Publishing and Linking WordNet using Lemon and RDF

John P. McCrae*, Christiane Fellbaum† and Philipp Cimiano*

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

Abstract

In this paper we provide a description of a dataset consisting of data from the Princeton WordNet. This version is intended to provide canonical URIs that can be used by a wide variety of lexical resources to express their linking as part of the Linguistic Linked Open Data Cloud. Furthermore, this is the first version to use the *lemon* model and we describe how we represent WordNet with this model.

Keywords: dataset description, WordNet, linked data

1. Introduction

WordNet is one of the oldest lexical resource for natural language processing and it continues to be widely used. In the time since the first version of WordNet was released many resources have been produced that represent complementary information (Schuler, 2005; Baker et al., 1998) or extend the WordNet model to new languages (Vossen, 1998; Bond and Foster, 2013). Meanwhile, in recent years we have seen the development of Web technologies for the representation of language resources and in particular, by the use of linked data. This has lead to the 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, 2011), has four main principles for publishing data: firstly, that it use 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. Finally, recent activity in the context of the ontology-lexicon interface has led to the creation of a number of models, most notably *lemon* (McCrae et al., 2012a), which give a principled method of relating words to ontological concepts. Given the wide popularity of WordNets in many languages it is important to clarify how WordNets can be integrated in the context of an ontology-lexicon model.

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 (Van Assem et al., 2006; McCrae et al., 2012b; Graves and Gutierrez, 2006), our version has several advantages: firstly that it is well-linked to many resources, secondly that 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

WordNet (Miller, 1995; Fellbaum, 1998; Fellbaum, 2010) is a large lexical database of English nouns, verbs, adjectives and adverbs. Word forms are grouped into more than 117,000 sets of roughly synonymous word forms ("synsets"). These are interconnected by bidirectional arcs that stand for lexical (word-word) and semantic (synset-synset) relations, including hyper/hyponymy (*tree-oak*), meronymy (*tree-branch*), antonymy (*long-short*) and various entailment relations (*buy-pay, show-see, untie- tie*).

WordNet's synsets and its network structure yield a rough measure of semantic similarity among words and concepts in terms of synset membership as well as the number of arcs separating synsets. Consequently, WordNet has become a popular tool for Word Sense Disambiguation (WSD) and Natural Language Processing in general. WordNets have been built for some 100 different languages. Most are mapped onto the Princeton WordNet, enabling translation on the lexical level as well as cross-lingual WSD and applications. WordNet continues to evolve both in terms of coverage and representation of meaning. Recent enhancements include the addition of internet language and partially compositional multiwordunits. Finally, WordNet has been mapped to formal ontologies, including SUMO (Niles and Pease, 2003) and KYOTO (Vossen et al., 2011).

2.2. lemon

lemon is a model that has been proposed (McCrae et al., 2012a) 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*.

In fact, in previous work (Eckle-Kohler et al., 2014) *lemon* has been used not only to represent WordNet but to integrate it with more syntactically sophisticated resources such as VerbNet. As such *lemon* shows potential to help in

	Number of triples
Links to VerbNet	26,353
Links to LexVo	458,907
Links to lemonUBY	475,502
Links to W3C WordNet	99,926
Total	8,903,345

Table 1: The number of links and total number of triples in WordNet-RDF

the integration of lexical data across many levels and languages.

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 (Chiarcos et al., 2011), who has been mapping the progress of this project by means of a cloud diagram showing how all the existing resources are linked. 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 (Auer et al., 2007) has for the wider cloud.
(Chiarcos et al., 2013) discuss the application to linguistic data and show that this model has notable advantages over standard approaches to data modelling, in particular 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.

3. Representing WordNets with lemon

It is not trivial to apply *lemon* to the case of a WordNet as there is no clear ontology, so we describe our mapping here. Clearly, WordNet's words 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

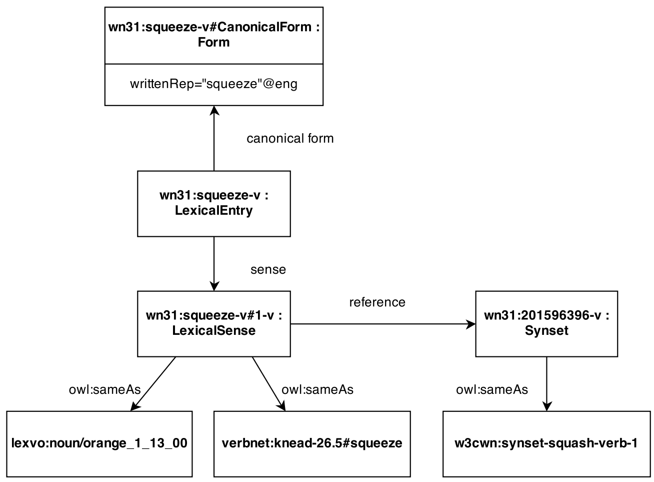

Figure 1: An example of the modelling a single word and synset and links to other resources

for each lemma and another form for each of these variants. Since 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 *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 individual) we introduce a new type `Synset`, which is typed as a `Concept` from the SKOS (Miles and Pérez-Agüera, 2007) vocabulary. This allows us to capture the nature of synsets without ontologizing the semantic network as in (Gangemi et al., 2003). Similarly, we introduce relations such as hypernymy, meronymy etc. as new properties rather than attempt to relate them to existing ontological properties such as OWL's sub class of. In order to capture the new properties we introduce an ontology[1] describing the new properties and classes and providing axioms for the use in the context of both *lemon* and SKOS. Furthermore, we link the elements in the ontology to data categories from ISOcat (Kemps-Snijders et al., 2008) following the guidelines of (Windhouwer and Wright, 2012).
Another key question concerns the identifiers we use for each element in the data. We do not follow previous exports such as (Van Assem et al., 2006) 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 consider it important to include the version number in the URI. As such, we use the following scheme for URIs, as exemplified in figure 2:

- 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)', 'adjective s(atellite)' or 'p(article)').

[1] `http://wordnet-rdf.princeton.edu/ontology`

```
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#2-n
http://wordnet-rdf.princeton.edu/wn31/00001740-a
```

Figure 2: URI scheme of RDF WordNet

- Senses and forms in the model use the entry URI and add a fragment identifier. For forms, for which there is no previous identifier in WordNet, we use `CanonicalForm` and `Form-n` where `n` is a number. For senses, the fragment is the index of the senses and the part of speech.

- Synsets are similarly the 8 or 9 figure[2] 'offset' code from the WordNet database, followed by a dash and the part of speech as a single letter.

A Python framework based on RDFlib[3] 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 incorporated a number of extra resources founded from other sources into the RDF data. In particular we include the following elements:

- For verbs, we include mappings to VerbNet (Schuler, 2005) where extant. As VerbNet does not currently have a linked data version, we link to the PHP page of the web site.

- We include translations from the Open Multilingual WordNet (Bond and Foster, 2013) collection as simple labels on the synsets, identified by the use of language codes.

- We have mapped previous mappings to LexVo (De Melo and Weikum, 2008) using the current identifiers in WordNet.

- We include links to the W3C WordNet 2.0 export (Van Assem et al., 2006).

- We have created new links to the lemonUby (Eckle-Kohler et al., 2014) resource.

In addition to these links we provide support for legacy resources by adding URL mappings from previous versions of WordNet identifiers to the most recent version, with mappings based on (Daudé et al., 2000). The number of triples in the resource can be seen in table 1 and an example of the data is illustrated in figure 1.

[2]The 9 figure codes include an extra initial digit for part-of-speech

[3]`https://github.com/RDFLib/rdflib`

5. Related Work

This work does not represent the first version of WordNet made available in RDF; (Van Assem et al., 2006; McCrae et al., 2012b) have made previous version available directly in WordNet. Furthermore, WordNet has been incorporated into various larger resources including BabelNet (Navigli and Ponzetto, 2010; Ehrmann et al., 2014) and 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 contrast, we view our work as more related to task of providing universal identifiers for words as in the ongoing work of the Global WordNet Grid (Pease et al., 2008).
There have been a number of attempts to interlink WordNets. (Pease and Fellbaum, 2009) used an upper-level ontology called SUMO to group WordNet concepts and these mappings are adopted by several language versions of WordNet. Similarly, attempts were made to integrate WordNets based around the Kyoto Ontology and LMF (Soria et al., 2009). Finally, the SemLink project (Palmer, 2009) has developed links among several resources, though it is not yet integrated into the linked data cloud.

6. Conclusion

We presented the creation of a WordNet RDF version that is linked with the development of the existing model and incorporates a large number of links to other resources on the web. Furthermore, we discussed the use of the *lemon* model to describe a WordNet and ameliorate the integration of not just other WordNets but also a wide variety of lexical resources that are integrated with WordNet. We believe that our WordNet RDF model will constitute a key central node for the expansion of the Linguistic Linked Open Data Cloud.

7. References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Tim Berners-Lee. 2011. Linked data-design issues (2006).

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, 52(3):245–275.

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

Jordi Daudé, Lluis Padro, and German Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 504–511. Association for Computational Linguistics.

Gerard De Melo and Gerhard Weikum. 2008. Language as a foundation of the semantic web. In *International Semantic Web Conference (Posters & Demos)*.

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*.

Maud Ehrmann, Danielle Vannela, John P. McCrae, Francesco Cecconi, Philipp Cimiano, and Roberto Navigli. 2014. Representing multilingual data as linked data: the case of babelnet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-14)*.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database.

Christiane Fellbaum. 2010. *WordNet*. Springer.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *On the move to meaningful internet systems 2003: CoopIS, DOA, and ODBASE*, pages 820–838. Springer.

Alvaro Graves and Caludio Gutierrez. 2006. Data representations for wordnet: A case for rdf. In *GWC 2006–Proceedings of the 3rd International WORDNET Conference*, pages 165–169. Citeseer.

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.

Marc Kemps-Snijders, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. 2008. Isocat: Corralling data categories in the wild. In *LREC*.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2012a. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

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

Alistair Miles and José R Pérez-Agüera. 2007. Skos: Simple knowledge organisation for the web. *Cataloging & Classification Quarterly*, 43(3-4):69–83.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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.

Ian Niles and A Pease. 2003. Mapping wordnet to the sumo ontology. Technical report, Teknowledge Technical Report.

Martha Palmer. 2009. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15.

Adam Pease and Christiane Fellbaum. 2009. Formal ontology as interlingua: The sumo and wordnet linking project and globalwordnet. *Ontology and the lexicon. A natural language processing perspective*, pages 31–45.

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

Karin Kipper Schuler. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-lmf: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 international workshop on Intercultural collaboration*, pages 139–146. ACM.

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*.

Piek Vossen, Eneko Agirre, Francis Bond, Wauter Bosma, C Fellbaum, Amanda Hicks, S Hsieh, Hitoshi Isahara, Ch Huang, Kyoko Kanzaki, et al. 2011. Kyoto: a wiki for establishing semantic interoperability for knowledge sharing across languages and cultures. *Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. Ed. Dr. E. Blanchard (Mc Gill University (Canada) and Dr. D. Allard (Dalhousie University), IGI Global USA*.

Piek Vossen. 1998. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Boston.

Menzo Windhouwer and Sue Ellen Wright. 2012. Linking to linguistic data categories in isocat. In *Linked Data in Linguistics*, pages 99–107. Springer.