# Linking Lexical Resources and Ontologies on the Semantic Web with lemon

John McCrae, Dennis Spohr, and Philipp Cimiano

AG Semantic Computing, CITEC, University of Bielefeld
{jmccrae,dspohr,cimiano}@cit-ec.uni-bielefeld.de

**Abstract.** There are a large number of ontologies currently available on the Semantic Web. However, in order to exploit them within natural language processing applications, more linguistic information than can be represented in current Semantic Web standards is required. Further, there are a large number of lexical resources available representing a wealth of linguistic information, but this data exists in various formats and is difficult to link to ontologies and other resources. We present a model we call *lemon* (Lexicon Model for Ontologies) that supports the sharing of terminological and lexicon resources on the Semantic Web as well as their linking to the existing semantic representations provided by ontologies. We demonstrate that *lemon* can succinctly represent existing lexical resources and in combination with standard NLP tools we can easily generate new lexica for domain ontologies according to the *lemon* model. We demonstrate that by combining generated and existing lexica we can collaboratively develop rich lexical descriptions of ontology entities. We also show that the adoption of Semantic Web standards can provide added value for lexicon models by supporting a rich axiomatization of linguistic categories that can be used to constrain the usage of the model and to perform consistency checks.

## 1 Introduction

The Semantic Web has made available a large amount of semantic data in the form of ontologies and there have been several attempts to apply this to NLP tasks such as question answering [17], information extraction [7] and text generation [2]. However, current standards such as RDFS and SKOS [18] only allow for limited linguistic information to be attached to an ontology, limiting the potential functionality of these applications. In contrast, there are a large number of rich sources of linguistic information that have been created including term bases, lexica (e.g., Lefff [20]) and machine readable dictionaries (e.g., WordNet [9]). However, much of this data is confined by the format and distribution methodology to "data silos" and as such cannot be easily shared or extended. This proves to be a specific disadvantage for the creation of lexical resources for specific domains (e.g., SNOMED [22]), as these terminologies inevitably need to reuse basic terms from general-domain resources. For these reasons, we propose a new model called *lemon* (Lexicon Model for Ontologies) that is designed to

allow lexical information to be represented relative to an ontology and shared on the Semantic Web. The *lemon* model has the following crucial features: i) it represents a concise and thus reusable model, ii) it is based on RDF(S), iii) it is "open" in the sense that it does not prescribe the usage of a particular inventory of linguistic categories and properties, but instead iv) supports the reuse of any linguistic ontology such as GOLD [8] or ISOcat [13], and v) assigns semantics to lexical entries by way of reference to ontological entities in line with Buitelaar [4].

There have already been several attempts to define linguistic ontologies, notably the GOLD ontology [8] and the OLiA ontologies [5]. However, these ontologies are primarily focused on providing specific linguistic categories and do not define a methodology for representing morphosyntactic information. Instead, we base our model on the Lexical Markup Framework (LMF) [10] with the goal of making lexica interoperable. In particular, we include the idea of *data categories* [19], which are uniquely identified concepts that can be used for computational linguistic tasks, such as those compiled by the ISOCat [13] project.

The paper is structured as follows: after discussing related work in Section 2, in Section 3 we introduce the *lemon* model as a basic model for representing linguistic information relative to ontologies, that uses existing ontologies and/or data category registries to represent specific linguistic categories. In Section 4 we present an extension of *lemon* called *lemon*-LexInfo which makes particular choices with respect to the linguistic categories and properties that can be modelled by importing categories from ISOCat and COMLEX [16], for instance. In Section 5 we present three experiments which show how *lemon* lexica can be created automatically as well as by reuse of WordNet. In a first experiment we show that legacy lexica such as WordNet can be easily converted to the *lemon* format. The main benefit here is that by this move, lexica can be linked to each other, extended and reused in a straightforward manner by exploiting the RDF datamodel and the Linked Data principles [1]. In a second experiment, we show how *lemon*-lexica for already existing ontologies can be created in an automatic fashion by building on standard NLP components, thus substantially reducing the costs of creating such lexica. Finally, in a third experiment, we show how general lexica such as WordNet can be reused when constructing a lexicon for a specific vocabulary such as FOAF, thus saving costs and resources in the creation of *lemon*-lexica. We conclude in Section 6.

## 2 Background

RDFS's label property provides a simple way to attach a lexical form to an ontological concept. The SKOS model [18] goes further by allowing to define a preference order on labels as "preferred", "alternative" and "hidden." However, modern lexica as developed in the lexical resources community, for example Lefff [20], contain more information than can be succinctly represented with these vocabularies, in particular morphology, phrase structure and subcategorization information. WordNet [9] is of course one of the most well-known lexica and there have been several attempts to adapt it to the Semantic Web, e.g. by transforming

it into an RDF format [24] and "ontologizing" it [12]. However, these lexica are limited by the actual amount of data available in WordNet and by the format of WordNet itself. In addition, the conceptual model used by WordNet has been identified as unsound from an ontological persepective (see [11]). In general, we wish for a model that is capable of representing a large variety of linguistic information and can do this for an arbitrary ontology.

While SKOS fails on the former, WordNet and similar domain-independent resources fail with respect to the latter. These two desiderata can be met by building on standardisation efforts carried out in the lexical resources community, in particular the Lexical Markup Framework [10]. LMF is capable of representing a wide variety of linguistic information, however it has no mechanism for relating lexica to ontologies and instead relies on a traditional word sense model as in WordNet, which has been criticised by Kilgariff[14].

The LexInfo model [6] is an ontology-lexicon model which has a clearly separate linguistic layer and a semantic-syntactic correspondence object. The LexInfo model was created by importing LMF. But the authors noted there were many technical issues with this, not least that there is still no canonical form of LMF that is usable for the Semantic Web, in the sense of being correct RDF and having dereferencable URIs. The authors fixed this by publishing their own version of LMF[1] and enhancing it by introducing names for the property relations: i.e., replacing LMF's 3 original properties (`isAssociated`,`isPartOf`,`isAdorned`) with more specific links such as `hasWordForm`.

| | WordNet (2.0) | GOLD | ISOcat |
|---|---|---|---|
| Number of values | 5 | 81 | 115 |
| | OLiA | LexInfo (1.0) | |
| Number of values | 174 | 11 | |

**Table 1.** Number of values for part of speech in some existing formalisms

An important problem in the representation of lexica is that there is significant disagreement between different models with respect to the properties and values of linguistic annotation that are needed to represent a lexicon. Take as an example the case of part of speech, which is a property that most lexica represent but often have a significant disagreement with respect to the number and granularity of part-of-speech tags. For example, in Table 1 we show the number of values of part of speech that can be represented by some language resource schemas. As we can see, the scope of the representation varies greatly. In addition, there is some disagreement about what constitutes a part of speech: for example, GOLD [8] considers "comparative adjective" to be a part of speech value, while the other formalisms consider the "comparative" value to be assigned to a property "degree." Furthermore, there is also disagreement about

[1] `http://www.lexinfo.net/lmf`

the hierarchy of these concepts: for example, GOLD has no class for "Verb" and instead groups adjectives, adverbs and verbs under the concept "predicator".

The OLiA project [5] attempted to solve this problem by aligning several linguistic annotation ontologies including GOLD, and ISOcat to a reference ontology. Within the ISO TC 37, the problem of different linguistic annotations has been handled by data categories, which are sets of values for such properties. These are currently being collected by the ISOcat project [13]. For our purposes, we require a formalism that does not distinguish between these different resources and can use any of them depending on the wishes of the lexicon creator.

## 3 The *lemon* model

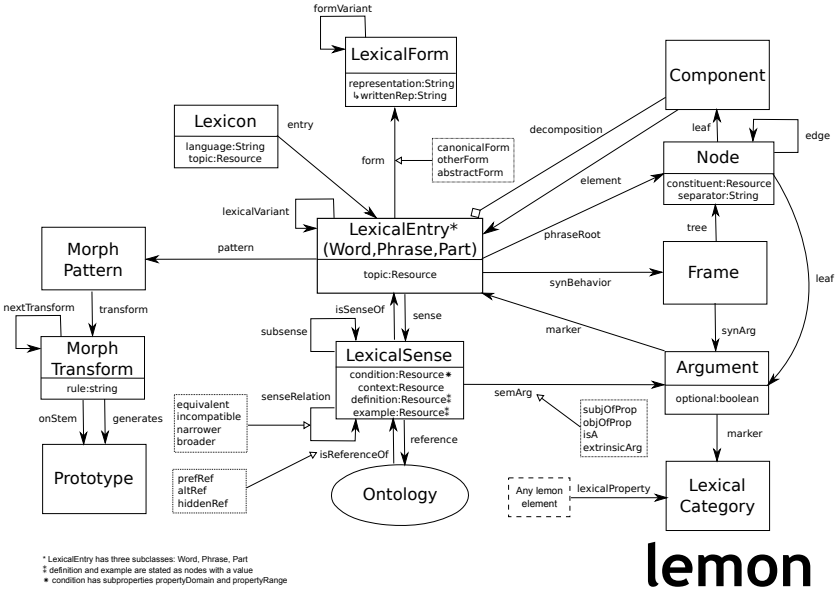



**Fig. 1.** The *lemon* model

We present the *lemon* model – illustrated in Figure 1 – as our proposal for a lexicon model for ontologies. The *lemon* model consists of a lexicon object with a number of (lexical) entries. Each of these entries can then be further described with morphosyntactic properties, and mapped via (lexical) sense objects to entities in the ontology. The core elements of the ontology are as follows

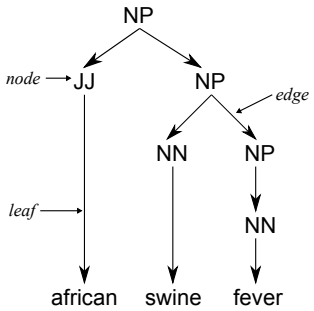
– **Lexicon**: The lexicon is realised as a resource. Each lexicon is mono-lingual and is marked with a language tag and optionally a topic
  - *Example:* A lexicon may consist of English names for diseases.

- **Lexical Entry**: The lexical entry represents a single term within the lexicon. As morphosyntactic information is attached to the lexical entry, each entry must have the same syntax. Hence term variants, such as abbreviation, are represented as separate lexical entries and marked as **lexicalVariant**s. Lexical entries are split up into three subclasses: **Word**, **Phrase** and **Part** (of word).
  - *Example:* "Cancer of the mouth" is a lexical entry in the lexicon. "Mouth cancer" would be another lexical entry, marked as a lexical variant of the first.
- **Form**: Each lexical entry consists of a number of forms. These represent different inflectional variants of the entry and may be marked as **canonical** (lemma), **other** or **abstract**.
  - *Example:* The lexical entry for "bacterium", may have two forms: the canonical "bacterium" and the other form "bacteria".
- **Representation**: Each form may have multiple representations, of which the most important is the written representation, but other representations such as a phonetic form are also possible.
  - *Example:* The written representation of the form of bacterium would be "bacterium". It may also have a phonetic representation bktim
- **Lexical Sense**: Unlike in other models, *lemon*'s senses are not assumed to be finite or disjoint. Instead, the sense represents the correspondence between the lexical entry and the ontology entity. It may include extra specification of this correspondence such as context and condition, or human-readable annotations such as definitions or examples. It may be indicated as the **preferred**, **alternative** or **hidden** lexicalisation of an ontology entity, by analogy to the preference order on labels defined by SKOS [18] in terms of *preferred*, *alternative* and *hidden*.
  - *Example:* The lexical entries for "influenza" and "flu" may both refer to an ontology entity `http://purl.org/obo/owl/DOID#DOID_8469` (in the OBO foundry ontologies [21]). Each entry would have a separate sense object. The former sense would be marked as used in a scientific context, and the latter as a layman term.
- **Reference**: The meanings of a lexical entry are specified through a "reference" to an ontology entity, and hence the lexical entry is linked to the semantic description given by the ontology.
- **Property**: Any element in a *lemon* model may be further described by a property. *lemon* offers a generic property `lexicalProperty`, which other linguistic properties should derive from, so that all lexical properties can be grouped.
  - *Example:* The forms, "bacterium" and "bacteria", may have a property "number" with values "singular" and "plural" respectively. The lexical entry may also be marked with part of speech noun.
- **Frame** and **Argument**: Subcategorization frames represent the valency of verbs and other lexical predicators, i.e., the number and type of arguments it can or should take. Each argument is also represented as a resource and is linked both from the frame, to indicate the syntactic role, and from the sense, to indicate the semantic role.

- *Example:* For example, the property `complicated_by` may be represented by a lexical entry with a frame corresponding to "Y complicates X" where "X" and "Y" are the arguments of this frame.

– **Component**: Each lexical entry may be split up into a list of components, each of which refers to other lexical entries. This is used for showing which words compose a multi-word expression or a compound word. The list of components are stated as an RDF list, hence the list of components is ordered and finite.
  - *Example:* The German term "hmorrhagisches Fieber" ("haemorrhagic fever"), is composed of two components "hmorrhagisch" and "Fieber." The first component may have properties to indicate that it is the form with neuter adjectival agreement. Decompositions may also be used with compound words, for example the German term, "Ebolavirus" ("Ebola virus"), may have a decomposition into "Ebola" and "Virus."

– **Node**: Each lexical entry may be associated with a phrase structure. This consists of a number of nodes linked by either **edge** or **leaf** arcs to components
  - *Example:* A parse tree may be constructed for the term "African swine fever" as below.




This is useful as it indicates that this term is understood as an "African" version of "swine fever" instead of a "fever" affecting "African swine."

The *lemon* model thus provides a general framework by which we can represent lexica linked to ontologies to specify the semantics of lexical entries. However, for most applications a specific vocabulary needs to be introduced to describe the specific linguistic categories used in the model. We do not have space to go into the full details of the usage of the model. A full technical report on the *lemon* model is available at `http://www.lexinfo.net/lemon-cookbook.pdf`.

It should be noted that *lemon* is not technically an instantiation of LMF as there are many differences in the modelling of semantics and optimizations due to the adoption of RDF. However, many aspects of *lemon* do correspond directly to LMF and in fact there is a *lemon*-LMF converter available at `http://www.lexinfo.net/lemon2lmf`.

In order to represent information such as part of speech, we can include this information by referencing URIs from a data category registry. GOLD, OLiA and ISOcat all provide URI based identifiers for their systems, such that we can reference any property and gain more information about this annotation by

deferencing this URI. As *lemon* is based on RDF, it is trivial to include these URIs as resources in our lexicon scheme. For example, the following represents a single lexical entry for the Dutch word "maag" ("stomach").

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix isocat: <http://www.isocat.org/datcat/> .

:maag
   lemon:canonicalForm [ lemon:writtenRep "maag"@nl ;
        isocat:DC-1298 isocat:DC-1387 ] ; # number=singular
   lemon:otherForm     [ lemon:writtenRep "magen"@nl ;
        isocat:DC-1298 isocat:DC-1354 ] ; # number=plural
   isocat:DC-1345 isocat:DC-1333 ; # partOfSpeech=noun
   isocat:DC-1297 isocat:DC-1880 ; # gender=feminine
   lemon:sense [
     lemon:reference <http://purl.org/obo/owl/EHDAA#EHDAA_2993> ] .

isocat:DC-1298 rdfs:subPropertyOf lemon:property .
isocat:DC-1345 rdfs:subPropertyOf lemon:property .
isocat:DC-1297 rdfs:subPropertyOf lemon:property .
```

Note that the prescribed URIs for ISOcat data categories are specified with a registration number. For legibility, we include comments to give a human-readable description of the properties used in the example.

Expanding on this example we can also model multilinguality by including lexical entries that have the same reference in different languages, for example:

```
:maag
  lemon:canonicalForm [ lemon:writtenRep "maag"@nl ] ;
  lemon:sense [
    lemon:reference <http://purl.org/obo/owl/EHDAA#EHDAA_2993> ] .

:stomach
  lemon:canonicalForm [ lemon:writtenRep "stomach"@en ] ;
  lemon:sense [
    lemon:reference <http://purl.org/obo/owl/EHDAA#EHDAA_2993> ] .
```

In this way we can publish translated versions of common lexicon without any need to modify the original lexicon or the original ontology.

## 4 The *lemon*-LexInfo model

*lemon* can accomodate any data category scheme, however the relations between the data categories and the *lemon* model must be specified in each lexicon. For this reason we adapted the LexInfo model[6], which was originally introduced as an extension of LMF. This second version of LexInfo was engineered from

the ground up using *lemon* and importing data categories from the ISOcat data category registry. For *lemon*-LexInfo we expand on the RDF schema used in *lemon* with OWL to create links to these external linguistic ontologies as well as to axiomatize certain linguistic types, subcategorization frames in particular and to further constrain their meaning and usage, thus supporting consistency checks.

- **Axiomatic definitions of linguistic categories**: From ISOcat [13] we converted the DCIF files for the morphosyntax section into RDF and imported them into the *lemon*-LexInfo model. We mapped the "complex" data categories to RDF properties and the "simple" data categories to RDF resources (OWL Individuals). This gave us a very large set of properties that can be used to describe entries in the lexicon. We then defined each of these properties as subproperties of the appropriate *lemon* property. Most of these came under the general **lexicalProperty** arc, but some were mapped elsewhere, e.g., **register** was modelled as a (sense) **context**, as it represents a semantic distinction on the usage of a term.
- **Instantiating a hierachy of categories**: For each of the properties adopted from ISOcat, their range was defined in terms of a class having as individuals all elements in the extension of this class according to ISOcat DCIF. In this way we can introduce hierarchy among the annotations, e.g., **properNoun** and **commonNoun** are both members of the classes **PartOfSpeech** and **NounPOS**. *lemon*'s three lexical entry classes (**Word**, **Phrase** and **Part**) were further subclassed into specific classes, e.g, **Verb**, **NounPhrase**. As each of these classes could be related to the properties introduced from ISOcat, we introduced appropriate axioms to define these classes. For example:

$$Noun \equiv \exists partOfSpeech.NounPOS$$

- **Compositional definition of subcategorization frames**: We define linguistic frames in a precise manner in terms of the sets of arguments they have. We introduced a set of syntactic role properties, for example, "subject", "object" and then created precise OWL definitions of a each frame from the COMLEX [16] vocabulary.
  We can now define an intransitive frame as a frame with a subject, no direct object and no indirect object as follows:

  $$IntransitiveFrame \equiv (= 1 subject) \sqcap (= 0 directObject) \sqcap (= 0 indirectObject)$$

  We found that we could further simplify the description of subcategorization frames by defining abstract frames such as **PrepositionalObjectFrame**, so we could then define a hierarchy of frames. E.g.,

  $$PrepositionalObjectFrame \equiv \exists prepositionalObject$$

  $$IntransitivePPFrame \equiv IntransitiveFrame \sqcap PrepositionalObjectFrame$$

  In this way, we reduced COMLEX's 163 frames to 36 basic frames and 4 modifiers to describe argument control. These are listed at `http://www.lexinfo.net/basic-frames`.

This illustrates the value of using Semantic Web standards, as we did not need to define specific vocabulary to define these linguistic concepts. Instead, OWL was sufficient to provide powerful modelling of linguistic concepts. To further illustrate this point, we note that it is also then possible to use OWL to define linguistic conditions, such as "every French noun is masculine and/or feminine," without requiring an extra modelling language such as in LMF.

## 5 Experiments

### 5.1 Converting WordNet to *lemon*

As we wish to use *lemon* to create Semantic Web lexica, we require that it is capable of representing legacy lexical resources. One of the largest, freely available lexica is WordNet [9] and it seems clear that it is necessary for *lemon* to be able to represent the information in this resource. We based our work on the existing RDF version of WordNet [24] and then simply aligned this to the *lemon* model. We proceeded as described below to yield a *lemon*-compatible version of WordNet.

**Methodology**

- We mapped WordNet's synsets to *lemon*'s references. This means that the synsets and the links between them form a quasi-ontology and replace the role that the ontology plays in normal usage of *lemon*, i.e. assigning meaning to lexical entries by reference to ontological entitites. The advantage of this separation is that we can introduce mappings to more sound semantic models such as OntoWordNet [12] without affecting the original data.
- The definition of word sense in WordNet and *lemon* corresponded well, as WordNet's word senses can be defined as the sub-meaning of a word belonging to a particular synset and *lemon*'s sense as the intersection between the lexical usage of the entry and the semantic usage of the ontology entity.
- We also found that the definition of word in WordNet and word in *lemon* corresponded, so mapped these appropriately. We note here that the original RDF version of WordNet actually loses information as alternative forms are listed in the original WordNet format. Hence, we manually extracted them and added them to the *lemon* representation.
- Finally, WordNet marks part of speech on the sense and synset level, whereas *lemon* does it on the word level. We switch the properties to the lexical entries using the morphosyntactic properties of LexInfo that were originally derived from ISOcat.

As such a brief example of a rewritten WordNet synset is as follows (`lwn` is the lemon-WordNet namespace and `wn20` the original WordNet-RDF mapping):

```
lwn:marmoset-noun-entry rdf:type lemon:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
```

```
  lemon:sense lwn:sense-marmoset-noun-1 ;
  lemon:canonicalForm lwn:word-marmoset-canonicalForm .

lwn:sense-marmoset-noun-1 lemon:reference wn20:synset-marmoset-noun-1 .

lwn:word-marmoset-canonicalForm lemon:writtenRep "Marmoset"@en .
```

**Discussion** We found that the *lemon* model was relatively close to WordNet. By mapping this to a common vocabulary, we believe it should make it easier to combine multiple lexica without losing information. In addition, as *lemon* can provide more complex representations of syntactic and morphological information we believe this could enable WordNet to be further extended vertically. In the future, we intend to extend this work by incorporating other open-source lexica, such as Wiktionary[2] and Lefff [20]. These resources are available at `http://www.lexinfo.net`.

### 5.2 Generating *lemon*-LexInfo models

The main goal of *lemon* is to create lexica that can be used to describe ontologies, as such, for the second experiment we chose to create a model for a widely used ontology. The "Friend of a friend" (FOAF) [3] is an ideal candidate for testing whether the model is concise and there is a large amount of FOAF data available on the Web. As the model is small, it seems feasible to develop a corresponding lexicon manually in order to guarantee a high quality result.

**Methodology** We used the *lemon*-LexInfo lexicon generation service available at `http://monnetproject.deri.ie/Lemon-Editor`, which provides an interface for working with *lemon* lexica and incorporates a number of basic NLP tasks so that it can auto-generate most of the information required for lexicon generation. In particular, the service has the following features:

- Extraction of labels from RDFS, SKOS or URI fragments.
- Tokenization yields sub-components. In particular, we used the standard tokenizer that is packaged with the Lucene information retrieval library[3].
- Part of speech tagging to give simple morphosyntactic features. We used the Stanford Tagger [23] for our experiments.
- Lemmatization to identify which forms are canonical. We again used the Stanford Tagger to perform this.
- Parsing to produce phrase structure. For this, we used the Stanford Parser [15].
- Subcategorization identification using a rule-based system. Each rule consists of the following:

[2] `http://www.wiktionary.org/`
[3] Available at `http://lucene.apache.org`

- A phrase structure pattern to detect the structure of the label, represented by the lexical entry. For example, the pattern, `FRAG (VP, PP)`, indicates a fragment consisting of a verb phrase and a prepositional phrase, such as "located in".
- A set of a classes which the ontology entity should be an instance of (by RDF's `type` property). These are generally basic OWL types such as `ObjectProperty`.
- The class of the generated frame.
- The definition of the arguments of the frame based on the syntactic and semantic roles it has.

| | Total Errors | Total Correct | Precision |
|---|---|---|---|
| Tokenizer | 1 | 112 | 99.1% |
| Tagger | 8 | 105 | 92.9% |
| Parser | 12 | 101 | 89.4% |
| Subcategorizer | 6 | 57 | 92.1% |
| Subcategorizer (with parses corrected) | 2 | 61 | 96.8% |
| Total | 21 | 92 | 81.5% |

**Table 2.** Results by component for lexicon generation on the FOAF ontology

We used the service by uploading a standard version of the FOAF ontology. This version of FOAF contained 63 entities (of which 12 were classes and 51 properties) and the generation process created 113 Lexical Entries (note that extra lexical entries were created to describe multiple word expressions). The results by component are described in Table 2. The results show that the lexicon generation is very accurate at different levels having accuracy levels between 89.4% (parsing of labels) and 99.1% (tokenization of labels). Overall, the number of lexical entries for which there is an error with respect to one level of linguistic analysis is 21 out of 103, thus corresponding to an overall accuracy of 81.5%. This is a very satisfactory result showing that we can generate lexica for a given ontology effectively and efficiently.

**Discussion** The tokenizer component was quite accurate producing only one error (splitting "E-commerce" into two words), and the tagger was relatively accurate. Most of the tagger's errors were related to not distinguishing correctly between common nouns and proper nouns. The parser was responsible for most of the errors, in particular the implementation we used was biased to produce full sentence parses. For example, the label "work info homepage" was interpreted as an imperative sentence instead of a noun phrase[4]. The subcategorizer

[4] Discarding sentence parses was not effective here as the next best parse was the same verb phrase

was generally correct, however, it should be noted that the vast majority of the labels in the source ontology (like with many ontologies) are simply noun phrases with the result that the subcategorization frames were mostly noun phrases with possessive adjunct (i.e., "X is the homepage of Y") or unary noun phrase predicates (i.e., "X is a homepage"). Once we corrected all the incorrect parses and reran the subcategorizer, the accuracy improved, generating only two incorrect frames: "X is the Myers Briggs of Y" and not recognizing "account" in "holds account" as the object of the verb. We would hope to conduct a more thorough evaluation of this component in later work, on an ontology with more complex labels and modelling for predicates with arity greater than two (e.g., donative structures).

### 5.3 Merging generated lexica with existing LR

Obviously, a large amount of the terminology used within the FOAF vocabulary is also found within WordNet. Thus, it seems that it would be advantageous to reuse the WordNet entries when defining a lexicon for FOAF. We can easily achieve this in *lemon* by creating a sense object for each meaning specified in the FOAF ontology and then linking it to our *lemon*-aligned version of WordNet if possible and only creating a new lexical entry if this is not possible, thus fostering reuse, producing more compact lexica and ultimately reducing the costs in lexicon creation.

**Methodology** We took the WordNet lexicon generated by the approach described in Section 5.1 and the lexicon for the FOAF ontology generated in section 5.2, and compared each of the entries in the two lexica. We used the following criteria to evaluate if two entries were equivalent:

- The written representation of the canonical from was the same or differed only by capitalization of the initial letter.
- The part of speech tag was equal, if specified.
- The two entries did not have a linguistic property with different values. Note that we still counted the entries as equal if one entry did not have a value for the linguistic property.
- The non-canonical forms could be matched in such a way that each corresponding pair had the same written representation and did not have contradictory property values. We need to search for similar pairs here as it is possible that one lexicon may have, for example, "made" as both the preterite and past participle form of the entry "make".

We then used this definition of equality to map FOAF lexical entries to WordNet, replacing the generated FOAF lexical entries with WordNet entries whenever an equivalent – as defined above – WordNet entry exists. Note that we only mapped to the words within the WordNet model and not to senses. Thus, for an ambigous term like "ID" we did not decide whether the meaning in FOAF was as an abbreviation of "identification" or "Idaho." We note here that

| | Number | Percentage |
|---|---|---|
| Mapped to WordNet | 78 | 69.0% |
| Not mapped (MWE) | 25 | 22.1% |
| Not mapped (Proper Noun) | 9 | 7.9% |
| Not mapped (other) | 1 | 0.9% |

**Table 3.** Number of lexical entries for FOAF lexicon mapped to WordNet

*lemon* does not require us to make this distinction, but we can as it is possible to reuse either lexical entries, which aren't semantically disambiguated, or senses, which are disambiguated. The results of this mapping process are depicted in Table 3. The table shows that we can successfully map 69% of the lexical entries derived from the FOAF ontology to appropriate WordNet lexical entries. The remaining 31% of the cases can be broken down into three groups: firstly, there were a number of multiple word expressions in FOAF that were not contained within WordNet, for example "past project" or "online gaming account." Secondly, FOAF contained a number of proper nouns to refer to specific social networking services such as "MSN" ("The Microsoft Network") or AIM ("AOL Instant Messenger") although some proper nouns were contained within WordNet, e.g., "Yahoo." Thirdly, we found one neologism, "weblog", that was not in WordNet.

**Discussion** These results show that there is significant value in reusing existing lexical resources in the creation of lexica for new domains, as the majority of terms used by the FOAF ontology were also found in WordNet. However, we also conclude that most domain ontologies will need to introduce new terminology, and as such there is a necessity to collaboratively expand lexical resources, through the use of linked data and semantic web search engines. In particular, this is most notable for multiple word expressions and proper nouns, both of which are contained in WordNet, but only to a limited degree.

## 6 Conclusion

We have presented our model *lemon*, which acts as a basic model for publishing lexica on the Semantic Web and connecting them to ontologies. The model's openness allows it to be a concise model and hence easy to use and work with. We have also introduced an extension of the model called *lemon*-LexInfo which makes specific design choices by reusing existing linguistic categories defined in ISOcat and COMLEX. The reuse of existing data categories exemplifies how *lemon* can be used to publish lexical resources in a way that avoids the data being confined to "silos." We have demonstrated that the RDF based foundations of *lemon* make it trivial to include these data categories. Furthermore, the use of RDF allows us to gain added value in the description of our lexica, as

we demonstrate by using OWL to simplify the process of describing subcategorization frames. By converting WordNet to *lemon*, we demonstrate the utility of *lemon* as an interchange format that could be used to bring complementary lexical resources together under a single framework. In this way, we believe this model could be used to bring together lexical resources with the semantic modelling on the Semantic Web. We also show that by the use of standard NLP components we can generate high-quality lexica. Finally, we show that for developing lexical resources for specific domains both the reuse of existing lexical resources and the generation of new lexical resources can be used together to effectively and collaboratively develop new resources.

## References


1. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
2. K. Bontcheva. Generating tailored textual summaries from ontologies. In *The Semantic Web: Research and Applications*, pages 531–545. Springer, 2005.
3. D. Brickley and L. Miller. *FOAF Vocabulary Specification 0.98*, 2010. Accessed 3 December 2010.
4. P. Buitelaar. Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. In *Ontology and the Lexicon*, pages 212–223. Cambridge University Press, 2010.
5. C. Chiarcos. Grounding an Ontology of Linguistic Annotations in the Data Category Registry. In *Proceedings of the 2010 International Conference on Language Resource and Evaluation (LREC)*, 2010.
6. P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29 – 51, 2011.
7. N. Collier, S. Doan, A. Kawazoe, R. Goodwin, M. Conway, Y. Tateno, Q. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi. BioCaster: detecting public health rumors with a Web-based text mining system. *Oxford Bioinformatics*, 24(24):2940–2941, 2008.
8. S. Farrar and D. Langendoen. Markup and the GOLD Ontology. In *Proceedings of Workshop on Digitizing and Annotating Text and Field Recordings*, 2003.
9. C. Fellbaum. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.
10. G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. Lexical markup framework (LMF). In *Proceedings of the Fifth International Conference on Language Resource and Evaluation (LREC'O6)*, 2006.
11. A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. Sweetening wordnet with DOLCE. *AI magazine*, 24(3):13, 2003.
12. A. Gangemi, R. Navigli, and P. Velardi. 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, 2003.
13. M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S. Wright. ISOcat: Corralling data categories in the wild. In *Proceedings of the 2008 International Conference on Language Resource and Evaluation (LREC)*, 2008.
14. A. Kilgarriff. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113, 1997.

15. D. Klein and C. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, 2003.
16. A. Korhonen, Y. Krymolowski, and T. Briscoe. A large subcategorization lexicon for natural language processing applications. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC06)*, 2006.
17. V. Lopez, M. Pasin, and E. Motta. Aqualog: An ontology-portable question answering system for the semantic web. In *The Semantic Web: Research and Applications*, pages 546–562. Springer, 2005.
18. A. Miles and S. Bechhofer. *SKOS Simple Knowledge Organization System Reference*, 2009. Accessed 19 October 2010.
19. L. Romary. Standardization of the formal representation of lexical information for NLP. In *Dictionaries: An International Encyclopedia of Lexicography.* Mouton de Gruyter, 2010.
20. B. Sagot. The Lefff, a freely available and large coverage morphological and syntactic lexicon for French. In *Proceedings of the 2010 International Conference on Language Resource and Evaluation (LREC)*, 2010.
21. B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255, 2007.
22. K. Spackman, K. Campbell, and R. Côté. SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA Annual Fall Symposium*, pages 640–644, 1997.
23. K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, 2003.
24. M. Van Assem, A. Gangemi, and G. Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC06)*, 2006.