Language Resources and Evaluation manuscript No.
(will be inserted by the editor)

Interchanging lexical resources on the Semantic Web

John McCrae · Guadalupe Aguado-de-Cea · Paul Buitelaar · Philipp Cimiano · Thierry Declerck · Asunción Goméz-Pérez · Jorge Gracia · Laura Hollink · Elena Montiel-Ponsoda · Dennis Spohr · Tobias Wunner

Received: date / Accepted: date

Abstract Lexica and terminology databases play a vital role in many NLP applications, but currently most such resources are published in application-specific formats, or with custom access interfaces, leading to the problem that much of this data is in "data silos" and hence difficult to access. The Semantic Web and in particular the Linked Data initiative provides effective solutions to this problem, as well as interesting possibilities for data reuse by inter-lexicon linking, and incorporation of data categories by dereferencable URIs. The Semantic Web focuses on the use of ontologies to describe semantics on the Web, but currently there is no standard for providing complex lexical information for such ontologies and for describing the relationship between the lexicon and the ontology. We present our model, *lemon*, which aims to address these gaps while building on existing work, in particular the Lexical Markup Framework, the ISOcat Data Category Registry, SKOS (Simple Knowledge Organization System) and the LexInfo and LIR ontology-lexicon models.

J. McCrae, P. Cimiano, D. Spohr
CITEC, University of Bielefeld
Universitätsstraße, Bielefeld, Germany
E-mail: {jmccrae, cimiano, dspohr}@techfak.uni-bielefeld.de

G. Aguado-de-Cea, A. Goméz-Pérez, J. Gracia, E. Montiel-Ponsoda
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, Boadilla del Monte, Spain
E-mail: {lupe, asun, jgracia, emontiel}@fi.upm.es

P. Buitelaar, T. Wunner
DERI, National University of Ireland, Galway
Galway DERI IDA Business Park, Galway, Ireland
E-mail: {paul.buitelaar, tobias.wunner}@deri.ie

T. Declerck
DFKI
Stuhlsatzenhausweg 3, Saarbrücken, Germany
E-mail: declerck@dfki.de

L. Hollink
Technical University of Delft
Mekelweg 4, Delft, Netherlands
E-mail: l.hollink@tudelft.nl

Keywords lexica · terminology · Semantic Web · Linked Data · ontologies

1 Introduction

Lexica and terminology databases form an essential part of many modern NLP systems and frequently consist of large amounts of highly detailed and well curated entries. Examples of such resources are the lexical semantic network WordNet (Fellbaum, 1998) and subcategorisation lexica such as COMLEX (Grishman et al., 1994) for English and Lefff (Sagot, 2010) for French. However, there is currently a great diversity of formats for representing these models. As such, the interchange of this data is challenging, requiring many inexact conversion programs leading to the creation of lexical "data silos." Current work on the Semantic Web, in particular that of the Linking Open Data project (Bizer et al., 2009), has focused on the challenge of using the Web to connect such "data silos" and allows for linking between different data sets.

Furthermore, by using such linking it is possible to re-use entries from a lexicon within another and to allow a third party to expand on an existing lexicon in a natural way. The standards for the Semantic Web from the W3C consortium do not only cover such linked open data, but also formats for representing ontologies. Furthermore, standards for rule systems are currently close to completion (Kifer, 2008). Moreover, the knowledge represented following Semantic Web and Linked Data representation formalisms is inherently language independent and would highly benefit from natural language interfaces provided by lexical resources serialised in Semantic Web standards. Such a semantic-lexicon interface represents an essential component in the scenario of the Semantic Web, since it will enable an appropriate exploitation of the available knowledge by end-user applications, which are frequently language-based. Ontologies and rule-based reasoning have been used as an integral part of state-of-the-art NLP systems, for example Kim et al. (2003), and as such it seems natural that any attempt to exchange lexica on the Semantic Web should use these extant technologies. In particular, while there exist many terminology resources, they rarely have sufficient semantic information to enable these resources to be used for challenging natural language processing tasks. Similarly, while there exist many large scale semantic resources, such as DBpedia (Auer et al., 2007), and in particular models of domain semantics such as the Gene Ontology (Ashburner et al., 2000), they are rarely connected to complex morpho-syntactic information.

We present a model, we call "*lemon*" (**Le**xicon **M**odel for **On**tologies), which is designed to represent lexical information about words and terms relative to an ontology, and by enabling such exchange of data on the Semantic Web. *lemon* is what we term an "ontology-lexicon", in that, following Buitelaar (2010), we use the ontology to provide a semantic framework and focus on the lexical information that needs to be stated to use the concepts in the ontology, instead of providing links such as hypernymy or synonymy, a principle we call "semantics by reference". We note here that *lemon* is not intended to be a collection of resources but instead a meta-model with which lexical resources can be exchanged by Semantic Web principles. We focus primarily on domain terminology, as ontologies generally refer to specific domains. However, *lemon* is not domain-specific and could also be used for more general tasks. The *lemon* model draws on research performed by the authors in the design of lexica for interfacing with ontologies, in particular that of the LexInfo (Buitelaar et al., 2009) and LIR (Montiel-Pondsoda et al., 2010) models. Like its predecessors LexInfo and LIR, *lemon* draws

from existing work on offline lexical resources, in particular from the Lexical Markup Framework (ISO-24613:2008) (Francopoulo et al., 2006) as well as the work that is currently being performed in using the Web to align lexical resources by the ISOcat (Kemps-Snijders et al., 2008) and OLiA projects (Chiarcos, 2010). The *lemon* model attempts to be a highly scalable format, in the sense that its modelling of lexical and linguistic information related to concepts in ontologies has to scale from very simple to quite complex lexical entries. In many ways, *lemon* is closely related to the work of the SKOS (Miles & Bechhofer, 2009) project, which attempts to model simple knowledge organisation systems such as thesauri, classification schemes and taxonomies on the Semantic Web. However, the model we propose differs from SKOS in that it is an independent and external model, intended to be published with arbitrary ontology-based conceptualisations, or any other type of knowledge organisation systems, in order to provide a richer description of the knowledge captured in those resources in one or several natural languages. As an RDF model, *lemon* allows powerful and novel representations, an example of which is the incorporation of data categories, such as part of speech, which has a potentially large number of values (ISOcat currently lists 115 values). *lemon* models this data category and its values as URIs, which means that each value of the property is unique and has clear ownership and extra information related to these values can be discovered by dereferencing this URI.

The *lemon* approach to modelling semantics in the lexicon is significantly different from the traditional word sense model used in WordNet and many other resources, which has come under criticism for its poor definition (Kilgarriff, 1997). Instead, *lemon* uses ontological entities, which we term "references", identified by URIs, which define the semantics of an element and further specify relations and axioms on the element. Hence, *lemon* has a clear separation between semantics and syntax. While enforcing this separation between syntax and semantics, it is important to state how syntactic structures correspond to semantics ones, stating in particular how syntactic arguments map to semantics predicates, and how pragmatic constraints can affect the meaning of a given word.

The rest of the paper is structured as follows. In section 2 we give a brief overview of the main standardisation initiatives for linguistic and terminological description that have inspired our work. We also provide a brief description of those models intended to interface ontologies we draw on, and of standards for interchanging linguistic and lexical information on the Web. Then, in section 3, we present the *lemon* model and provide several examples of linking possibilities provided by the model that contribute to the reuse of and interoperability with existing standards. Further, in section 4 we report on available tools that support the use of models instantiated from *lemon*. Finally, in section 5 we summarise the main benefits of the model and conclude the paper.

2 Foundations and Related Work

In this section we will briefly describe some of the extant models for the representation of lexica and work on establishing the correspondence between syntactic and semantic resources.

2.1 WordNet and FrameNet

WordNet (Fellbaum, 1998) is probably the most significant lexical database for English, in which word senses are organised in sets of semantic equivalents (so-called "synsets"). Over the years, the WordNet model has been applied to many other languages besides English, and as part of the EuroWordNet project (Vossen, 1998), many of these multilingual WordNets have been linked by means of an interlingual index, a set of core meanings assumed to exist in all languages. More recently, WordNet has been adapted to RDF and published on the Semantic Web as linked data (Van Assem et al., 2006). As WordNet aims to be a general lexicon for English, however, it does not contain many domain terms – although Vossen et al. (1999) have outlined how such could be added to the interlingual index. In addition, WordNet assumes a rather informal interpretation of its lexical-semantic relations (e.g. synonymy, antonymy, hypernymy and meronymy), and does not provide for sophisticated linguistic information. In fact, WordNet only models four parts of speech: nouns, verbs, adjectives and adverbs. As such, its data model is not easy to carry over to lexica with significantly different purposes.

FrameNet (Baker et al., 1998) is a hierarchically structured collection of prototypical situations (called "semantic frames") used to organise lexical units. Each frame has a set of slots or roles (called "frame elements") describing participants or props involved in a particular situation. In contrast to WordNet, semantic relations are expressed not between senses, but between frames and frame elements, and Scheffczyk et al. (2006) have shown how these can be represented in OWL and linked to ontologies like SUMO. As with WordNet, however, FrameNet is not intended to be a general lexicon model, and as such does not provide vocabulary for deeper linguistic description, nor a clear methodology how such could be integrated. In fact, FrameNet is mainly concerned with defining a repertoire of cognitively inspired linguistic frames and not to provide a general model for the ontology-lexicon interface.

2.2 LMF

The Lexical Markup Framework (LMF) is an ISO standard for representing lexica in XML/UML and is conceptually related to the way lexical information is related in *lemon*. The framework is also available as an OWL download, however as noted in Cimiano et al. (2011), this is not currently a valid OWL file, and is not available at a fixed dereferencable URI. It provides a framework for modelling and representing lexical objects, including morphological, syntactic, and semantic aspects. It was conceived with the purpose of providing a common model for the representation of electronic lexical resources in order to permit data exchange and interoperability. The description of an entry is very detailed and relies on previous standards for linguistic category description, namely ISO 12620 Data Categories or ISOcat (see section 2.5), thus making the data highly reusable. In this sense, the LMF standard has been conceived as a metamodel for representing the whole lexicon of a language, in which all possible senses of a word are accounted for. Instead, *lemon*'s purpose is to enrich the conceptualisation represented by the ontology, by means of a lexico-terminological layer.

A simple example of an LMF entry in RDF is given below in Turtle syntax (Beckett & Berners-Lee, 2008): a lexicon consisting of a single entry, a common noun with lemma "tax".

```
@prefix lmf: <http://www.tagmatica.fr/lmf#> .
```

```
:lexicon1 a lmf:Lexicon .

:entry1 a lmf:LexicalEntry ;
  lmf:isPartOf :lexicon1 ;
  lmf:isAdorned [ lmf:att "partOfSpeech" ;
                  lmf:val "commonNoun" ] .

:lemma1 a lmf:Lemma ;
  lmf:isPartOf :entry1 ;
  lmf:isAdorned [ lmf:att "writtenForm" ;
                  lmf:val "tax" ] .
```

The RDF/OWL version of LMF however uses only the properties `isAssociated`, `isPartOf` and `isAdorned`, and the size of the RDF models generated is very large due to the number of unnecessary elements introduced by the conversion to RDF. Moreover, lexical properties like "writtenForm" and data categories like "partOfSpeech" or "commonNoun" are hidden inside literal values. As such, the format does not exploit the full potential of RDF, and as a result, it is very difficult to query and work with lexica represented using this schema. *lemon* takes an RDF-native approach in using a different name for each property, and as such in *lemon* the above example can be written as follows. (Note that "lemma" approximately corresponds to "canonical form" in *lemon* and we specify the `xml:lang` special property on each string in *lemon*).

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

:lexicon lemon:entry :tax .

:tax lemon:canonicalForm [ lemon:writtenRep "tax"@en ] ;
     :partOfSpeech :commonNoun .
```

2.3 SKOS

SKOS (Simple Knowledge Organisation System) was developed as a system to provide a way to formalise many knowledge organisation systems, and share them on the Semantic Web: (from Miles & Bechhofer (2009))

> Different families of knowledge organisation systems, including thesauri, classification schemes, subject heading systems, and taxonomies are widely recognised and applied in both modern and traditional information systems. In practise it can be hard to draw an absolute distinction between thesauri and classification schemes or taxonomies, although some properties can be used to broadly characterise these different families. The important point for SKOS is that, in addition to their unique features, each of these families shares much in common, and can often be used in similar ways. *However, there is currently no widely deployed standard for representing these knowledge organisation systems as data and exchanging them between computer systems.* (emphasis added)

We focus on SKOS here as it is based on RDF and is the most widely used such format. However, it is also important to note that there exist other models for representing terminologies such as OTR (Reymont et al., 2007) and TBX (ISO 30042),

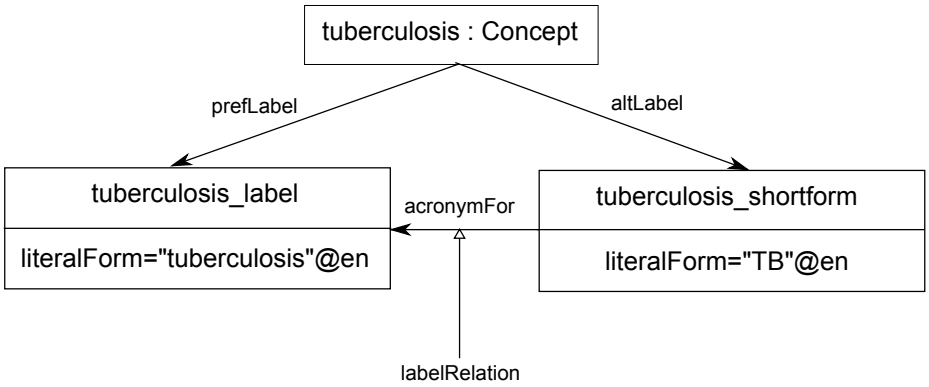

Fig. 1 Example of adding lexical information with SKOS-XL

which are based on Linked Data and XML standards. In many ways *lemon* aims to achieve similar goals to SKOS in making lexica available on the Semantic Web.

Within the SKOS use case (Isaac et al., 2009), it was identified that it is important for many of these knowledge organisation systems to have labels that relate to one another, for example to indicate one label as being the acronym of another. For this reason, an extension called SKOS-XL (SKOS eXtension for Labels) was introduced, in which the label property is 'reified', so that further properties of labels can be specified. In this context, a concept like tuberculosis with a label "tuberculosis" can be additionally associated to an alternative label "TB", and thus indicate that one is an acronym of the other.

```
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#> .

:tuberculosis skosxl:prefLabel :tuberculosis_label ;
              skosxl:altLabel :tuberculosis_shortform .
:tuberculosis_label skosxl:literalForm "tuberculosis"@en .
:tuberculosis_shortform skosxl:literalForm "TB"@en .
:tuberculosis_shortform dcr:acronymFor :tuberculosis_label .

dcr:acronymFor rdfs:subPropertyOf skosxl:labelRelation .
```

The main modelling decisions of *lemon* are based on SKOS's model, however we extend this by introducing a well defined "textual-conceptual" path, which is defined simply as *the number of nodes between the string literal value and the concept* (in the above example, illustrated in figure 1, `:tuberculosis`). In SKOS-XL an extra node is introduced between the text and the concept, and this allows for greater description of the labels. There is no clear principle for the intention of this node, and as such one of the key aspects of *lemon* is introducing a longer but more principled chain from a linguistic and lexical point of view, in particular differentiating between syntactic and terminological variation and clearly separating pragmatic and syntactic constraints.

2.4 LexInfo and LIR

As lemon follows the principle of "semantics by reference", no complex semantic information needs to be stated in the lexicon. Consequently, we build on our previous models to represent the interface between lexica and ontologies and in particular how syntactic information in the lexicon can be linked to semantic information in the ontology. In the LexInfo project we identified the key requirements of a lexicon-ontology model as follows:

1. **Separation of the lexicon and ontology**. There must be a clear separation of the lexical layer and the ontological layer, and lexica must be interchangeable (i.e., multiple lexica can describe the same ontology).
2. **Structured linguistic information**. It should be possible to represent linguistic descriptions, e.g., part of speech.
3. **Syntactic behaviour**. The model must represent the syntactic behaviour of its entries, e.g., valency of verbs.
4. **Morphological decomposition**. Terms can be represented as a decomposition of other terms.
5. **Arbitrary ontologies**. The lexicon should be reusable for different ontologies.

The LIR (Linguistic Information Repository) model (Montiel-Pondsoda et al., 2010) had similar goals to the first, second and fifth goals above, but in addition focused strongly on multilinguality. LIR provides mechanisms to establish links among lexical and terminological elements within and across different natural languages. Thus, LIR and LexInfo can be viewed as complementary models, however these models have both direct contradictions and shared flaws. We note that both models give a large number of categories for things such as a part of speech, and this may lead to difficulties in adapting them to non-European languages. There are also more specific problems, such as that many of the properties in LIR, such as distinction between scientific/layman terms, were introduced for a single domain. LexInfo on the other hand, due to the way it adapts LMF was very verbose and had a very large but incomplete set of subcategorisation frames. As such, we designed *lemon* as a model that avoided some of the issues in our previous models and would prove to be more suitable for interchange of lexica, by avoiding specific categories and keeping the model as concise as possible.

2.5 ISOcat, OLiA, GOLD

There have a been a number of attempts to enable the exchange of computer lexica and as described by Romary (2010), there is increasing convergence among the formats. One of the key challenges he identifies in developing an exchange lexicon is whether to give a specific model or general guidelines. In particular he notes that "the choice to provide an actual format, potentially facilitates immediate interoperability across applications, but bears the risk of not being flexible enough if some phenomena occur, that have not been anticipated in the standard." One of the key solutions to this issue is the idea of *data categories*, that aim to provide the following (again from Romary (2010)).

- A generic entry point and unique identifier for sharing concepts
- Fine grained information about a linguistic concept that may only be relevant to certain languages or resources

	Values	Properties
GOLD	506	83
ISOcat	1140	613
OLiA	595	45

Table 1 Size of existing resources for data categories of linguistic properties and values. Note that for ISOcat values are called "simple" DCs and properties "complex" DCs, for OLiA and GOLD values corresponds to the classes and individuals in the OWL files. All values valid June 2010.

The goals of data category projects can be seen to parallel those of the Semantic Web in particular as identified in Shadbolt et al. (2006). The first goal of data categories parallels the usage of dereferenceable URIs to identify resources on the Semantic Web, while the second goal seems similar to the creation of large scale RDF(S) taxonomies and OWL ontologies for describing particular domains.

There has been some work on bringing large scale data categories together on the Web. One of the first of these is the GOLD ontology (Farrar et al., 2003), which combines many of the most common lexical categories into a single large ontology. A similar but more recent project is that of ISOcat, that follows from the standardisation work of LMF. This is done primarily through a format called DCIF (ISO 12620). However, RDF versions of each data category are also published allowing for some interface with existing Semantic Web standards. Finally, OLiA (Chiarcos, 2010) is an ontology that derives from existing taxonomies of linguistic annotation and provides a core reference model that covers similar ground to GOLD and ISOcat. More interestingly, OLiA has annotation linking models that are used to describe alignment between the OLiA reference model and other annotation schemes (for example Penn Treebank tags). The OLiA project is also working on publishing links between the OLiA reference model and the GOLD and ISOcat models. The relative sizes of each resource are given in table 1.

3 The *lemon* model

In the light of current existing linguistic resource standards, we propose lemon as a model for exchanging lexicon resources on the Web with the following goals:

- LMF-like structure to enable conversion to existing offline formats (TBX, TEI, TIGER etc.).
- RDF-native form to enable leverage of existing Semantic Web technologies (SPARQL, OWL, RIF etc.).
- Separation of the lexicon vs. ontology layers, so that the semantic information and lexical information are well separated. This modularity enables straightforward exchange, addition, extension and substitution of lexica.
- The semantic inventory (ontology) is external to the lexicon model. Thus the model does not prescribe a representation of the meaning of entries and is open to any semantic distinction the user of the lexicon requires.
- Linking to data categories, in order to allow for arbitrarily complex linguistic description.
- A small model using the *principle of least power* - the less expressive the language, the more reusable the data (from Shadbolt et al. (2006)).

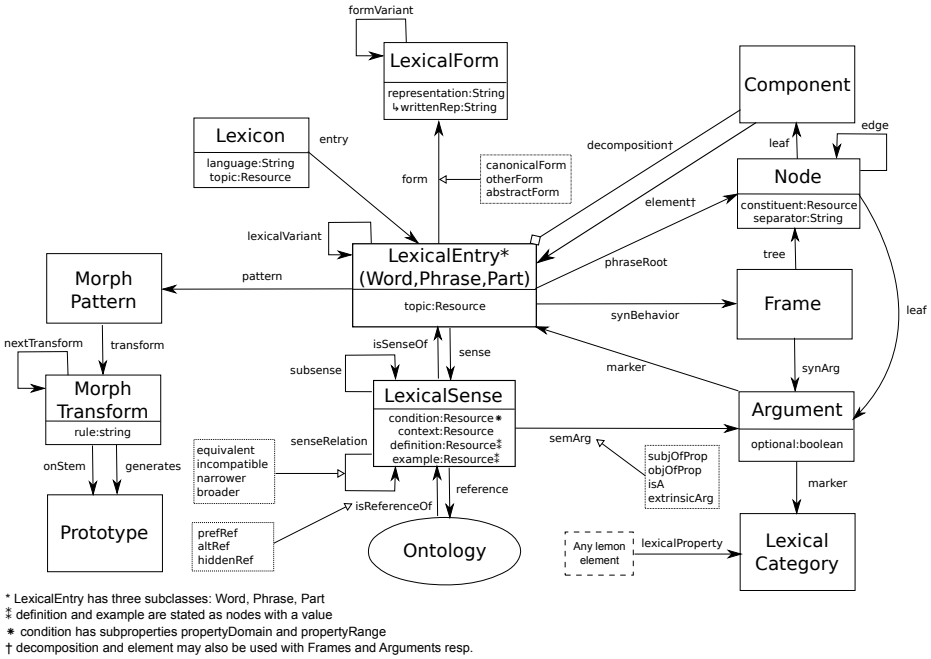

Fig. 2 The *lemon* model

The *lemon* model, as illustrated in figure 2, is available in RDF with extra OWL constraints at `http://www.monnet-project.eu/lemon`.

3.1 The core

The core of *lemon* covers the basic elements of a *lemon* lexicon, that is linking entries to particular lexical forms and to particular reference senses. This is done primarily by defining the following entities.

- **Lexicon:** The object representing the lexicon as a whole. This must be marked with a language, hence all lexicon objects in *lemon* are assumed to be monolingual.
- **Lexical Entry:** An entry in a lexicon is a container for one or several forms and one or several meanings of a lexeme. All forms of an entry must be realised with the same part of speech, and while an entry may have multiple meanings, homonyms are treated as separate lexical entries.
- **Lexical Form:** An inflectional form of an entry. The entry must have one canonical form and may have any number of other forms. It may also have abstract forms, which are intended for stems and other partial morphological units.
- **Representation:** A given lexical form may have several representations in different orthographies, for example a phonetic representation in addition to a standard written representation.
- **Lexical Sense:** A sense links the lexical entry to the **reference** used to describe its meaning, i.e, the ontology.
- **Component:** A lexical entry may also be broken up into a number of components.

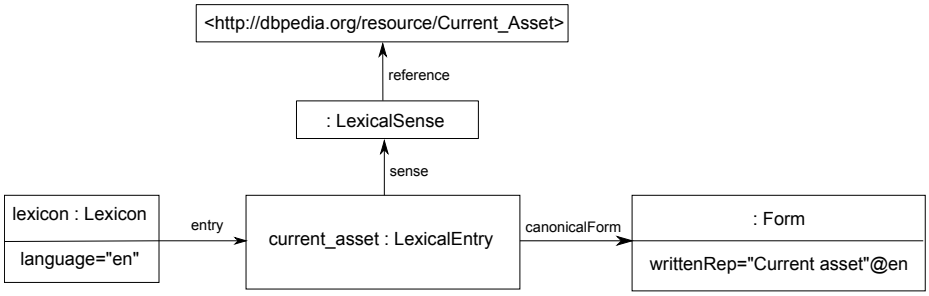

Fig. 3 A simple example of a *lemon* lexicon with a single entry "Current asset"

In this way we give a clearer textual-conceptual path than is possible with SKOS. The following example gives a simple lexicon with a single lexical entry as follows:

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

:lexicon lemon:entry :current_asset ;
   lemon:language "en" .

:current_asset
   lemon:canonicalForm [ lemon:writtenRep "Current asset"@en ] ;
   lemon:sense [ lemon:ref dbpedia:Current_asset ] .
```

In this example (illustrated in figure 3), we have an English lexicon with a single entry, with canonical form "Current asset", and a sense that refers to the entry in the Linked Data resource DBpedia (Auer et al., 2007), from which further semantic information about the entry can be obtained.

3.2 Linking to data categories

While the core is fairly useful for representing many aspects of lexical information, it is frequently necessary to include more information about morphology, syntax, terminological distinctions, versioning, authorship information etc. It would be very difficult to include all such categories in a way that would satisfy all users of the model and as such we do not wish to do this in the *lemon* model. However, the Semantic Web presents a solution to this, as we can link to such large collections of *data categories* by referencing their URIs. This is the approach taken by *lemon* and consequently, arbitrarily complex linguistic information can be included in the *lemon* model by referencing sources such as ISOcat, OLiA, GOLD for linguistic information, and vocabularies such as Dublin core[1] for authorship information. It is important to note that this does not solve the interoperability problem between category ontologies. This is not our goal, instead people may use whatever category system they want, at any granularity. However the use of a unique identifier and the ability to dereference these identifiers to find further information and restrictions on these properties should aid in aligning categories.

[1] See `http://dublincore.org/`

For example, we will show an entry for the Dutch feminine noun "vergunning" ("permit"), with plural form "vergunningen."[2]

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix isocat: <http://www.isocat.org/datcat/> .
@prefix dublincore: <http://purl.org/dc/elements/1.1/> .

:vergunning
   lemon:canonicalForm [ lemon:writtenRep "vergunning"@nl ;
                           # number=singular
                         isocat:DC-1298 isocat:DC-1387 ] ;
   lemon:altForm       [ lemon:writtenRep "vergunningen"@nl ;
                           # number=plural
                         isocat:DC-1298 isocat:DC-1354 ] ;
   isocat:DC-1345 isocat:DC-1333 ; # partOfSpeech=noun
   isocat:DC-1297 isocat:DC-1880 ; # gender=feminine
   dublincore:contributor "John McCrae" .

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

Here we use ISOcat URIs to reference each of the properties, so that extra information about the data category can be obtained by dereferencing this link. Each of these properties is also linked back into the *lemon* model by declaring them as subproperties of *lemon*'s `property`, so that the role of the property in the *lemon* model is defined, and as such the semantics of the property is completely defined. The use of URIs means that the specification of the linguistic category becomes unambiguous. Furthermore, the source and provenance as well as ownership and responsibility for the data category are clearly defined. In addition, we use the Dublin Core vocabulary to provide non-linguistic annotations, such as the author of the entry. We note here that the use of RDF for data categories may allow ontological relationships between data categories to be expressed. This is potentially an interesting direction and is explored further in McCrae et al. (2011),

3.3 Linking between lexica

One of the most interesting aspects of using RDF and Semantic Web standards is that there are possibilities of data re-use not available to static resources. For example the medical term "hospital-acquired pneumonia", is composed of the words "hospital", "acquired" and "pneumonia", and we can provide appropriate morpho-syntactic and terminological information for each of these entries. However, it is inefficient for every single lexicon to repeat non-domain-specific words like "acquired". Thus, we shall expand on our previous example to show how RDF can aid in data re-use.

```
@base <http://www.example.org/biomedical_lexicon> .
@prefix common: <http://www.example.org/common_lexicon#> .
```

[2] We reference ISOcat by the use of the data category number, and put a readable comment to each property. In the diagrams, we put only the readable description

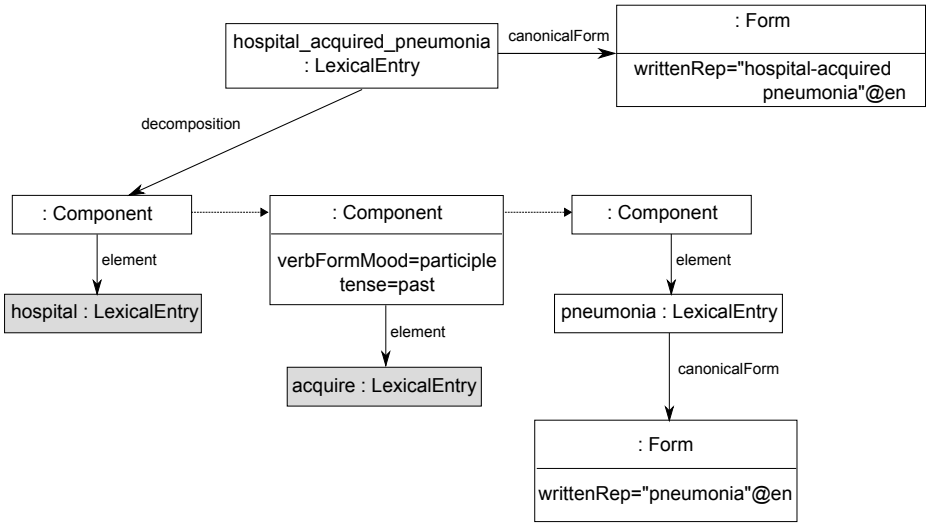

Fig. 4 Linking between lexica. The entries in white are part of the biomedical lexicon and the greyed entries are part of the general lexicon. Note "acquired" is modelled as the past participle of the verb "acquire"

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

:hospital_acquired_pneumonia
   lemon:canonicalForm
     [ lemon:writtenRep "hospital-acquired pneumonia"@en ] ;
   lemon:decomposition (
     [ lemon:element common:hospital ]
     [ lemon:element common:acquire ;
       isocat:DC-1427 isocat:DC-1341 ; # mood=participle
       isocat:DC-1286 isocat:DC-1347 ] # tense=past
     [ lemon:element :pneumonia ]
   ) .

:pneumonia
  lemon:canonicalForm [ lemon:writtenRep "pneumonia"@en ] .
```

In this example (illustrated in figure 4), we see that "hospital-acquired pneumonia" is stated as being composed of an ordered list of components each of which refers to a lexical entry.[3] Two of these entries have URIs in the "common lexicon" (identified by, for example, `http://www.example.org/common_lexicon#hospital`) and one in the same lexicon, the "biomedical lexicon" (identified by the URI `http://www.example.org/biomedical_lexicon`). As such, any extra information that is stated in the common lexicon about the entries is then automatically available for users of the domain

[3] We note that more precise modelling of the phrase structure of the term is possible using the *lemon* model and this is described further at `http://www.monnet-project.eu/docs/lemon-cookbook.pdf`

lexicon. Because these lexical entries are included by use of their URIs, they can be imported from any lexicon published on the Semantic Web, not just those controlled by the same author. This has the advantage that if the lexical entries are updated, the lexicon importing it will also automatically update these changes, a clear benefit of referencing in contrast to static import or duplication.

3.4 Lexicon-Ontology Mapping

The *lemon* model does not intend to be a semantic model, but instead it allows semantics to be represented by referencing extant semantic resources, in particular ontologies. The *lemon* model approaches this by means of its "(lexical) sense" object, which differs significantly from the concept of a word sense found in existing models such as WordNet. Technically, the sense object is unique for every pair of lexical entry and reference, i.e., the sense refers to a single ontology entity and a single lexical entry. Thus, each word has a different sense for each distinct reference. In fact, a sense may have multiple reference URI values, but this infers that the reference URIs represent ontologically equivalent entities[4]. The sense object in lemon plays three roles: first, the set of all senses define a many-to-many mapping between lexical entries and ontological entities. This models the fact that lexical entries can have different meanings with respect to a given ontology and the fact that ontology elements can be verbalised linguistically in various ways. Second, the sense object represent the (lexical) meaning of the lexical entry when interpreted as the given ontological concept. Third, the sense also represents an ontological specialisation of the referenced ontology entity which accounts for the specific lexico-semantic connotations that the lexical entry introduces

As this relationship does not state that the meaning of the entity and the lexicalisation are equivalent, but rather indicates that there are some times when this lexical entry is used with this meaning and conversely this entity may sometimes be lexicalised with this entry. Therefore, it follows that the sense object belongs neither truly to the lexicon nor the ontology but instead acts as a bridge between the two and represents an *underspecified* relationship in that it represents only those uses where the given lexical entry is used to refer to the given ontology element. This mapping can be further specified by a number of contexts when the given lexicalisation corresponds to an ontology entity. Words may have different meanings based on the register they are used in, and any conditions on the usage of a particular word, for example the use of "fressen" and "essen" in German to indicate eating by animals and humans respectively.

One of the other key aspects of this modelling is to state a correspondence between the syntax and the predicates within the ontology, that is between the arguments given by the valency of the verb and the subject/object of properties.The *lemon* model represents subcategorisation with a frame object that can have a number of syntactic arguments indicated with the `synArg` property, which may be sub-typed to indicate specific roles played by syntactic arguments. The link to the ontology is then represented by linking the sense to each of these arguments with `subjOfProp`, `objOfProp` and `isA` (used for classes which we model as unary predicates). An example of such a

[4] For example, s `lemon:reference` x_1 , s `lemon:reference` $x_2 \vdash x_1$ `owl:sameAs` x_2

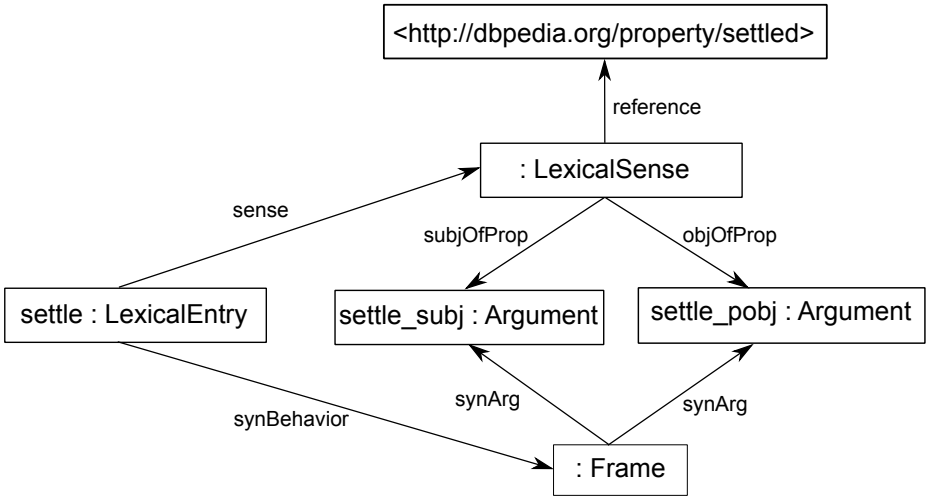

Fig. 5 Linking a verb's subcategorisation to an ontology property

mapping for the subcategorisation "X was settled in Y" is as follows, where "X" is the subject entity and "Y" the object entity.[5]

```
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix dbpedia: <http://dbpedia.org/property/> .
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .

:settle lemon:syntacticBehavior [
       lexinfo:subject :settle_subj ;
       lexinfo:prepositionalObject :settle_pobj ] ;
     lemon:sense [
       lemon:reference dbpedia:settled ;
       lemon:subjOfProp :settle_subj ;
       lemon:objOfProp :settle_pobj ] .
```

This example (illustrated in figure 5), shows how we define a subcategorisation frame for a verb, in this case by indicating its arguments with ISOcat data categories that are specified as sub-properties of `synArg`. These arguments are then also linked to the sense, and indicated as the subject and object of the property referred to by this sense. In this way we can precisely describe the correspondence between a lexical entry and an ontology property or class.

These examples cover only a small part of the model, a full technical manual is available at `http://www.monnet-project.eu/docs/lemon-cookbook.pdf`, which also covers other features of the *lemon* model including:

- Mapping with ternary (e.g., "donative") and other higher order subcategorisations
- Relations between lexical entries

[5] Here we use our *lemon*-aligned version of LexInfo, as ISOcat does not currently have many data categories for subcategorisation. Note, it is not necessary to state these properties as subproperties of *lemon*, as they are already published as such.

– Representing syntax trees
– Combining syntax trees with subcategorisations
– Specifying sense contexts and conditions
– Assigning lexica and entries topics ("subject fields")
– Asserting global lexicon constraints
– Providing compact representations of inflection and agglutination

4 Using *lemon*

In general, the most important step for instantiating a *lemon* model is identifying the sets of data categories that we wish to (re-) use in the model. Unlike other formats, there is no need to create a data category selection file to state which set of data categories are used in a given file. Instead as each data category is uniquely identified by a URI, they can simply be used without prior identification. As such, in order to use *lemon* to represent a lexicon, the following steps should be carried out:

1. Identify which properties/relations you wish to use to define specific linguistic concepts.
2. Look up appropriate data categories from some source (e.g., ISOcat) and include them in the lexicon by stating them as subproperties of the appropriate *lemon* property.
3. If there are properties that are not covered by any standardised source, you may define them yourself. The URIs of the properties should be dereferenceable, i.e., an RDF description of it should be available at the address.
4. Align the data source with the *lemon* core model. For example it is commonly necessary to identify how canonical and alternative labels are identified in the source.
5. Publish the lexicon as an RDF/XML document. The URIs for each entry should be resolvable at the given URI.

We have also created a number of tools to support the use of *lemon* models. Firstly, as *lemon* is developed from LMF we have a method to convert the models to and from the LMF format; this is available at `http://www.lexinfo.net/lemon2lmf`. We are also working on import/export facilities to a number of other formats including XLIFF and TBX. We have also developed a web interface that allows people to upload and modify *lemon* models, this is available at `http://monnetproject.deri.ie/lemonsource`. This service can also create *lemon* models automatically from OWL ontology files. This works by extracting the labels for each concept from the ontology through an annotation such as `rdfs:label` or `skos:prefLabel`. Otherwise the system uses the URI of the entity to attempt to obtain a label for the concept, for example by de-camel-casing the fragment. Then, the system applies a tokeniser and then a part-of-speech tagger and uses this to create the core structure of the *lemon* entry. Finally, as in the work of Cimiano et al. (2011), we apply syntactic analysis to deduce the subcategorisation frame of the term and the phrase structure (if desired).

Another important aspects of *lemon* is that it is based on established Semantic Web technologies and hence a number of tools already exists to enable the sharing and integration of models on the Web. For example, Sindice[6] maintains an index of

[6] `http://sindice.com`

all RDF data published on the Semantic Web. As such, if someone chooses to publish their *lemon* lexicon on the Web, it can be submitted to Sindice. It is then easy for other users to find lexica and share them, as Sindice allows for particular properties to be queried. For example, querying for all triples using the property `lemon:writtenRep` and value `"cat"` would find all *lemon* lexica that use the word "cat."

5 Conclusion

The *lemon* RDF model allows for lexical data to be shared and interlinked on the Web, allowing for greater reuse of existing data than is possible using current lexicon formats. The *lemon* model is based on several existing resources, in particular LMF, SKOS, LIR, and LexInfo and as such maintains a high degree of compatibility with these models. However, its focus on compactness and expressivity allows for a large amount of linguistic information to be represented, while keeping the models quite small. It maintains a high degree of flexibility and extensibility by the use of data categories, allowing the model to act as a lexicon meta-model as well as a format in its own right. We have also discussed tools that facilitate easy usage of the *lemon* model and interaction with existing standards in both lexicography and the Semantic Web. As such we hope that this will lead to a consensus model for the exchange of lexica on the Semantic Web and we are working towards building a community that can continue to develop and apply the model.

Acknowledgements The *lemon* model and associated tools have been developed in the context of the Monnet project, which is funded by the European Union FP7 program under grant number 248458 and the CITEC excellence initiative funded by the European Union and the DFG (Deutsche Forschungsgemeinschaft). We would like to thank the following people for their contributions and advice: Axel Polleres (DERI), Antoine Zimmermann (DERI), Dimitra Anastasiou (CNGL), Susan Marie Thomas (SAP), Christina Unger (CITEC), Sue Ellen Wright (Kent State University), Menzo Windhouwer (Universiteit van Amsterdam).

References

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25(1):25

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2007) Dbpedia: A nucleus for a web of open data. In: Proceedings of the 6th International Semantic Web Conference

Baker C, Fillmore C, Lowe J (1998) The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL)

Beckett D, Berners-Lee T (2008) Turtle - Terse RDF Triple Language. URL http://www.w3.org/TeamSubmission/turtle/, Accessed 19 October 2010

Bizer C, Heath T, Berners-Lee T (2009) Linked data - the story so far. International Journal on Semantic Web and Information Systems 5:1–22

Buitelaar P (2010) Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. In: Ontology and the Lexicon, Cambridge University Press, pp 212–223

Buitelaar P, Cimiano P, Haase P, Sintek M (2009) Towards linguistically grounded ontologies. In: The Semantic Web: Research and Applications, pp 111–125

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)

Cimiano P, Buitelaar P, McCrae J, Sintek M (2011) LexInfo: A Declarative Model for the Lexicon-Ontology Interface. Journal of Web Semantics 9(1):29–51

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

Francopoulo G, George M, Calzolari N, Monachini M, Bel N, Pet M, Soria C (2006) Lexical markup framework (LMF). In: Proceedings of the 2006 International Conference on Language Resource and Evaluation (LREC)

Grishman R, Macleod C, Meyers A (1994) COMLEX syntax: Building a computational lexicon. In: Proceedings of the 15th International Conference on Computational Linguistics (COLING)

Isaac A, Phipps J, Rubin D (2009) SKOS Use Cases and Requirements. URL http://www.w3.org/TR/2009/NOTE-skos-ucr-20090818/, Accessed 19 October 2010

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)

Kifer M (2008) Rule interchange format: The framework. In: Proceedings of the 2nd International Conference on Web Reasoning and Rule Systems

Kilgarriff A (1997) I Dont Believe in Word Senses. Computers and the Humanities 31(2):91–113

Kim J, Ohta T, Tateisi Y, Tsujii J (2003) GENIA corpus-a semantically annotated corpus for bio-textmining. Bioinformatics 19(1):180–182

McCrae J, Spohr D, Cimiano, P (2011) Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC-11)

Miles A, Bechhofer S (2009) SKOS Simple Knowledge Organization System Reference. URL http://www.w3.org/TR/skos-reference/, Accessed 19 October 2010

Montiel-Ponsoda E, Aguado de Cea G, Gómez Pérez A, Peters W (2010) Enriching Ontologies with Multilingual Information. Natural Language Engineering, Available on CJO 2010 doi:10.1017/S1351324910000082

Reymonet A, Thomas J, Aussenac-Gilles N (2007) Modelling ontological and terminological resources in OWL-DL. In: Proceedings of the 6th International Semantic Web Conference (ISWC)

Romary L (2010) Standardization of the formal representation of lexical information for NLP. In: Dictionaries: An International Encyclopedia of Lexicography., Mouton de Gruyter, URL http://arxiv.org/abs/0911.5116v1

Sagot B (2010) 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)

Shadbolt N, Hall W, Berners-Lee T (2006) The semantic web revisited. IEEE intelligent systems 21(3):96–101

Scheffczyk J, Pease A, Ellsworth M (2006) Linking FrameNet to the Suggested Upper Merged Ontology. In: Formal Ontology in Information Systems (FOIS-2006), pp 289–300

Van Assem M, Gangemi A, Schreiber G (2006) Conversion of WordNet to a standard RDF/OWL representation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)

Vossen P (1998) EuroWordNet: a multilingual database with lexical semantic networks. Computational Linguistics 25(4):628-630

Vossen P, Bloksma L, Peters W, Kunze C, Wagner A, Pala K, Vider K, Bertagna F (1999) Extending the Inter-Lingual-Index with new Concepts. Deliverable 2D010, EuroWordNet, LE2-4003