Undefined 0 (2012) 1–0
IOS Press

Converting the Parole Simple Clips Lexicon into RDF using the Lemon model

Riccardo Del Gratta, Francesca Frontini, Fahad Khan, Monica Monachini [a,*]
[a] *Istituto Di Linguistica Computazionale 'A. Zampolli' - Consiglio Nazionale delle Ricerche, Via Moruzzi 1 Pisa, Italy*
E-mail: first.last@ilc.cnr.it

Abstract. This paper reports on the publication of parts of an Italian Lexicon Parole Simple Clips (PSC) as linked data using the Lemon linked data model. The main problem dealt with during the conversion process related to the mismatch between the Lemon view of lexical sense objects, as a reified pairing of a lexical item and the concept in an ontology that provides a meaning for it, and the corresponding notion within PSC of semantic units, which can take part in a broader class of semantic relations. The solution outlined in this paper was to instantiate the semantic units of PSC semantic layer as units in an OWL ontology called Simple OWL along with the relations holding between them in the PSC semantic layer, and also to duplicate them as Lemon lexical sense objects. The details of the solution and the organisation of the data are given in the paper, as well as a discussion of possible improvements and future work.

Keywords: Lemon Model, Linked Data, Generative Lexicon, RDF, OWL, Lexical Resource

1. Introduction

The central aim of the linked data movement is to make it easier to use and to share collections of data distributed at various locations across the web by setting up a standardized way of structuring, describing, and interlinking this data [1]. In the linked data model, data is formatted according to the Resource Description Framework (RDF)[1] and is therefore structured in the form of *subject-predicate-object* triples. These triples are used to link together data items using their so called Unique Resource IdentifiersUnique Resources Identifiers (URIs) making it far simpler to use description standards and formats such as Ontology Web Language (OWL)[2] for organising and reasoning about such distributed, interlinked data. The Language Resources and Technology (LRT) community is becoming increasingly active within the linked data movement. This is the result of a greater awareness of the opportunities that linked data offers for setting up the kind of general Language Resources and Technology infrastructure variously described in the Language Resources and Technology literature as the Lexical Web [2] and as a lexical linked space [4]. Language Resources and Technology research has traditionally put great emphasis on the standardisation, linking, and reusability of Lexical Resources and the linked data movement makes it far easier to achieve these core aims. This increased realisation of the importance of linked data within the LRT community, has resulted, in the specific context of computational lexicons, in a trend towards the conversion of existing lexica and the output of lexicon creation tools to the RDF format. It is then much more straightforward to connect lexicons to other relevant data resources such as web-based ontologies. The work described in this paper had as an aim the conversion of a subset of the lexical items in a large-scale, multi-layered Italian language lexicon Parole Simple Clips (PSC), namely all of the nouns in the lexicon, into the RDF format, using

*Corresponding author. E-mail: riccardo.delgratta@ilc.cnr.it.
[1] http://www.w3.org/RDF/
[2] http://www.w3.org/OWL/

the Lemon lexicon description model. This process entailed the full conversion of the semantic layer of the lexicon into OWL, as well as the creation of a new lexicon using the Lemon model containing all the nouns in the Parole Simple Clips lexicon and the subsequent linking of these two resources using the Lemon model. These different stages are described below.

2. Lexical Ontologies with Lemon

Lemon [6] is a descriptive model that supports the linking up of a computational lexical resource with the semantic information stored in one or more ontologies, as well as enabling the publishing of such lexical resources on the web. At its heart, Lemon defines a set of core modules that serve to describe the basic aspects of the entries in most lexicons such as for example the phrase structure of complex expressions, or the syntactic frames associated with a verb. At the same time, in Lemon each lexical entry l of a lexicon is mapped onto a concept c in an ontology that provides reference or meaning for l via a mediating lexical sense object $\sigma^{l,c}$. This lexical sense object $\sigma^{l,c}$ is best understood as the *reification* of the pairing (l, c) of the lexical entry l with the ontological concept c. Within the Lemon model, in fact, each lexical sense object can be seen either as a subset of the uses of the entry l in a collection of sentences where l has the meaning c, or as representing a hypothetical concept that captures the full lexical meaning of an entry, see [3] for full details.
An important advantage of making such a clear separation between linguistic and ontological levels is that it allows a clearer understanding of the effect that the granularity of the conceptual distinctions within an ontology has on this relationship between a lexical entry and its reference. It ensures that the reference or meaning of a lexical term l can be seen as something contextual to the specific needs of the lexicon ontology mapping so that the reference c does not even need to account for the full meaning. Thus if we are mapping a lexicon to a urban planning ontology, both synagogue and mosque could be mapped onto the ontological type ReligiousBuilding each via a separate lexical sense where, as mentioned above, this lexical sense is viewed "a 'hypothetical' concept that, if added to the ontology, would be a subclass of the evoked concept" [3].

3. Parole Simple Clips and Simple OWL

Parole Simple Clips (PSC) is a multi-layered Italian language lexicon that was built in successive stages within the framework of three major lexical resource projects. Parole [8] and Simple[5] were two consecutive European projects which resulted in the creation of a wide ranging Italian language lexicon (as well as similar lexicons in 11 other European languages) structured into different, interconnected layers; CLIPS[3] was an Italian national project which enlarged and refined the Italian Parole-Simple lexicon.
The lexical information in PSC is encoded at different descriptive levels. These are the phonetic, morphological, syntactic and semantic layers, each one being built up in units, with one-to-one, one-to-many, or many-to-one mappings interconnecting the different layers. The semantic layer of PSC includes a language independent ontology of 153 semantic types as well as $\sim 60k$ so called "semantic units" or *USems* representing the meaning of the items in this layer. From a theoretical point of view the semantics of PSC is based on Pustejovsky's Generative Lexicon (GL) theory [7]. In GL theory, the meaning of each lexical entry is structured into components, one of which, the *qualia structure*, consists of a bundle of four orthogonal dimensions, allowing for the encoding of four different aspects of the sense: the formal, namely the dimension which allows the identification of an entity, i.e., what it is; the constitutive, what an entity is made of; the telic, that which specifies the function of an entity; and finally the agentive, which specifies the origin of an entity[4].
The formal aspect of a sense allows it to be placed in a hierarchical, hyponymic relation with other senses, in the PSC ontology this sense aspect are represented by so called simple types; the more complex, multidimensional aspects of a sense are represented by unified types. Here is a small example of the relations between Semantic Units within PSC. Lexical entry *libro* (book) has two USems. USem4046libro points to the semantic type Semiotic_artifact and USem4047libro points to the type Information.
Both USems have several relations to other USems[5]:

[3]CLIPS stands for Corpora e Lessici dell'Italiano Parlato e Scritto

[4]The PSC ontology is based on the notion of an *extended qualia structure*, which as the name suggests is an extension of the qualia structure notion found in GL.

[5]Here and after, we simplify things by representing each of the two Usems in the example using only the number part of the name.

```
4047libro,Information,Isa,D5496testo
4046libro,Semiotic_artifact,Contains,D5496testo
```

Fig. 1. How Usems partecipate in semantic relations.

The Simple OWL project [9] involved the extraction of all the semantic types (e.g.,"Entity", "Semiotic_artifact" and "Information") from PSC, as well as the relations between semantic units (e.g., "isa" and "contains"), the features associated with semantic units (e.g., "PLUS_EDIBLE"), along with a number of well-formedness constraints from PSC in order to construct a consistent OWL ontology, Simple OWL. Simple OWL is a multidimensional ontology, in that each relation can be mapped onto one of the dimensions of the GL *qualia structure*. Thus the "isa" relation is mapped under the Formal axis, while the "Contains" relation is mapped under the "Constitutive" one.

4. Converting the Parole Simple Clips Lexicon into Lemon and linking to Simple OWL

Having decided to convert at least a subset of the PSC lexicon into Lemon, the challenge then arose of how to model the link between the morphologic and semantic layers of PSC. As described above the Lemon model requires a lexical sense object, seen as a *reified* lexical-semantic pairing, to mediate between a lexical entry and the concept or meaning of that entry as provided in an ontology. Unfortunately this particular distinction, has not been observed (or at least not observed systematically in the same way) in the mappings between the PSC lexical and semantic objects: the lexical sense objects in Lemon don't take part in semantic relations, but the Usems in PSC do, as in the example reported in figure 1. This means that it is not always possible to identify PSC Usems with lexical sense objects in Lemon. This becomes evident when one comes to consider the restriction on the kinds of relations that can hold between lexical sense objects in the Lemon model. At the same time in the lemon model certain properties such as *synonymy* and *antonymy* should only occur at the level of senses and not between the concepts that they point to in an ontology, since in the Lemon philosophy these distinctions should not affect the conceptualisation of the domain in an ontology.

On the other hand in the PSC semantic layer, semantic units can be related both by using properties corresponding to what would be regarded as sense relations such as *synonymy* and *polysemy*, as well as those that apply at a more conceptual, ontological level. For example, the lexicon entry gladiolo_N representing the Italian noun "gladiolo", translated in English as gladiolus in the singular, or as gladioli in the plural, points to two different Usems, one of which, 1617, represents the gladiolo as a flower (as in "*Ho comprato un mazzo di gladioli rosa dal fioraio*", "I bought a bouquet of pink gladioli from the flower seller"), so that the `rdf:type` of 1617 is `simple:Flower`[6] and the other, 1616, represents the gladiolo as a plant (as in "*Il gladiolo fiorisce una sola volta*", "The Gladiolus flowers only once") so that the `rdf:type` of 1616 is `simple:Plant`. Between the two Usems 1616 and 1617, we have both the conceptual relation that "1616 produces 1617", i.e., that the gladiolo (as) plant produces the gladiolo (as) flower, and the "sense" property that there is a systematic polysemy relation between them.

The solution implemented as part of the work described in this paper was to convert the semantic units in PSC into corresponding objects in the Simple-OWL lexicon. These semantic unit objects were then linked to their associated semantic types and the relations between these objects corresponding to those in the PSC lexicon were also implemented. This process served to complete the conversion of the PSC semantic layer into an OWL ontology that had begun with the Simple OWL project. The Usems of PSC, having been transferred into individuals in the Simple-OWL ontology, had to be then distinguished from the lexical sense objects of Lemon.

A new lexical resource, *pscLemon*, was created using the noun lexical entries from the Parole Simple Clips. These noun entries became the lexical entries of *pscLemon*. Each such noun entry n in PSC points to one or more Usems, u_1^n,...,u_k^n, in the semantic layer of PSC. These Usems were now represented by individuals $o(u_1^n), ..., o(u_k^n)$ in the extended Simple OWL ontology. A new lexical sense object was then created for each one of these Usems, namely, $s(u_1^n)$,...,$s(u_k^n)$, and the lexical entry corresponding to n in the PSC_lexicon was linked to each one of these lexical sense objects using the `lemon:sense` relation. These sense objects was linked in turn to the corresponding individual in the Simple OWL ontology using the `lemon:reference` relation. In effect each of the semantic units in the PSC semantic layer was

[6] For the definition of the namespaces, see 5.1

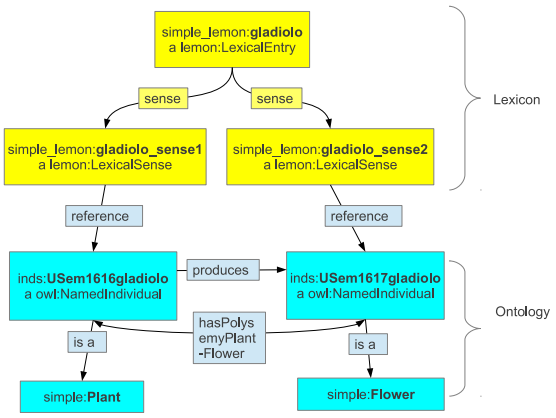

Fig. 2. Schema of the example.

copied twice: in the first case as an ontology individual, preserving all its its PSC relations and properties, and secondly as an object that provided an intermediary between the lexical item and the ontology individual. This was felt as being necessary in order to bring the lexicon into line with the Lemon description model. Figure 2 reports the duplication of the Usems and their use in the Lemon description model and in the Simple OWL ontology.

5. Structure of the data and distribution

The whole dataset produced for this paper is available under the Data Hub catalogue[7,8]. All the resources which belong to the dataset are licensed with a " Open Data Commons Attribution License"[9].
The dataset consists of the following three resources:

- the *SimpleOntology* ontology. This is the Simple-OWL ontology as described in [9];
- the *SimpleEntries* ontology. This is the ontology which collects the semantic units of PSC promoted as individuals, `namedIndividuals`; *SimpleEntries* imports *SimpleOntology* so that all semantic relations and properties which were defined in the PSC semantic level are preserved and resolved;
- the *pscLemon* lexicon. This is the lexicon which is conformal to the Lemon model. In this lexicon, the collection of the PSC lexical entries are `lemon:LexicalEntries`.

5.1. Namespaces

The three resources listed in section 5 are hosted under the common namespace:
`http://www.languagelibrary.eu/owl/simple` hereafter `base`.
The three resources have been stored under sub-folders of `base`, according to their specific content. Table 1 gives the resources, their namespaces and URIs.

Ontology	namespace	URI
SimpleOntology	base	base/SimpleOntology
SimpleEntries	base/inds	base/inds/SimpleEntries
pscLemon	base/psc	base/psc/pscLemon

Table 1
Resources, namespaces and URIs

Using the above namepaces, the example shown in Figure 2 can also be described using turtle[10] syntax, as shown in Figure 3.

```
@prefix psc_lemon:base/psc/pscLemon
@prefix inds:base/inds/SimpleEntries
@prefix simple:base/SimpleOntology
psc_lemon:gladiolo
lexinfo:partOfSpeech lexinfo:noun ;
lemon:canonicalForm [
lemon:writtenRep "gladiolo"@it ;
a lemon:Form
] ;
lemon:sense psc_lemon:gladiolo_sense1;
lemon:sense psc_lemon:gladiolo_sense2;
a lemon:LexicalEntry .
psc_lemon:gladiolo_sense1;
lemon:reference inds:USem1616gladiolo ;
a lemon:LexicalSense .
psc_lemon:gladiolo_sense2;
lemon:reference inds:USem1617gladiolo ;
a lemon:LexicalSense .
inds:USem1616gladiolo
simple:hasIsa inds:USem1428pianta ;
simple:hasPolysemyPlant-Flower inds:USem1617gladiolo ;
simple:hasProduces inds:USem1617gladiolo ;
a simple:Plant, owl:NamedIndividual ;
inds:USem1617gladiolo
simple:hasIsa inds:USemD2389fiore ;
simple:hasPolysemyPlant-Flower inds:USem1616gladiolo ;
simple:hasProducedby inds:USem1616gladiolo ;
a simple:Flower, owl:NamedIndividual ;
```

Fig. 3. Example of encoding.

Figure 3 shows how the polysemous lexical entry "gladiolo" is represented. The lexical entry and the two lexical senses are implemented according to the Lemon model in the *pscLemon* lexicon: in fact

[7] http://www.datahub.io/dataset/simple
[8] See http://www.datahub.io/about for more information on the Data Hub project.
[9] http://www.opendefinition.org/licenses/odc-by
[10] www.w3.org/TR/turtle/

their namespace is `psc_lemon` which is resolved to `base/psc/pscLemon`. The Lemon LexicalSense objects each refer to an individual in the `inds` namespace which points to a single entry in the *SimpleEntries* resource; finally each referenced individual is linked to the *SimpleOntology* resource through the `simple` namespace.

5.2. Data figures and Obtained Triples

As explained in Section 2, so far, in the work described in this paper, only the nouns have been extracted from the Parole Simple Clips lexicon. The number of entries that have been processed are 31232 Usems, out of an original total of $\sim\ 60k$, corresponding to 18610 lexical entries. Once processed, the data provided a different number of effective *subject-predicate-object* triples, as shown in Table 2:

file	Original Units	Triples
SimpleOntology	153	6332
SimpleEntries	31232	105674
pscLemon	18610	138610

Table 2
Files, units and triples

5.3. Data Organization

The *SimpleEntries* and *pscLemon* resources are not very usable, since they collect all entries in the same file[11]. Users are required to download all the resources and post-process the data to extract the information they need. It was therefore decided to organize the data according to the Linked Data paradigm: in such a way that each single entry in both resources points to a different file. For example there is a file *gladiolo* which contains the Lemon lexical entry for "gladiolo"; and there are two distinct files (one file for each lexical sense of "gladiolo"): "Usem1616gladiolo" and "Usem1617gladiolo".
A file system structure was created based on the first characters of the hash coding of the lexical entry, for example $gladiolo \rightarrow f/f6c$. Under the namespaces `base/psc` and `base/inds` were added structures similar to $f/f6c$, single entry files were also inserted, as shown in Figure 4:

```
base/psc/
  f/f6c/
    gladiolo
base/inds/
  f/f6c/
    Usem1616gladiolo
    Usem1617gladiolo
```

Fig. 4. Example of folder structures

It will be noticed that different lexical entries can generate the same folder structure, for instance $f/fe9$, is generated by both "famiglia" and "zolfo", so that, under the same folder unrelated lexical entries can be found. However this is not a problem since the structure is coherent through all namespaces:

```
base/psc/
  f/fe9/
    famiglia
    zolfo
base/inds/
  f/fe9/
    USemD5487famiglia
    USem3427zolfo
```

Fig. 5. Example of folder structure with multiple entries

As a consequence, all entries from *SimpleEntries* and *pscLemon* were extracted and two resources were added to the provided dataset, see the table below.

Resource	Comment
base/inds/SimpleListOfEntries	Contains the list of Usems along with their physical location
base/psc/pscLemonListOfEntries	Contains the list of Lemon entries along with their physical location

Table 3
Additional Resources

6. Conclusion and future work

The solution presented above seems to go a large part of the way towards reconciling the Lemon philosophy of separating the lexical and ontological layers of lexico-semantic resources with the representation of the multiple dimensions of meaning instantiated by the semantic layer of Parole Simple Clips. This differentiates the present solution from other ways of using Lemon to represent the Parole-Simple lexicon, in particular those which do not treat a lexical sense object as being solely a relation between a lexical entry and a meaning, but instead implement Parole-Simple semantic relations directly among lexical sense objects without reference to an external ontology. See Fig-

[11] In fact all references are prefixed by #, meaning that each entry is referenced with an inner anchor.

ure 6 which gives an example from the Spanish Parole/Simple dataset (http://datahub.io/it/dataset/parole-simple-ont).

```
lex:libro_NOUN a lemon:LexicalEntry ;
rdfs:label "libro"@es ;
lemon:form [ lemon:writtenRep "libro"@es ] ;
parole:id "UMNO035656" ;
parole:attestation "simple" ;
lexinfo:partOfSpeech lexinfo:commonNoun ;
.
lex:libro_NOUN lemon:sense lex:libro_Semioticartifact .
lex:libro_Semioticartifact parole:countability
parole:CountableNoun .
lex:libro_Semioticartifact a lemon:LexicalSense ;
parole:id "libro_Semioticartifact" ;
lemon:example [ lemon:value "e.g., libro
(Conjunto de hojas manuscritas
o impresas ordenadas para su lectura y
reunidas formando un volumen; Me comprÃl'
un libro sobre perros)" ] ;
parole:template parole:TemplSemioticartifact ;
parole:semanticClass parole:SemClassARTIFACT ;
parole:semanticFeature parole:SemanticFeatureSEMIOTIC ;
parole:semRelationCreatedby lex:hacer_X ;
parole:semRelationUsedfor lex:comunicar_SpeechAct ;
```

Fig. 6. Example of encoding of Spanish Parole/Simple lexicon

There remains the difficulty however that there exist relations within our lexical-semantic resource, such as those denoting polysemy and synonymy, that hold between conceptual objects within the Simple OWL ontology but which according to the lemon understanding of the link between lexica and ontologies should hold between Lemon lexical sense objects. Future work could explore the possibility of completely reconciling Parole Simple Clips with the Lemon model while fully preserving the semantic relations within the PSC.

There are also a number of other important issues to consider.The Usems that have been turned into individuals could actually be promoted to types; thus "UsemCane1" could be promoted to the class of all individuals that are dogs. In order to do this it would be necessary to collapse those individuals in the Simple OWL ontology between which synonymy holds The new elements that have been added to the ontology are labeled in Italian; the top ontology instead has English labels and can be mapped by all of the European language projects falling under the Simple banner. Thus some mapping is require to merge "UsemCane1" and "Usemperro1" when they represent the same class. With this approach lexical semantic relations are promoted to ontological relations; relations that were stated in Simple as being templates among senses are now represented as relations among their referents. Other future work could be to import further elements of the morpho-syntactic information in PSC into the Lemon model and to convert other lexical categories, namely verbs, which present a more complex structure.

References

[1] T. Berners-Lee. Linked data. *W3C Design Issues*, 2006.

[2] N. Calzolari. Approaches towards a 'Lexical Web': the Role of Interoperability. In J. Webster, N. Ide, and A. C. Fang, editors, *Proceedings of The First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pages 18–25, 2008.

[3] P. Cimiano, J. McCrae, P. Buitelaar, and E. Montiel-Ponsoda. *On the Role of Senses in the Ontology-Lexicon*. 2012.

[4] Y. Hayashi. Direct and indirect linking of lexical objects for evolving lexical linked data. In *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011)*, 10 2011.

[5] A. Lenci, N. Bel, F. Busa, N. Calzolari, E. Gola, M. Monachini, A. Ogonowski, I. Peters, W. Peters, N. Ruimy, M. Villegas, and A. Zampolli. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263, 2000.

[6] J. McCrae, D. Spohr, and P. Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I*, ESWC'11, pages 245–259, Berlin, Heidelberg, 2011. Springer-Verlag.

[7] J. Pustejovsky. The generative lexicon. *Comput. Linguist.*, 17(4):409–441, dec 1991.

[8] N. Ruimy, O. Corazzari, E. Gola, A. Spanu, N. Calzolari, and A. Zampolli. The european le-parole project: The italian syntactic lexicon. In *Proceedings of the First International Conference on Language resources and Evaluation*, pages 241–248, 1998.

[9] A. Toral and M. Monachini. Simple-owl: a generative lexicon ontology for nlp and the semantic web. In *Workshop of Cooperative Construction of Linguistic Knowledge Bases, 10th Congress of Italian Association for Artificial Intelligence.*, 2007.