Integrating WordNet and Wiktionary with *lemon*

John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano

Abstract Nowadays, there is a significant quantity of linguistic data available on the Web. However, linguistic resources are often published using proprietary formats and, as such, it can be difficult to interface with one another and they end up confined in "data silos". The creation of web standards for the publishing of data on the Web and projects to create Linked Data have lead to interest in the creation of resources that can be published using Web principles. One of the most important aspects of "Lexical Linked Data" is the sharing of lexica and machine readable dictionaries. It is for this reason, that the *lemon* format has been proposed, which we briefly describe. We then consider two resources that seem ideal candidates for the Linked Data cloud, namely WordNet 3.0 and Wiktionary, a large document based dictionary. We discuss the challenges of converting both resources to lemon , and in particular for Wiktionary, the challenge of processing the mark-up, and handling inconsistencies and underspecification in the source material. Finally, we turn to the task of creating links between the two resources and present a novel algorithm for linking lexica as lexical Linked Data.

1 Introduction

In the last decade, a large amount of linguistic and lexical resources in particular have been created. These resources are confined however to what Tim Berners-Lee has named "data silos", as either they are publicly available, albeit in proprietary

John McCrae
CITEC, Universität Bielefeld, e-mail: `jmccrae@cit-ec.uni-bielefeld.de`

Elena Montiel-Ponsoda
Ontology Engineering Group, Universidad Politécnica de Madrid, e-mail: `emontiel@delicias.dia.fi.upm.es`

Philipp Cimiano
CITEC, Universität Bielefeld, e-mail: `cimiano@cit-ec.uni-bielefeld.de`

formats, or access to them is restricted. This leads to a situation in which the integration of various linguistic data becomes cumbersome. The *Linking Open Data* project (Berners-Lee, 2009) has aimed to solve these issues by fostering the publication of data on the Web using the RDF data model and, most importantly, linking data across sites. In this paper, we discuss how the principles of Linked Data can be applied to the publication of linguistic data. We discuss in detail the conversion of WordNet and Wiktionary to Linked Data resources using the *lemon* model as a use case. While WordNet has been already converted to the RDF data model, there are significant challenges in converting a semi-structured resource such as Wiktionary into the RDF data model. We discuss these challenges and how we addressed them. Our use cases demonstrate that *lemon* can be used as a uniform, principled and simple model for the publication of lexical resources as Linked Data as well as their linking. All resources described in this paper are available at `http://monnetproject.deri.ie/lemonsource`.

2 Related Work

There is a high interest in Natural Language Processing to exploit not only curated resources such as WordNet, but also collaboratively created resources such as Wiktionary[1] or Wikipedia.[2] These collaboratively created resources are especially interesting due to their coverage and due to the fact that they contain linguistic knowledge for a plethora of languages. Further, in spite of not having been created by linguists, they are still highly interesting for them as they contain huge amounts of semantically structured knowledge that is not typically available in standard linguistic resources (Zesch et al, 2008). A good example of a project integrating and linking various lexical resources is the NULEX project.[3] NULEX is a lexical resource derived automatically from from WordNet (Fellbaum, 1998), VerbNet (Kipper-Schuler, 2005) and Wiktionary. It reuses lexical information from WordNet and syntactic knowledge as well as subcategorization frames from VerbNet (an extension of Levin's (1993) verbs classification). By mapping these two resources, WordNet verbs are complemented with information about subcategorization. Finally, tense information is obtained from Wiktionary. However, the publication of linguistic resources as Linked Data does not solve the interoperability problem *per se*, as categories still need to be aligned to each other. Chiarcos has presented the OLiA framework (Chiarcos, 2010, also see Chiarcos, this vol.) for this purpose. It consists of a number of OWL DL ontologies that formalize the mapping between annotations of existing terminology repositories, such as GOLD (Farrar and Langendoen, 2003) or the ISOcat category registry (Kemps-Snijders et al (2008), also

[1] `http://www.wiktionary.org`

[2] `http://www.wikipedia.org`

[3] `http://www.qrg.northwestern.edu/resources/nulex.html`

see Windhouwer and Wright, this volume). OLiA thus facilitates the mapping of various annotation schemes.

3 The *lemon* Model

lemon(LExicon Model for ONtologies) (McCrae et al, in press) is an RDF model that allows to specify lexica for ontologies and allows to publish these lexica on the Web[4]. In contrast to the existing WordNet 2.0 RDF model, *lemon* is not intended to be a model for a single lexical resource, but a method by which multiple models with complementary purposes can be published, linked and shared on the Web.

The main features of the model can be summarised as follows:

- **Semantics By Reference**: Linguistic descriptions are separate from the ontology, but their semantics are defined by pointing to the corresponding semantic objects in the ontology
- **Modular Architecture:** The model consists of a core model and a set of complementary modules. Linguistic descriptions are grouped into 5 modules:
    1. Linguistic properties (e.g., part-of-speech, gender, number),
    2. Lexical and terminological variation,
    3. Decompositions of phrase structures,
    4. Syntactic frames and their mappings to the logical predicates in the ontology, and
    5. Morphological decomposition of lexical forms.
- **Openness:** *lemon* is a descriptive model that does not prescribe the usage of specific linguistic categories. Thus, the data categories or linguistic annotations used to define lexical information in the model are not captured in the *lemon* model proper, but have to be specified by reusing URIs from other dictionaries and repositories such as ISOcat or the GOLD ontology.

The core classes of the *lemon* model can be seen in Fig. 1. The core classes are the ones that form the main path between the `Ontology` and the lexical realisation represented in the `LexicalEntry` class. A `LexicalEntry` may also have multiple `LexicalForm`s representing morphological variants, each of which is associated with a written representation (`writtenRep`). The `LexicalSense` class provides a principled link between an ontology concept and its lexical realization. Since ‘concepts’ or world objects, as defined in ontologies, and ‘lexical entries’, as defined in lexicons, can rarely be said to truly overlap, the `LexicalSense` class provides the adequate restrictions (usage, context, register, etc.) that make a certain lexical entry appropriate for naming a certain concept in the specific context of the ontology being lexicalised.

[4] Technical details of the model have been described in `http://lexinfo.net/lemon-cookbook.pdf`.

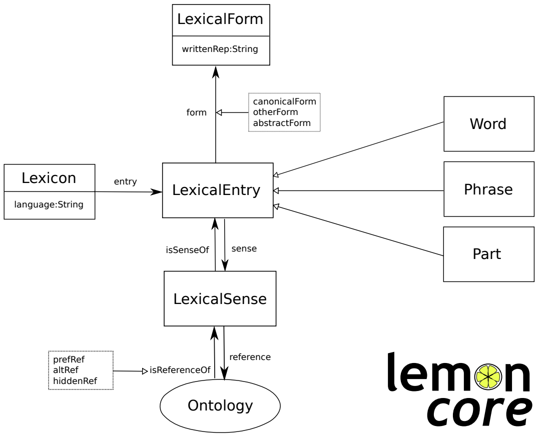

Fig. 1 The core *lemon* model

The design principles of this model make it ideal for interchanging lexica on the Web. Since *lemon* builds on the RDF data model, URIs are used to name and dereference linguistic annotations, and links can be easily created between lexicons using RDF triples. Moreover, the model is modular in the sense that, according to the final application needs, certain modules can be used or not. This also allows for new modules to be created if this is required by a certain application. In this sense, the *lemon* model can be said to be suited for the publication and linking of lexical resources on the Web.

Multilingualism is also foreseen in *lemon*, as several lexica in different languages can be associated to one and the same ontology. Moreover, translation relations can be established at the `LexicalSense` class, even allowing for conceptualization mismatches between languages to be represented, if needed. In fact, a specific module for representing translations has been proposed for *lemon* (Montiel-Ponsoda et al, 2011). The main idea of this module is to provide metadata about translations (such as provenance, confidence level, etc.), as well as to capture different types of translations (descriptive translations vs. culturally equivalent translations).

4 Methods

This section describes the methods employed to transform WordNet and Wiktionary into *lemon*, and the linking of lexical entries that are common to both resources.

4.1 WordNet

The transformation of WordNet into *lemon* has been described before by McCrae et al (2011). This conversion was performed automatically based on the manual alignment of the WordNet vocabulary to the *lemon* vocabulary. Hereby, synsets in WordNet were essentially converted into ontology concepts, words into *lemon* lexical entries, and senses into *lemon* lexical senses, respectively. The major change was the modelling of *forms* as RDF resources, in contrast to treating them as properties. A disadvantage of using ad-hoc formats when publishing lexical resources as Linked Data is the fact that schema changes might be required when the schema of the underlying resources changes. For example, when using an ad-hoc conversion to RDF schema, the conversion of WordNet 3.0 and WordNet 2.0 would yield different schemas as form variants are specified in WordNet 3.0 in extra files. Having a principled and uniform format such as *lemon* would overcome this issue of changing RDF schemas.

4.2 Wiktionary

Wiktionary is a human-readable lexicon that is publicly available on the Web, hosted by the WikiMedia foundation. It is maintained by an active community that collaboratively edits the lexicon using the 'wiki' principles. Due to its broad scope it has become an important resource for NLP research (Zesch et al, 2008). Thus, there is a general interest in converting Wiktionary into a standard machine-readable form that can be directly exploited by NLP applications. As the pages in Wiktionary are actually very regularly structured, it is is in principle straightforward to extract the data. A Wiktionary page in particular consists of at least the following sections:

- A language block, containing all entries with the same orthographic form. For example the page *cat* contains the English word as well as the Indonesian word *cat* (meaning 'paint') and the Romanian word (meaning 'storey').
- Under each language block, the entries are then grouped by part of speech, i.e., the page for *bank* has both the noun and the verb listed together.
- Alternative forms.
- Pronunciations.
- The etymology.
- The body of each entry then consists of:
  - The inflectional information for the entry, e.g., "free (*comparative* freer, *superlative* freest)".
  - An enumerated list of definitions, often with usage notes such as "archaic" or "slang".
  - A list of synonym links.
  - A list of antonyms.
  - A list of derived terms.

– A list of translations.

We have developed a parser that works as a robust finite state automaton for parsing the XML dumps of Wiktionary. The automaton is illustrated in Fig. 3; it works for pages in English, German, French, Spanish, Dutch and Japanese.

Wiktionary:

```
<page>
<title>free</title>
==English==
===Adjective===
{{en-adj}}

# Not [[imprisoned]] or [[enslaved]].
# Obtainable without any [[payment]].

====Synonyms====
* {{sense|obtainable without payment}}:
    [[free of charge]], [[gratis]]

====Translations====
{{trans-top|not imprisoned}}
* German: {{t+|de|frei}}
{{trans-bot}}
</page>
```

lemon:

```
:free_en_adj lemon:canonicalForm [
  lemon:writtenRep "free"@en ] ;
  lexinfo:partOfSpeech lexinfo:adjective ;
  lemon:sense :free_en_adj_sense0 ;
  lemon:sense :free_en_adj_sense1 ;
  lemon:sense :free_en_sense_def .

:free_en_adj_sense0 lemon:definition [
  lemon:value "Not imprisoned or enslaved"@en ] ;
  lemon:reference
    <http://en.wiktionary.org/wiki/free> ;
  lexinfo:translation :frei_de_sense_def .

:free_en_adj_sense1 lemon:definition [
  lemon:value "Obtainable without any payment"@en ] ;
  lemon:reference
    <http://en.wiktionary.org/wiki/free> ;
  lexinfo:synonym :free_of_charge_en_sense_def .
```

Fig. 2 An example of a Wiktionary entry and the corresponding *lemon* generated

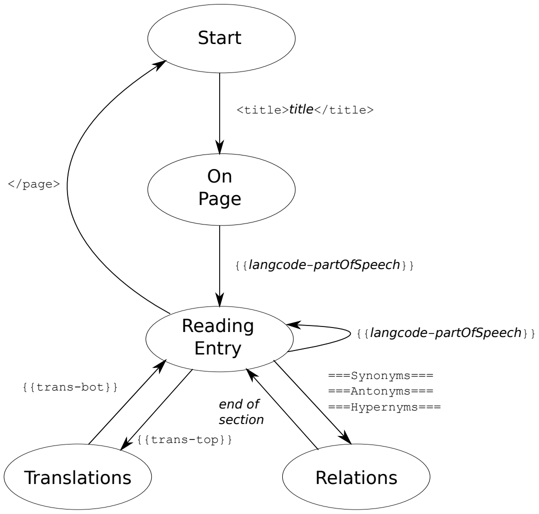

Fig. 3 The algorithm for extracting information from Wiktionary pages

Before a *lemon* model can be created for Wiktionary, a major issue is the definition of appropriate senses. As Wiktionary lists a number of definitions of the term, one could assume that they could be directly mapped to concepts in *lemon*. However, as there are different definitions per section of the article and there is no direct correspondence between these definitions, the task of collapsing various senses into an appropriate subset is not trivial. In a first step, we thus create one *lemon* sense for each definition. Then, we attempt to align the different definitions by computing the Levenshtein distance between the definitions in various sections. In the sections synonym, antonym, derived forms and translation (henceforth "SADT"), each sense used there has a gloss; *gratis*, for example, is specified as a synonym of *free* with a gloss 'obtainable without payment', and we assume that this corresponds to a definition given in the main definition section. We have observed that the ordering of definitions is similar to that of SADT senses and glosses are often short substrings of SADT glosses. Thus our algorithm for finding alignments, given a threshold λ, is:

- For each SADT sense s:
  - For each main sense m that is not already equated to some SADT sense:
    - If s's gloss is a substring of m's gloss, equate s and m; go to next sense;
    - Else, calculate the normalized Levenshtein distance between the glosses of s and m.
  - Select that main sense m that minimizes the Levensthein distance to s. If the corresponding Levenshtein distance is lower than λ, equate s and m.

The evaluation results of this algorithms that collapses senses together are presented in Tab. 1 in terms of "coverage" (indicating the percentage of senses that were mapped to a sense from the definition section), and "precision" (indicating the correctness of the mappings based on a sample of 100 randomly selected examples at each threshold level). The precision is indicated for various thresholds of the Levenshtein distance. Precision obviously increases with higher values for the threshold, but never drops below 71%. This is due to the fact that many entries have only one sense in the main definition, such that there is only one mapping candidate for the senses in the sections corresponding to SADT words.

Table 1 The results of merging duplicate senses found within Wiktionary.

	Merged	Coverage	Precision
Substring	36 595	37.8%	99.5%
> 0.9	6 842	44.9%	100%
> 0.8	3 398	48.4%	99%
> 0.7	2 669	51.2%	99%
> 0.6	3 243	54.5%	97%
> 0.5	7 128	61.9%	97%
> 0.4	4 612	66.6%	98%
> 0.3	6 295	73.1%	91%
> 0.2	7 983	81.4%	92%
> 0.1	6 934	88.5%	73%
> 0.0	3 862	92.5%	71%

4.3 Linking

As there will be many lexical entries that are common to both resources, a further goal is to identify these duplicates and merge them. This is clearly not the same as finding synonyms or equivalent synsets/sense across resources. We apply an entry linking criterion for this purpose that was previously described in McCrae et al (2011). This method proceeds by first finding entries that have the same canonical form in both resources (case was ignored). Then, we compare the part-of-speech tags of both lexical entries. Note that, as we have aligned both resources to LexInfo (Cimiano et al, 2010), this amounts to a simple string comparison. If the tags differ, we infer that the entries are different. We then check whether the remaining properties of the entry are *similar* as follows: for each property p with value v, if the other entry has a different value for p, consider the entries as different. We then also check each (non-canonical) form of the entries; for each of these forms, we find those that on the other entry are *(property) similar*. We then reject the entry if there is such a similar form on the other entry with a different written representation. In Tab. 2, we present the results of linking in terms of the number of entries that were linked against those that were not linked.

Table 2 The results of linking Wiktionary to WordNet

	#Entries	Percent (WN)	Percent (Wikt)
Linked	63,478	21.0%	26.9%
Not Linked (Wiktionary)	172,674	-	73.1%
Not Linked (WordNet)	238,408	79.0%	-
Ambiguous	1,741	0.6%	0.7%

We found that the overlap WordNet and Wiktionary in terms of lexical entries amounts to roughly between 20% and 25%. To investigate the reason for the low overlap between Wiktionary and WordNet, we took 50 entries from Wiktionary and analysed the mapping to WordNet. Out of these, 28 were also contained in WordNet. Of the remaining 32, 9 were single words which were simply absent from WordNet, e.g. *polysemic* or *abaciscus*. Further, 10 entries were compounds not present in WordNet, e.g. *false friend* and *apples and pears*. Two further entries were contained in Wordnet but with a different part-of-speech i.e. *raven* as an adjective and *to minute* as a verb. Finally, Wiktionary had a separate lexical entry for the plural noun *wares*,[5] while WordNet correctly only listed the singular form as a lexical entry. Thus, the resources seem largely complementary, a surprising result. Combining them might thus yield a lexicon that has significantly better coverage, and would therefore be of more use to applications that rely on machine readable dictio-

[5] In spite of having two entries for the plural and singular of *ware*, Wiktionary specified that *wares* is the plural of *ware*.

nary/lexica. This ultimately corroborate the usefulness of creating / linking lexica following the Linked Data principles.

5 Discussion and Conclusion

In this paper we presented a case study showcasing the publication and linking of lexical resources following the Linked Data principles. The conversion of WordNet to Linked Data was rather straightforward due to the fact that an existing RDF export was available. We have argued that when using an ad-hoc RDF format for publishing resources, changes to the RDF schema might become necessary if the underlying data structures change. This can be alleviated by using a principled model such as *lemon*.

The conversion of Wiktionary to Linked Data was more intricate as it represents a semi-structured resource that needs to be parsed appropriately before. We found that *lemon* seems an adequate model which revealed important flaws in the design of Wiktionary, i.e. the fact that correspondences between sense definitions in various sub-sections of the article are not explicitly modelled. To address this, we have proposed a simple yet effective algorithm to align the definitions across sections. Finally, we proposed an approach to linking lexical entries across WordNet and Wiktionary, showing that the overlap between the two resources was lower than expected. Integrating both resources promises to create a wide-coverage resource that can be exploited in NLP applications. Instead of creating a new lexical resource from these two resources, we have shown how we can create a virtual new resource by applying the Linked Data technologies, linking lexical entries across both resources. In our view, the adoption of Linked Data principles is thus a promising method for extending the life cycle of linguistic resources In order to integrate resources in a principled manner, a common model is needed. In this paper, we have proposed the *lemon* model for this and shown that it provides a principled model to which the lexical resources we have considered (WordNet and Witkionary) could be straightforwardly converted to.

Acknowledgements This work was developed in the context of the Monnet project, which is funded by the European Union FP7 program under grant number 248458, the CITEC excellence initiative funded by the DFG (Deutsche Forschungsgemeinschaft), and the Spanish national project BabeLData (TIN2010-17550).

References

Berners-Lee T (2009) Linked Data-The Story So Far. International Journal on Semantic Web and Information Systems 5(3):1–22

Chiarcos C (2010) Grounding an Ontology of Linguistic Annotations in the Data Category Registry. In: Proceedings of the 2010 International Conference on Language Resource and Evaluation (LREC)

Chiarcos C (this vol.) Interoperability of corpora and annotations. P. 161-179

Cimiano P, Buitelaar P, McCrae J, Sintek M (2010) Lexinfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web

Farrar S, Langendoen D (2003) Markup and the GOLD Ontology. In: Proceedings of Workshop on Digitizing and Annotating Text and Field Recordings

Fellbaum C (1998) WordNet: An electronic lexical database. MIT press Cambridge, MA

Kemps-Snijders M, Windhouwer M, Wittenburg P, Wright S (2008) ISOcat: Corralling data categories in the wild. In: Proceedings of the 2008 International Conference on Language Resource and Evaluation (LREC)

Kipper-Schuler K (2005) Verbnet: A broad coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania

Levin B (1993) English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago

McCrae J, Spohr D, Cimiano P (2011) Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. The Semantic Web: Research and Applications pp 245–259

McCrae J, Aguado-de Cea G, Buitelaar P, Cimiano P, Declerck T, Gomez-Perez A, Gracia J, Hollink L, Montiel-Ponsoda E, Spohr D, Wunner T (in press) Interchanging lexical resources on the semantic web. Language Resources and Evaluation

Montiel-Ponsoda E, Gracia J, Aguado de Cea G, Gómez-Pérez A (2011) Representing translations on the semantic web. In: Proceedings CW (ed) Proceedings of the 2nd International Workshop on the Multilingual Semantic Web 2011 (MSW 2011), vol 775, pp 25–37

Zesch T, Müller C, Gurevych I (2008) Extracting lexical semantic knowledge from wikipedia and wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC), Citeseer, pp 1646–1652