Publishing and Linking WordNet using lemon and RDF

John P. M^{c}Crae[*], Christiane Fellbaum[†] and Philipp Cimiano[*]

[*]University Bielefeld, [†]Princeton University
CITEC Building, Inspiration 1, 33615 Bielefeld; 35 Olden Street, Princeton
{jmccrae, cimiano}@cit-ec.uni-bielefeld.de, fellbaum@princeton.edu

Abstract

Each article must include an abstract of 150 to 200 words in Times 9 pt with interlinear spacing of 10 pt. The heading Abstract should be centred, font Times 10 bold. This short abstract will also be used for printing a Booklet of Abstracts containing the abstracts of all papers presented at the Conference.

Keywords: keyword A, keyword B, keyword C

1. Introduction

WordNet is one the first and still most widely used resources for natural language processing. In the time since the first version of WordNet was released many resources have been produced that represent complementary information(**?**; **?**) or extend the WordNet model to new languages(**?**; **?**). In contrast, in recent years we have seen the development of Web Technologies for the representation of language resources and in particular, the use of linked data. This has lead to a linguistic linked open data cloud, which is constructed by linking resources and publishing them on the web using RDF. Linked data, as proposed by Berners-Lee(**?**), has four main principles for publishing data: firstly, the use of URIs to identify objects; secondly, that these URIs should be resolvable; thirdly, that semantic information is returned, using standards such as RDF and finally, that links are provided to other resources. Chiarcos *et al.*(**?**) discuss the application of this to linguistic data and show that this model has notable advantages over standard approaches to data modelling, in particular they outline the following:

1. Representation: Graph-based models are a method that can represent any form of language resource.
2. Structural interoperability: By using RDF graphs and URIs datasets can be merged with no effort.
3. Federation: Multiple datasets can easily be drawn from different sources in the web and used together seamlessly.
4. Conceptual interoperability: Linking to common data category repositories allows common definitions to be inferred.
5. Ecosystem: Building on standards such as RDF, allows the use of common tools, including databases.
6. Dynamic import: Data on the web is not static and as such errors can be corrected after publication.
7. Expressivity: The use of other Semantic Web models allows the easy expression of metadata, provenance and ontological constraints on the data

In this paper we describe our experience in publishing WordNet following the linked data principles. While this is not the first version of WordNet to be published as linked data (**?**; **?**; **?**), this version has several advantages, firstly that it is well-linked to many resources, secondly it uses an open model in *lemon* and finally, that as it is integrated with the development of WordNet, and as such will be updated alongside future releases of WordNet.

2. Background

2.1. WordNet

Introduce descriptio of WordN

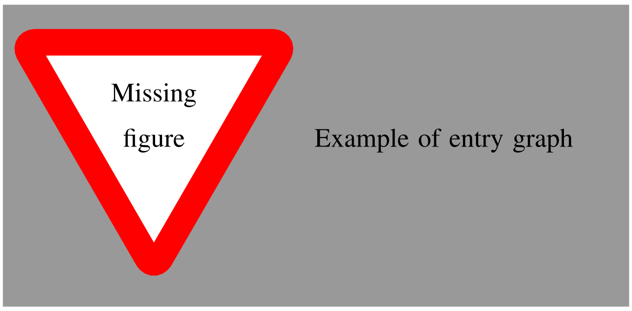
2.2. lemon

lemon is a model that has been proposed() for the representation of lexicons relative to ontologies. As such, this model is well suited to the representation of semantic networks such as WordNet and defines many useful features for linking a WordNet to wider objects in the Semantic Web. *lemon* models lexicons by means of a core consisting of the following elements:

- A *lexical entry* which represents a single word or multi-word unit.
- A *lexical sense*, representing a meaning of that word, which contains a *reference* to a concept in the ontology.
- *Forms*, which are inflected version of the entry, and associated with a string *representation*.

2.3. Linguistic Linked Data

The application of linked data technology to the representation of linguistics resources has been spearheaded by the OKFN Working Group on Linguistics(), and they have been mapping the progress of this project by means of a cloud diagram showing how all the existing resources are linked (figure **??**). As outlined above, there are many key advantages to the use of this technology for language resources, however the cloud has until now lacked a central node. As WordNet is the most widely referenced language resource we believe that WordNet can act as a nucleus for linguistic linked data in the way that DBpedia has for the wider cloud.



	Number of triples
Links to VerbNet	?
Links to DBpedia	?
Links to LexVo	?
Links to lemonUBY	?
Other	?
Total	?

3. Representing WordNets with lemon

As *lemon* is a model for representing lexica relative to ontologies it is not obviously clear how it can be used to represent a WordNet. It is clear that the words of WordNet can be called lexical entries and the word senses correspond well to the concept of lexical senses. WordNet has lemmas and a separate list of variants of these, and as such we create a canonical form for each lemma and an other form for each of these variants. As there is currently no indication in WordNet of what grammatical properties these variants have we do not distinguish these forms by means of annotation. As a *lemon* is a model for ontology-lexica, the main question is what the reference of the sense should be. We approach this by saying that the synsets of WordNet are the ontological references, but instead of assigning them a formal ontological type (e.g., class, property or indiviual) we instead use the SKOS() vocabulary and type them as concepts. This allows us to capture the nature of synsets without ontologizing the semantic network as in ().

The other key question is the identifiers we use for each element in the data. We do not follow previous exports such as () in assigning new identifiers but instead attempt to use the existing identifiers in WordNet. Furthermore, as WordNet has released several versions and is still under development, we feel it is important to include the version number in the URI. As such, we use the following scheme for URIs:

- Each lexical entry is represented by means of the URL-encoded lemma and then a dash followed by the part of speech as a single letter (i.e., 'n(oun)', 'v(erb)', 'a(djective)', 'r (adverb)', 's(entence)' or 'p(article)').
- Senses and forms in the model use the entry URI and add a fragment identifier. For forms, as there is no previous identifier in WordNet we simply use `CanonicalForm` and `Form-n` where `n` is a number. For senses, the fragment is simply the sense identifier from WordNet
- Synsets are similarly the 8 figure 'offset' code from the WordNet database, followed by a dash and the part of speech as a single letter.

As such an example of the URI scheme is shown in figure **??**.

A Python framework based on RDFlib() is used to serve the website and provide SPARQL access to the data.

4. Linking WordNet

In addition to providing a RDF version of WordNet we also incorporated a number of extra resources founded from other sources into the RDF data. In particular we include the following elements

- For verbs, where extant we include mappings to VerbNet(). As VerbNet does not currently have a linked data version, we simply link to the PHP page of the web site.
- We include translations from the Open Multilingual WordNet() collection as simple labels on the synsets, identified by the use of language codes.
- We have mapped previous mappings to LexVo() and DBpedia() to use the current identifiers in WordNet.
- We include links to the W3C WordNet 2.0 export().
- We have created new links to the lemonUby() resource.

In addition to these links we also provide support for legacy resources by adding URL mappings from previous versions of WordNet identifiers to the most recent version. These mappings are based on the mappings of ()

5. Related Work

This work does not represent the first version of WordNet made available in RDF, in fact (Van Assem et al., 2006; McCrae et al., 2012) have made previous version available directly in WordNet. Furthermore, WordNet has been incorporated into various larger resources including BabelNet (Navigli and Ponzetto, 2010; **?**), UBY (Gurevych et al., 2012; Eckle-Kohler et al., 2014). These projects however have mostly been fixed to using a single version of WordNet. In constrast, we view our work as more related to task of providing universal identifiers for words as in the work of the Global WordNet Grid (Pease et al., 2008).

6. Conclusion

In this paper we have presented the creation of a WordNet RDF version that is linked with the development of the existing model as well as incorporates a large number of links to other resources on the web. As such we believe that this node will constitute a key central node for the expansion of the linguistic linked open data cloud

7. References

Judith Eckle-Kohler, John McCrae, and Christian Chiarcos. 2014. lemonuby-a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, submitted. special issue on Multilingual Linked Open Data.*

```
http://wordnet-rdf.princeton.edu/wn31/cat-n
http://wordnet-rdf.princeton.edu/wn31/cat-n#CanonicalForm
http://wordnet-rdf.princeton.edu/wn31/cat-n#???
http://wordnet-rdf.princeton.edu/wn31/00001740-a
```

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M Meyer, and Christian Wirth. 2012. Uby: A large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.

John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. 2012. Integrating wordnet and wiktionary with lemon. In *Linked Data in Linguistics*, pages 25–34. Springer.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Adam Pease, Christiane Fellbaum, and Piek Vossen. 2008. Building the global wordnet grid. *CIL18*.

Mark Van Assem, Aldo Gangemi, and Guus Schreiber. 2006. Conversion of wordnet to a standard rdf/owl representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC06), Genoa, Italy*.