain**, so that the instance that occurs as the syntactic complement in the VP is passed on to the verb's `:complement` value. The chain links the inverse of the `:headDtr` relation (getting to the mother VP from the V) and the `:compDtr` relation (getting from the VP to the complement NP).

```
:complement owl:propertyChainAxiom (
     [ owl:inverseOf :headDtr ]  :compDtr )
```

This is how the notion of **structure-sharing**, central to constraint-based theories like HPSG, can be implemented in OWL. If a sentence has a syntactic VP complement that does not match the restriction on the verb's subclass, as in (2), then the ontology becomes inconsistent. This result has been

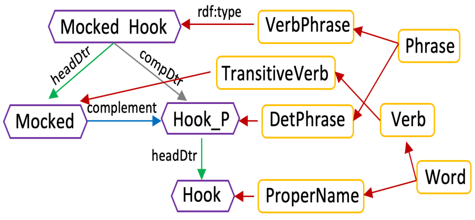

Figure 1: Ontology schematic for "mocked Hook"

[1] As instances, constituents need to be given unique identifiers, like "Mocked_Hook", and not generic class names like VP.

[2] This is similar to the use of syntactic frames in LexInfo [8].

confirmed in an ontology designed with the help of the Protégé editor.[3]

4 Consequences and conclusions

Syntactic theory has an important role to play in the development of the Semantic Web. Automatic sentence parsing, for instance, is a component of systems that allow users to access semantic content through natural language queries, which must be converted into formal SPARQL queries [9]. But these approaches use syntactic tools that are external to the ontology itself, and are usually procedural. My proposal, by contrast, develops a declarative approach to sentence structure which is built using native OWL constructs, and is formalized as an ontology.

Moreover, there are ontologies of linguistics (e.g. GOLD [10], LexInfo [8]), but the purpose of these is mainly to define the concepts that linguists use in their discipline, with the associated terminology, rather than as a generative model (i.e. a system that defines the set of grammatical sentences of a language with their associated structures). Here is where an important ontological difference with the current proposal stems from. While models like GOLD or LexInfo treat parts of speech and other linguistic categories as instances, I treat them as classes. In my system, the only instances are concrete utterances, with their latent structure.[4] That is because my goal is to formalize the *implicit* knowledge that a speaker has of their language (the Chomskyan notion of *competence*, if you may), while other ontologies formalize the *explicit* knowledge that a linguist has about their discipline. To the extent that I use that explicit knowledge to model the implicit knowledge, there should be a point of contact between the approaches.

The approach sketched here is not intended to compete with statistical models of language in terms of scale and empirical coverage. Rather, it offers a method to model the constructs that syntactic theory has proposed to account for sentence structure. There are two directions in which this method can be extended. First, there are other verb classes besides transitive verbs that should be modeled with similar tools: intransitives (*glitter*, *work*), ditransitives (*give*, *tell*), prepositional (*rely on*), verbs with sentential complements (*hope*, *think*), etc. Each of these classes is defined by a different restriction. Second, there are different contexts in which a verb may or may not appear (including verbs with variable behavior). In this paper I have dealt with the problem of a verb with a complement of the wrong class (a PP instead of an NP). But a sentence may also be ungrammatical if a verb has fewer complements than it requires (e.g. a transitive verb with no complements) or more than it requires (an intransitive verb with a complement of any kind). Working out those aspects of the problem should be the matter of future work.

What I have shown is that an OWL ontology can be used to model a grammar that goes beyond simple context free generation to account for strict subcategorization. This result is important, in that it shows that ontologies can be used to generate context-

[3] https://github.com/RaulAranovich/OnSyDE/blob/main/OnSyDE.owl

[4] My treatment of individual utterances as instances is similar to efforts to serialize syntactically-annotated corpora as RDF documents, sharable as Linguistic Linked Data [11, 12].

sensitive languages. The knowledge representation methods that have been developed for the Semantic Web may prove to be a useful tool to model syntactic competence, understood as the implicit knowledge that defines the human linguistic abilities.

References

1. Joshi, A. K., Shanker, K. V., Weir, D.: The Convergence of Mildly Context-Sensitive Grammar Formalisms. University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-90-01 (1990).
2. Hitzler, P.: A review of the Semantic Web Field. Communications of the ACM 64(2), 76-83 (2021).
3. Cimiano, P., Unger, C., McCrae, J.: Ontology-Based Interpretation of Natural Language. Morgan & Claypool, San Rafael (2014).
4. Schalley, A. C.: Ontologies and ontological methods in linguistics. Language and Linguist Compass 13(11) (2019). DOI: 10.1111/lnc3.12356.
5. Pollard, C., Sag, I.: Head-Driven Phrase Structure Grammar. U. of Chicago Press, Chicago (1994).
6. Copestake, A.: Implementing Typed Feature Structure Grammars. CSLI Publications, Stanford (2002).
7. Francez, N., Wintner, S.: Unification Grammars. Cambridge U. Press, Cambridge (2012).
8. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: LexInfo: A declarative model for the lexicon-ontology interface. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 9(1), 29–51 (2011).
9. Unger, C., Cimiano, P.: Pythia: Compositional Meaning Construction for Ontology-Based Question Answering on the Semantic Web. In: R. Muñoz et al. (Eds.), Proceedings of NLDB 2011, pp. 153-160. Springer-Verlag, Berlin/Heidelberg (2011).
10. Farrar, S., Lewis, W.D.: The GOLD Community of Practice: An Infrastructure for Linguistic Data on the Web. Language Resources and Evaluation, 41(1), 45–60 (2007).
11. Chiarcos, C., Glaser, L: A Tree Extension for CoNLL-RDF. In: Proceedings of the 12th Conference on Language Resources and Evaluation, pp. 7161–7169. ELRA (2020).
12. Chiarcos, C.: POWLA: Modeling linguistic corpora in OWL/DL. In: 9th Extended Semantic Web Conference, pp. 225–239. Heraklion (2012).