Undefined 0 (2013) 1–0
IOS Press

Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with *lemon*

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 describes the publication and linking of (parts of) PAROLE SIMPLE CLIPS (PSC), a large scale Italian lexicon, to the Semantic Web and the Linked Data cloud using the *lemon* model. The main challenge of the conversion is discussed, namely the reconciliation between the PSC semantic structure which contains richly encoded semantic information, following the qualia structure of the generative lexicon theory and the *lemon* view of lexical sense as a reified pairing of a lexical item and a concept in an ontology. The result is two datasets: one consists of a list of *lemon* lexical entries with their lexical properties, relations and senses; the other consists of a list of OWL individuals representing the referents for the lexical senses. These OWL individuals are linked to each other by a set of semantic relations and mapped onto the SIMPLE OWL ontology of higher level semantic types.

Keywords: *lemon*, 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 [3]. In the linked data model, data is formatted according to the Resource Description Framework (RDF)[1] model and is therefore structured in the form of *subject-predicate-object* triples. These triples are used to link together data items using their Uniform Resource Identifiers (URIs) making it far simpler to use description standards and formats such as Web Ontology 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 LRT infrastructure variously described in the LRT literature as the "Lexical Web" [4] and as a "Lexical Linked Space" [9]. LRT research has traditionally put great emphasis on the standardisation, linking, and reusability of lexical resources (LRs) 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 lexicons and the development of lexicon creation tools that use the RDF format. It then becomes much more straightforward to connect lexicons to other relevant data resources such as web-based ontologies.

By now there has been extensive work carried out on the ontologisation and publication of lexical resources as linked open data. Among the most important studies in this area are [8,1] for the Princeton WordNet, and [11] for the multilingual resource EuroWordNet.

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

The work described in this paper aims at the conversion of a subset of the lexical items, namely the nouns, in the large-scale, multi-layered Italian lexicon PAROLE SIMPLE CLIPS (PSC) into the RDF format, using the *lemon* model. This process entailed the full conversion of the semantic layer of the lexicon into OWL, as well as the creation of a new version of the lexicon modeled according to *lemon* and containing all the nouns of PSC lexicon; these two resources were then linked by using the `LexicalSense` object of *lemon* to map between them.

2. Lexical Ontologies with *lemon*

lemon (LExicon Model for ONtologies)[3] [12] 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 help to describe the basic aspects of the entries in most lexicons such as those relating to morphology, the phrase structure of complex expressions, and the syntactic frames associated with a verb. It also allows the addition of semantic information to a given lexical entry by mapping it to a concept in an ontology via an intermediate lexical sense object. This entails a clear separation between the linguistic and ontological levels of a lexical resource and facilitates the ”plugging-in" of different ontologies into the same lexicon. This is particularly useful when it comes to modelling the meaning of terms in different domains.

As mentioned above this paper describes the (partial) conversion of a lexical resource into the RDF format. The *lemon* framework was adopted for this purpose for a number of reasons. The core reason being of course that *lemon* was explicitly developed for just this kind of work, its creators striving both to meet the requirements of the language resources community and to adhere to the principles of the emerging linked open data paradigm[12]. It was also designed to be efficient and easy to use and capable of being adapted to a wide range of differently conceived lexicons. Indeed these efforts have met with a good deal of success, and the use of *lemon* is fast becoming widespread[4] among the LR community, which is in itself a compelling enough reason to work with it. Most notably it has been taken up by the Ubiquitous Knowledge Processing (UKP) Lab at the Technische Universität Darmstadt and the Universitat Pompeu Fabra (UPF). In addition the linguistic structure of *lemon* is based on Lexical Markup Framework (LMF)[6,7], a format with which the authors of this paper have had substantial previous experience.

3. Generative Lexicon, SIMPLE and SIMPLE-OWL

PSC the lexical resource whose conversion into RDF we will be describing is a multi-layered Italian language lexicon that was built in successive stages within the framework of three major lexical resource projects. PAROLE [14] and SIMPLE[10] 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[5] 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. The semantic layer of PSC, SIMPLE, is largely based on Pustejovsky’s Generative Lexicon (GL) theory[13] [2].

GL theory posits that the meaning of each word in a lexicon is structured into components, one of which, the *qualia structure*, consists of a bundle of four orthogonal dimensions. These dimensions allow for the encoding of four separate aspects of the meaning of a word: the formal, namely that 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, that which specifies the origin of an entity. These qualia structures plays an important role within GL in explaining the phenomena of polysemy in natural languages. In fact SIMPLE is actually based on the notion of an *extended qualia structure*[15], which as the name suggests is an extension of the qualia structure notion found in GL. So that there is hierarchy of constitutive, telic, and agentive relations that can hold between semantic units.

[3] http://www.lemon-model.net/.

[4] An extensive list of *lemon* users can be found at http://www.lemon-model.net/

[5] CLIPS stands for Corpora e Lessici dell’Italiano Parlato e Scritto

SIMPLE contains a language independent ontology of 153 semantic types as well as $\sim 60k$ so called "semantic units" or *USems* representing the meanings of lexical entries in the lexicon. SIMPLE also contains 66 relations organized in a hierarchy of types and subtypes all subsumed by one of the four main qualia roles:

- FORMAL (is-a)
- CONSTITUTIVE, such as ACTIVITY → produced-by
- TELIC, such as INSTRUMENTAL → used-for
- AGENTIVE, such as ARTIFACTUAL → caused-by

Fig. 1. SIMPLE extended qualia structure

as well as a series of lexical relation relations organized into 5 main classes: and four sets of lexical relations

- SYNONYMY e.g. *car/automobile*
- POLYSEMY e.g. *chestnut* for fruit and color
- ANTONYM e.g. *fast/slow*
- DERIVATION e.g. *jewel/jewelry*
- METAPHOR e.g. *chicken* for coward

Fig. 2. SIMPLE Lexical relations

Here is a small example of the relations among USems within PSC. The lexical entry *limone* (lemon) has three USems[6] each one linked to a different semantic type.

1450limone type: Fruit
76884limone type: Color
D2244limone type: Plant

Among these three USems, the PSC semantic framework implements different types of relations; qualia relations such as:

1450limone is-a (FORMAL) D2369frutto
1450limone produced-by (CONSTITUTIVE) D2244limone
1450limone object-of-the-activity (TELIC) D598mangiare

and lexical relations such as:

1450limone polysemy-plant-fruit D2244limone
1450limone polysemy-vegetal-entity-color 76884limone

The construction of the SIMPLE-OWL ontology [16] began with the extraction of the SIMPLE semantic types (e.g.,"Plant", "Flower", "Color" ...). Relations were then induced between these semantic types by generalising relations between USems (e.g., "is-a" and "contains") and the features associated with them (e.g., "plus_edible"), and adding a number of well-formedness constraints. SIMPLE-OWL was induced from the SIMPLE lexicon using a "bottom-up" strategy and can therefore be regarded as a linguistically-motivated ontology. So that as well as formalizing the typical ontological relations derived from the qualia structure (see Figure 1, SIMPLE-OWL also contains the lexical relations (see Figure 2).

4. Converting the PAROLE SIMPLE CLIPS Lexicon into *lemon* and linking to SIMPLE-OWL

In this section we explain how the (partial) conversion of PSC into *lemon* was carried out, paying particular attention to the distinction between the meaning of `LexicalSense` in *lemon* and the concept of USem in PSC.

As described above the *lemon* model requires a lexical sense object, seen a reified lexical-semantic pairing, to mediate between a lexical entry and the concept or meaning of that entry as provided in an ontology.

The main problem faced in this conversion related to the fact that it was not always possible to identify PSC USems with lexical sense objects in *lemon*. This becomes evident when one comes to consider the fact that only certain limited kinds of relations can hold between lexical sense objects in *lemon*, e.g., that two senses are incompatible or that two senses are equivalent; whereas conceptual relations relating to the meanings or referents of words belong in the ontology, since according to the *lemon* philosophy these distinctions should not affect the conceptualisation of the words in a domain [5]. In SIMPLE on the other hand no such division is made and USems can be linked by both lexical relations (polysemy, derivation, ...) and conceptual relations relating to the meaning of lexical entries (such as producedby, hasparts, ...).

For this reason the decision was taken to duplicate each USem from SIMPLE both as a *lemon* lexical sense and as an individual in an ontology; the former was then linked to the latter using the *lemon* reference relation. The aforementioned ontological individuals were then mapped onto their types in the SIMPLE-OWL ontology.

This allows one to properly distinguish between the SIMPLE relations: so that SIMPLE lexical relations are now encoded between *lemon* lexical senses in a

[6] Here and afterwards, we simplify things by representing each USem in the examples using only the number part of the name.

lexicon, whereas SIMPLE qualia relations now pertain to individuals in an ontology. In addition to this it was decided to use the "is-a" relation among USems also to induce the narrower/broader relations among lexical items as defined by the *lemon* model.

The final output of this conversion has been partitioned into the following datasets:

- **SIMPLE-OWL** types, which contains the definitions of both semantic types and relations.
- **SIMPLE** Entries which contains the list of all USems in SIMPLE converted into OWL named individuals. These concepts are connected to their semantic type in SIMPLE-OWL through `rdf:type`.
- **pscLemon** which contains the lexical items of SIMPLE converted into *lemon* lexical entries, with part of speech information and list of senses.

The following sets of relations holds between items in these datasets:

- Extended qualia relations as defined in SIMPLE-OWL, holding between individuals;
- Lexical relations, as defined in SIMPLE-OWL, holding between lexical senses;
- Induced narrower/broader relations, as defined by the *lemon* model, holding between lexical senses.

Here a set of examples shown are given to clarify the procedure. Please note that turtle[7] notation is used. First of all the lexical entries and their senses need to be instantiated:

```
limone a lemon:LexicalEntry.
limone_1 a lemon:LexicalSense.
limone_2 a lemon:LexicalSense.
limone_3 a lemon:LexicalSense.
```

Each lexical sense $\sigma^{l,c}$ connects the lexical entry *limone* (l) to the corresponding USem (c) in SIMPLE Entry through a `lemon:reference`:[8]

```
limone_1 a lemon:LexicalSense;
     lemon:reference inds:USem1450limone
```

Then lexical relations are instantiated among lexical senses in the pscLemon resource. In *lemon* we have:

```
limone_1 a lemon:LexicalSense;
   lemon:reference inds:USem1450limone;
   simple:PolysemyPlant-Fruit  limone_2.
```

[7] www.w3.org/TR/turtle/

[8] The namespaces *inds* and *simple* in the following examples are defined in section 5.1.

The last information to be added to the pscLemon resource concerns the narrower/broader relations. Using the the "is-a" *qualia* relation it is inferred that the sense limone_1 is narrower than the sense frutto_1 which gives:

```
limone_1 a lemon:LexicalSense;
   lemon:reference  inds:USem1450limone;
   lemon:narrower  frutto_1;
   simple:PolysemyPlant-Fruit  limone_2.
```

The SIMPLE Entry resource contains the relations among concepts (USems) and the link between each concepts and the general ontological types defined in SIMPLE-OWL ontology. As stated above, this resource contains only the set qualia relations.

```
1450limone
   a simple:Fruit, owl:NamedIndividual;
   simple:hasProducedby  D2244limone;
   simple:hasIsa  D2369frutto.
```

Figure 3 represents the interrelations among the three resources described above.

5. Structure of the data and distribution

The whole dataset produced for this paper is available under the Data Hub catalogue.[9,10] All the resources which belong to the dataset are licensed with a "Open Data Commons Attribution License".[11]
The dataset consists of the three resources described in section 4, which were made available in two formats: a big file containing all of the entries of each resource, and a set of individual files, one for each entry. In the Data Hub catalogue separate lists of entries are also provided, for each resource in the different modalities:

- **lemon entries pointing to big file** A list of *lemon* lexical entries, each list item pointing to one file;
- **lemon entries pointing to single file** A list of *lemon* lexical entries each list item pointing to a single file containing the corresponging entry;
- **individuals pointing to big file** A list of SIMPLE entries, each pointing to one file;
- **individuals pointing to single file** A list of SIMPLE entries, each pointing to a single file containing the corresponging entry;

[9] http://www.datahub.io/dataset/simple

[10] See http://www.datahub.io/about for more information on the Data Hub project.

[11] http://www.opendefinition.org/licenses/odc-by

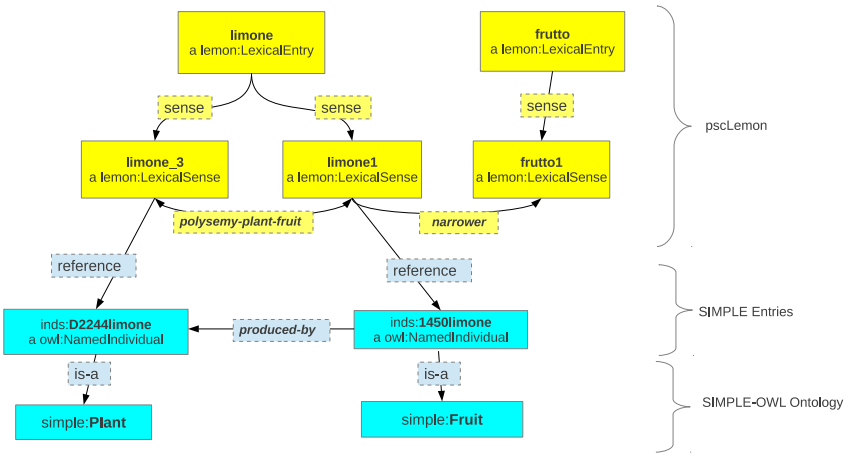

Fig. 3. Schema of the example.

5.1. Namespaces

The resources listed in sections 5 are hosted under the common namespace: `http://www.languagelibrary.eu/owl/simple` hereafter `base`.

The resources have been stored under sub-folders of `base`, according to their specific content. Table 1 shows the resources and their namespaces. The corresponding URIs of the resources are the concatenation of `base` with `name`

Table 1

Resources and namespaces

Resource	Namespace
SIMPLE-OWL Ontology	base
SIMPLE Entries	base/inds
pscLemon	base/psc
individuals pointing to single file	base/inds
individuals pointing to big file	base/inds
lemon entries pointing to big file	base/psc
lemon entries pointing to single file	base/psc

5.2. Data figures and Obtained Triples

As explained in Section 2, so far only the nouns have been extracted from the PAROLE SIMPLE CLIPS lexicon; this was purely due to time constraints. The number of entries that have been processed is 31232 USems (corresponding to all of the nouns in the lexicon), 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:

Table 2

Files, units and triples

File	Original Units	Triples
SIMPLE-OWL Ontology	153	6,332
SIMPLE Entries	50,502	247,264
pscLemon	31,444	372,296

5.3. Data Organization

The *SIMPLE Entries* and *pscLemon* resources are not very usable, since they collect all entries in the same file. In fact, in these resources, all references are prefixed by #, meaning that each entry is referenced with an inner anchor. Users are then 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 also 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 *limone* which contains the *lemon* lexical entry for “limone”; and there are three distinct files (one file for each lexical sense of “limone”): “USem1450limone”, “USemD2244limone” and “USem76884limone” to describe the information extracted from PSC.

A file system structure was created based on the first characters of the hash coding of the lexical entry, for example $limone \rightarrow 2/299$. Under the namespaces `base/psc` and `base/inds` were added structures similar to $2/299$, single entry files were also inserted, as shown in Figure 4: As a consequence, all entries

```
base/psc/
  2/299/
    limone
base/inds/
   2/299/
    USem1450limone
    USemD2244limone
    USem76884limone
```

Fig. 4. Example of folder structures

from *SIMPLE Entries* and *pscLemon* were extracted and four resources containing the lists of entries were added to the provided dataset, as detailed above.

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 lexical resources with the representation of the multiple dimensions of meaning instantiated by SIMPLE. This differentiates the present solution from possible others, in which the all SIMPLE semantic relations are encoded directly among lexical sense objects without reference to an external ontology.

The ontological individuals that were added during the conversion of the SIMPLE USems are labeled in Italian; the top ontology instead has English labels and can be mapped to by all of the European language projects falling under the SIMPLE banner. Thus some additional mapping is required to merge "USemCane1" and "USemperro1" where they represent the same concept. Other proposed future work includes importing further elements of the morpho-syntactic information in PSC into the *lemon* model and of course converting the other lexical categories in PSC, in particular the verbs, which present a more complex structure into *lemon*.

References

[1] M. V. Assem, A. Gangemi, and G. Schreiber. Conversion of wordnet to a standard rdf/owl representation. In *Proceedings of LREC 2006*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[2] N. Bel, F. Busa, N. Calzolari, E. Gola, A. Lenci, 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. In *Proceedings of LREC 2000*, Athens, Greece, 2000.

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

[4] 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, Hong Kong, 2008.

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

[6] G. Francopoulo. *LMF - Lexical Markup Framework*. ISTE Ltd + John Wiley & sons, Inc, 1 edition, 2013.

[7] G. Francopulo, R. Laurent, M. Monachini, and N. Calzolari. Lexical markup framework (lmf iso-24613). In *Proceedings of LREC 2006*, Genova, Italy, 2006.

[8] A. Gangemi, R. Navigli, and P. Velardi. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer, 2003.

[9] Y. Hayashi. Direct and indirect linking of lexical objects for evolving lexical linked data. In E. Montiel-Ponsoda, J. McCrae, P. Buitelaar, and P. Cimiano, editors, *MSW*, volume 775 of *CEUR Workshop Proceedings*, pages 62–67. CEUR-WS.org, 2011.

[10] 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.

[11] E. W. D. Luca, M. Eul, and A. Nürnberger. Converting eurowordnet in owl and extending it with domain ontologies. In C. Kunze, L. Lemnitzer, and R. Osswald, editors, *Proceedings of the GLDV-2007 Workshop on Lexical-Semantic and Ontological Resources*, pages 39–48, 2007.

[12] 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.

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

[14] 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.

[15] N. Ruimy, M. Monachini, R. Distante, E. Guazzini, S. Molino, M. Ulivieri, N. Calzolari, and A. Zampolli. Clips, a multi-level italian computational lexicon: a glimpse to data. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain, 2002.

[16] 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.