Publishing and Linking WordNet using lemon and RDF

John P. McCrae[◇], Christiane Fellbaum[♣] and Philipp Cimiano[◇]

[◇]University Bielefeld, CITEC Building, Inspiration 1, Bielefeld, Germany
[♣]Princeton University, 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, language resources, lexicon, ontology, WordNet, linked data

1. Introduction

WordNet is still one of the most widely used lexical resources within natural language processing. From the time since the first version of WordNet was released, many resources have been produced that represent complementary information to WordNet (**?**; **?**) or extend it to other languages (**?**; **?**). Meanwhile, new Web technologies, in particular Linked Data, have fostered the publication of data on the Web. This has lead to the emergence of the so called *Linguistic Linked Open Data (LLOD)* cloud (**?**) in which resources and datasets are represented in RDF format and linked to each other. Linked data, as proposed by (**?**), has four main principles for publishing data: firstly, that it relies on 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. Recent activity in the context of the ontology-lexicon interface has led to the creation of a number of models, most notably *lemon* (**?**), providing a principled method for publishing lexical data in combination with ontologies that define the semantics of the lexical entries. Given the popularity and availability of WordNets in many languages, it is important to clarify how WordNets can be integrated into the Linked Data cloud.

In this paper we describe our experience in publishing WordNet following the linked data principles and using the *lemon* model. While this is not the first version of WordNet to be published as linked data (**?**; **?**; **?**), our version has several advantages: firstly that it is linked to many resources, secondly that it uses an open model, *lemon*, and, most importantly, that it is directly hosted by Princeton and will thus directly be automatically in sync with further releases of WordNet. As such we intend that this data will provide a central hub for the linkage of not only many WordNets in many languages, thus enabling the vision of a Global WordNet Grid (**?**), but also for many other kinds of lexical resources.

2. Background

2.1. WordNet

WordNet (**?**; **?**; **?**) 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, so called *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. Due to its availability under open licenses, WordNet has become a popular tool for Word Sense Disambiguation (WSD) and Natural Language Processing in general. WordNets have been built for around 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 multi word units. Finally, WordNet has been mapped to formal ontologies, including SUMO (**?**) and KYOTO (**?**).

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/Linked Open Data Cloud. *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 versions of the entry, and associated with a string *representation*.

In fact, in previous work (**?**) *lemon* has been used not only to represent WordNet but to integrate it with more syntac-

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

tically sophisticated resources such as VerbNet. As such *lemon* shows potential to help in the integration of lexical data across many levels and languages. The *lemon* model is highly compatible with the ISO standard LMF (**?**) and forms the basis of the work of the W3C OntoLex Community Group[1].

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 (**?**), who has been documenting the progress of this project by means of a cloud diagram showing all the linguistic resources available on Linguistic Linked Open Data Cloud as well as the links between these.

There are many key advantages to the use of this technology for language resources. However, the cloud has until now lacked a central point of reference. 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.

Chiarcos et al. (**?**) have discussed the application to linguistic data and argue 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 from a syntactic point of view straightforwardly.
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 supports interoperability.
5. Ecosystem: Building on standards such as RDF supports the use of common tools, including databases.
6. Dynamic publishing: Data on the web is not static, such that different versions can be exposed and errors can be corrected.
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 in WordNet. Clearly, WordNet's words can be regarded as *lemon* lexical entries and the word senses correspond well to *lemon*'s 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 a *Form* object for each of these variants. Since there is currently no indication in WordNet of what grammatical properties these variants have, we do not attach additional properties to these variants/forms. As *lemon* is a model for ontology-lexica, the main question is what the reference of the lexical senses should be. We choose to regard WordNet's synsets as ontological references, but instead of assigning them a formal ontological type (e.g., class, property or individual), we introduce a new type `Synset` as a subclass of *Concept* in SKOS (**?**).

This allows us to capture the nature of synsets without ontologizing the semantic network as in (**?**). 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 subClassOf. In order to capture the new properties, we introduce an ontology[2] describing the new properties and classes and provide axioms for the use in the context of both *lemon* and SKOS. These axioms including stating transitivity constraints and equivalence to other vocabularies, e.g., WordNet's *hypernym* to SKOS's *broader*. Furthermore, we link the elements in the ontology to data categories from ISOcat (**?**) following the guidelines of (**?**).

Another key question concerns 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 consider it important to include the version number in the URI. As such, we use the following scheme for URIs:, as exemplified below:

- 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)'), e.g. `http://wordnet-rdf.princeton.edu/wn31/cat-n`.
- 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

	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

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

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

part of speech, e.g. `http://wordnet-rdf.princeton.edu/wn31/cat-n#2-n`.

- Synsets are similarly are identified by a number consisting of 8 or 9 digits corresponding to offset codes in the WordNet database[3], followed by a dash and the part of speech as a single letter, e.g. `http://wordnet-rdf.princeton.edu/wn31/300001740-a`.

A Python framework based on RDFlib[4] is used to serve the website and provide SPARQL access to the data.

4. Linking WordNet

In addition to providing a RDF/Linked Data version of WordNet, we have incorporated a number of links to other resources. In particular we include the following elements:

- For verbs, we include mappings to VerbNet (**?**) if they exist. As VerbNet does not currently have a linked data version, we link to the PHP page of the web site.
- We include translations from Open Multilingual WordNet (**?**) as simple labels on the synsets, identified by the use of language codes.
- We have included previous mappings to LexVo (**?**) using the current identifiers in WordNet.
- We include links to the W3C WordNet 2.0 export (**?**).
- We have created new links to lemonUby (**?**).

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 (**?**). The number of triples in the resource and an example can be seen in table 1 and figure 1. We intend to continue to expand this linksets with contributions from the community.

5. Related Work

This work does not represent the first version of WordNet published in RDF. Previous versions include the one by van Assem et al. (**?**) as well as McCrae et al. (**?**). Furthermore, WordNet has been incorporated into various larger resources including BabelNet (**?**; **?**) and UBY (**?**; **?**). These projects however have mostly been fixed to using a single version of WordNet. In contrast, we view our work as more related to the task of providing universal identifiers for words as in the ongoing work of the Global WordNet Grid (**?**).

There have been a number of attempts to interlink WordNets. Pease and Fellbaum (**?**) 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 (**?**). Finally, the SemLink project (**?**) has developed links among several resources, though it is not yet integrated into the linked data cloud.

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

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

6. Conclusion

We have described a Linked Data version of Princeton WordNet that is expressed in RDF and linked to other relevant resources and is directly sychronized with the WordNet project. Furthermore, we have 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.