Representing Multiword Expressions on the Web with the OntoLex-Lemon model

John P. McCrae[*] Philipp Cimiano[†] Paul Buitelaar[*]
Georgeta Bordea[*]

[*]Insight Centre for Data Analytics, National University of Ireland, Galway
[†]Cognitive Interaction Technology Centre of Excellence, Bielefeld University

1 Introduction

Lexical resources are increasingly published on the Web and they are linked to semantic information contained in ontologies. This trend has lead to the development of several new models for the representation of lexical resources notably the *lemon* model [2], which since 2012 has been further developed by the W3C OntoLex community group[1]. In this paper we discuss the developments of this model in particular with respect to how the semantics of multi-word expressions are encoded and give practical examples of the modelling.

2 The OntoLex-Lemon Model

2.1 Overview

The OntoLex-Lemon model has been developed based on the *lemon* model [2] and aims to represent a lexicon relative to an ontology expressed in a language such as OWL [4]. An OntoLex-Lemon lexicon is centered around a *lexical entry*, which represents all morphological and semantic variants that can be represented by a term, which each meaning associated to a *lexical sense*. The *lexical entry* is further divided into the classes: *word*, *multiword expression* and *affix*. The entry is linked to a concept in the ontology, which may be a *class*, *property*, *individual* or an *event class*[3]. The lexical entry may be associated with a number of forms, which give alternative grammatical forms of a word.

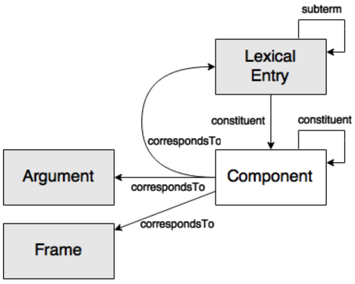

Figure 1: The OntoLex-Lemon Decomposition Module

2.2 Decomposition

The OntoLex-Lemon model (Figure 1) models decomposition principally by means of the *Component* element which must uniquely *correspond to* a lexical entry, frame or an argument. Each lexical entry then has a number of *constituents* which indicate the elements of the entry. It is assumed that a lexical entry only has a single decomposition. In case of a multiword expression, the decomposition indicates the tokenization, while for the case of a compound noun it indicates the actual decomposition into sub-word units. Components may be annotated with morphosyntactic annotations that indicate the inflectional form used, for example in the Irish term "Poblacht na hÉireann" (Republic of Ireland) the third word can be marked as being the genitive, lenited form of the lexical entry "Éire" (Ireland). The order of the words can be indicated by using the `rdf:_n` properties and this was chosen over the use of a linked list based mechanism as it makes querying of the data using SPARQL [5] and similar methods easier. In addition, for lexicons that do not need to represent individual components it is possible to use a property called *subterm* which is equivalent to indicating that there is a component. An example[2] of this is as follows:

```
:PoblachtNahEireann a ontolex:MultiwordExpression ;
  rdfs:label "Poblacht na hÉireann"@ga ;
  decomp:constituent :poblacht, :na, :hEireann ;
```

[1] http://www.w3.org/community/ontolex/

[2] More examples are available at http://cimiano.github.io/ontolex/specification.html

```
  rdf:_1 :poblachtr; rdf:_2 :na; rdf:_3 :hEireann ;
  ontolex:denotes dbpedia:Republic_of_Ireland .

:hEireann a decomp:Component ;
  lexinfo:case lexinfo:genitive ;
  lexinfo:lenition true ;
  decomp:correspondsTo :Eire .
```

2.3 Linking decomposition to semantics

Besides representing simple term decomposition, the OntoLex-Lemon model allows one to associate subunits of a decomposition to an element in an ontology. In Ontolex-Lemon only the arguments in the ontology are modelled, thus more a phrase like "kick the bucket" is modelled as an intransitive verb that categorizes for subject. Take the example of the term "give up", for which we might want to specify that it denotes the class of `QuittingEvents` (see Figure 2). In order to do this, we would say that the entry has the constituents 'give' and 'up', each of which is associated with a lexical entry for the word, and that these entries are also associated with a syntactic frame that is attached to the entry. This frame is composed of not only the words but also of two arguments which are mapped via atomic senses[3] to roles in the ontology. The lexical entry can thus be said to *denote* the event class maintaining a separation between the syntactic object (the words) and the semantic object.

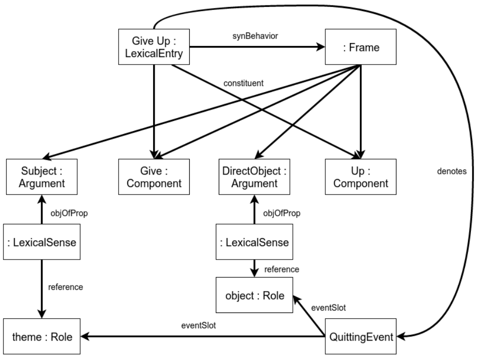

Figure 2: An example of modelling a complex multiword expression in a lexicon

3 Conclusion

As shown in this paper the OntoLex-Lemon model has a number of advantages that follow from the use of a flexible data model such as RDF and the ability to link to a sophisticated logical representations of concepts such as those in OWL[1]. In particular the graph-based structure allows us to represent the links within a complex lexical entry efficiently, while still clearly and unambiguously grounding it in an ontological basis, thus ensuring that further reasoning can be applied for applications such as question answering.

References

[1] Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. Towards open data for linguistics: Lexical Linked Data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. 2013.

[2] John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709, 2012.

[3] John P. McCrae, Christina Unger, Francesca Quattri, and Philipp Cimiano. Modelling the Semantics of Adjectives in the Ontology-Lexicon Interface. In *Proceedings of 4th Workshop on Cognitive Aspects of the Lexicon*, 2014.

[4] Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. W3C Recommendation, World Wide Web Consortium, 2014.

[5] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C Recommendation, World Wide Web Consortium, 2008.

[3] Due to the limitation of OWL to predicates of only two arguments, entries are modelled as compound senses with many atomic senses corresponding to properties in the ontology