Semantic Web 1 (2009) 1–5
IOS Press

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

Editor(s): Sebastian Hellmann, AKSW, University of Leipzig, Germany; Steven Moran, LMU Munich, Germany; Martin Brümmer, AKSW, University of Leipzig, Germany; John P. McCrae, CITEC, Bielefeld University, Germany
Solicited review(s): Sebastian Hellmann, AKSW, University of Leipzig, Germany; John P. McCrae, CITEC, Bielefeld University, Germany; three anonymous reviewers

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 data distributed at various locations across the web by setting up a standardized way of structuring, describing, and interlinking datasets [3]. In the linked data model, data is formatted according to the Resource Description Framework (RDF)[1] model. The RDF triples are used to link together data resources which are referred to by their Uniform Resource Identifiers (URIs). Linked open data refers to linked data freely available for download and use.

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" [10]. 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.

The increased awareness of the importance of linked data within the LRT community has resulted in a trend towards the conversion of language resources, in particular lexicons, using the RDF format. This has the added benefit that these language resources can also be linked to other kinds of resources on the linked data cloud, such as for example DBpedia.

By now there has been extensive work carried out on the publication of lexical resources as linked data.

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

Among the most important studies in this area are [9,1] describing the conversion of the Princeton WordNet, and [13] for the multilingual resource EuroWordNet.

This paper describes the conversion of a subset of the lexical entries, namely all of the nouns, from a large-scale, multi-layered Italian lexicon PAROLE SIMPLE CLIPS (PSC) as linked open data using the *lemon* model. This process included the full conversion of the semantic layer of the lexicon into Web Ontology Language (OWL), as well as the creation of a resource containing the lexical entries of PSC and containing all the nouns of PSC lexicon; these two resources were then linked using the `LexicalSense` object of *lemon* to map between them.

2. Lexical Ontologies with *lemon*

lemon (LExicon Model for ONtologies)[2] [14] is a descriptive model that supports the linking together of a computational lexicon with with an ontology, where the ontology provides the semantics for the lexical entries in the lexicon.

lemon defines a core and a set of additional modules that together serve to describe the basic morphological and syntactic data typically associated with the lexical entries in a lexicon. It also allows the addition of semantic data to a given lexical entry though mapping the lexical entry 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 which in turn enables the reuse and plugging in of different ontologies to the same lexical resource.

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. It was, the authors felt, an extremely efficient and easy to use model. In addition it has become one of the popular models available for publishing computational lexicon as linked data, indeed, one might argue it is almost a de facto standard[3] among the LR community. 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) for the conversion of several resources [19,6]. Finally the definition *lemon* is heavily based on Lexical Markup Framework (LMF)[7,8], a format with which the authors of this paper have had substantial previous experience.

3. SIMPLE and SIMPLE-OWL

PAROLE SIMPLE CLIPS (PSC) is a multi-layered Italian language lexicon that was developed in successive stages within the framework of three major lexical resource projects, PAROLE, SIMPLE, CLIPS. PAROLE [16] and SIMPLE [12] were consecutive European projects which resulted in the creation of a wide ranging Italian language lexicon PAROLE-SIMPLE (in addition to similar lexicons in 11 other European languages) that was structured into different, interconnected layers; CLIPS[4] 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 [15,2].

GL theory is based on the idea 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 within a hierarchy; 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 are used within GL theory in order to explain polysemy in natural languages.

The semantic layer of PSC, which we will refer to as SIMPLE, is actually based on the notion of an *extended qualia structure* [17], which is, as the name suggests, an extension of the notion of qualia structure found in GL in which a hierarchy of different constitutive, telic, and agentive relations has been defined, as we will see below.

SIMPLE contains a language independent ontology of 153 semantic types common to all of the different language lexicons that were developed as part of the PAROLE SIMPLE projects, as well as $\sim 60k$ so called

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

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

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

"semantic units" or *USems* representing the meanings of lexical entries in the lexicon. SIMPLE also contains 66 relations organised in a hierarchy which is structured around the four main qualia roles:

- FORMAL (is-a);
- CONSTITUTIVE, such as produced-by;
- TELIC, such as used-for object-of-the-activity;
- AGENTIVE, such as caused-by.

as well as a series of lexical relation relations organized into 5 main classes: and 4 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

For example, the lexical entry *limone* (which means *lemon* in English) has three USems each one subsumed by a different semantic type.

USem1450limone type: Fruit
USem76884limone type: Color
USemD2244limone type: Plant

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

USem1450limone is-a USemD2369frutto,
USem1450limone produced-by USemD2244limone,
USem1450limone object-of-the-activity
USemD598mangiare,

and lexical relations such as:

USem1450limone polysemy-plant-fruit
USemD2244limone,
USem1450limone polysemy-vegetal-entity-color
USem76884limone.

Previous work [18] on the construction of an OWL ontology, SIMPLE-OWL, based on the semantic type ontology that was informally presented in the SIMPLE specifications, began with the extraction of the semantic types (e.g.,"Plant", "Flower", "Color" etc.). 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. As well as formalizing the typical ontological relations derived from the qualia structure, SIMPLE-OWL also contains the lexical relations. The SIMPLE-OWL ontology was the starting point of the work described in this paper.

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 to mediate between a lexical entry and the meaning of that entry as provided by a vocabulary item 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 how the *lemon* model is defined and its various components descibed in works such as [5] and the lemon cookbook. Given that one of the definitions of a lexical sense is as a reification of word-meaning pairings, it would seem that explicitly lexical relations such as those relating to synonymy and hyponymy, were better placed between lexical senses; whereas more conceptual relations, namely those explicitly pertaining, or that seem to pertain to the extensions of words were better placed in an ontology.

Of course there are grey areas,[5] but, given the emphasis placed on the distinction between a lexicon and an ontology, the authors of this paper felt that it was better to place such as relations such as producedby or object-of-the-activity in the ontology itself.

In SIMPLE however no such division is made and USems can be linked both by relations which which are clearly lexical (polysemy, derivation, ...) and those which relate to the meaning of lexical entries (such as producedby, hasparts, ...) rather than as senses qua reified word-meaning pairings.

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

[5] Most works on lexical semantics for example consider that meronymy relations to be lexical relations.

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 lexicon, whereas SIMPLE qualia relations now relate items 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. We partizioned the final conversion 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 are then 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.

With the following sets of relations among items:

- 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 are given to clarify the procedure.[6] 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 connects a lexical entry to a corresponding USem in SIMPLE Entry through a `lemon:reference`:[7]

```
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.
```

[6] Here and in other examples, we have used the Turtle notation. See http://www.w3.org/TR/turtle/

[7] The namespaces *inds* and *simple* in the following examples are defined in section 5.

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 1 represents the interrelations among the three resources described above.

5. Structure of the data and distribution

The whole dataset produced for this paper is hosted at http://www.languagelibrary.eu/owl/simple/ (hereafter `base`) and licensed with an "Open Data Commons Attribution License". [8] The access page contains links to:

- the indexes of PSC lexical entries and of PSC ontological referents, containing the lists of individual URIs for online access;
- a compressed version of lexicon and ontology for download.

The resource files are stored under sub-folders of `base`, according to their specific content. Table 1 shows the resources and their namespaces. The corresponding Uniform Resource Identifiers (URIs) of the resources are the concatenation of `base` with `name`

As explained in Section 2, so far only the nouns have been extracted from the PAROLE SIMPLE CLIPS lexicon. 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 Ta-

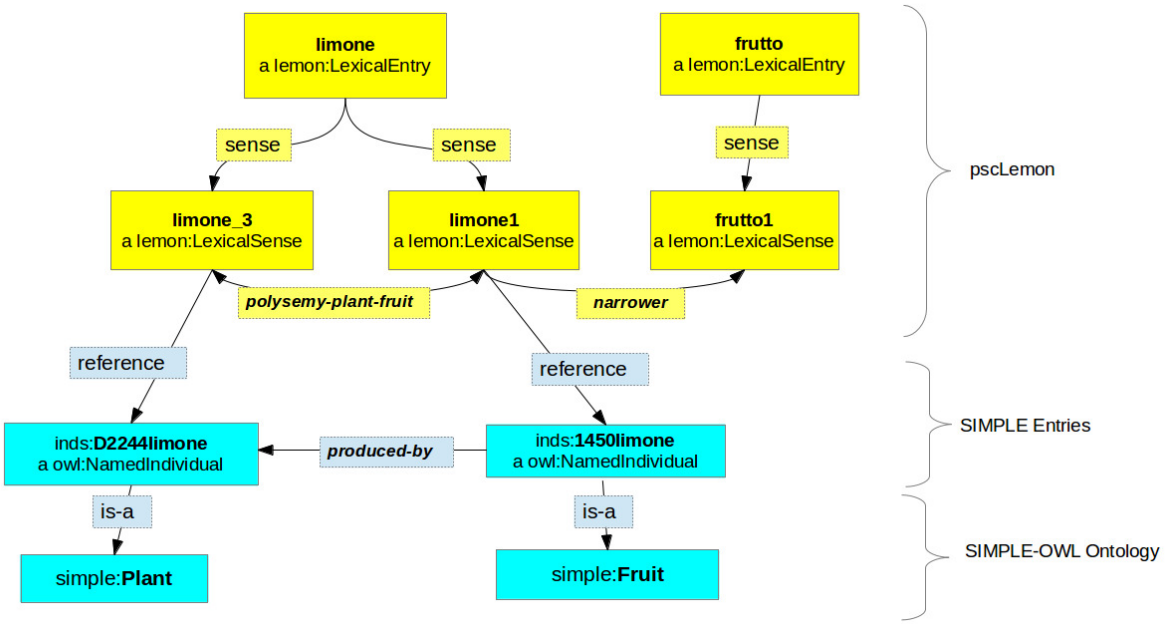

Fig. 1. Schema of the example.

Table 1
Resources and namespaces

Resource	Namespace
SIMPLE-OWL Ontology	base
SIMPLE Entries	base/inds
pscLemon	base/psc

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

ble 2: Notice that, while the downloadable versions of lexicon and ontology are collected in two big files of entries, the online version is organized according to the linked data paradigm in such a way that each entry constitutes a single 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 sense of "limone"): "USem1450limone", "USemD2244limone" and "USem76884limone" to describe the information extracted from PSC.

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

To limit the number of files in a each folder, a file system structure was created under the namespaces `base/psc` and `base/inds`, based on the first characters of the hash coding of the lexical entry, as shown in Figure 2. The correct URIs for each entry can be found in the two aforementioned index files.

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

Fig. 2. Example of folder structures

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 solutions in which all SIMPLE semantic relations are encoded directly among lexical sense objects without reference to an external ontology.

In a recently submitted paper[11], a proposal was presented for translating the PSC verbs into RDF. In

this proposed model, a verb sense can have an associated syntactic frame as well as a predicative semantic frame. These syntactic and semantic frames are then further specified as regards their arguments using for example the LMF property `'has Argument'`. Syntactic frames are described independently of specific verbs in PSC (for example the class of transive frames that take *avere* as an auxiliary). In addition different mappings between syntactic and semantic frames have the status of independent objects in PSC. Both of these design decisions have been closely adhered to in the proposed model.

References

[1] Mark Van Assem, Aldo Gangemi, and Guus Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the 5th International Conference on Language resources and Evaluation (LREC 2006)*, pages 237–242, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[2] Núria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta, and Antonio Zampolli. SIMPLE: A General Framework for the Development of Multilingual Lexicons. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000. European Language Resources Association (ELRA).

[3] Tim Berners-Lee. Linked Data. *W3C Design Issues*, 2006.

[4] N. Calzolari. Approaches towards a 'Lexical Web': the Role of Interoperability. In Jonathan Webster, Nancy Ide, and Alex Chengyu Fang, editors, *Proceedings of the First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, pages 18–25, Hong Kong, 2008. City University, City University.

[5] Philipp Cimiano, John McCrae, Paul Buitelaar, and Elena Montiel-Ponsoda. *On the Role of Senses in the Ontology-Lexicon*. Springer-Verlag, Berlin - Heidelberg, 2013.

[6] Judith Eckle-Kohler, John McCrae, and Christian Chiarcos. lemonUby - a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal. Multilingual LOD 2012 MSS*, 2014.

[7] Gil Francopoulo. Computer Engineering and IT series. ISTE Ltd + John Wiley & sons, Inc, 1 edition, 2013.

[8] Gil Francopulo, Romary Laurent, Monica Monachini, and Nicoletta Calzolari. Lexical Markup Framework (LMF ISO-24613). In *Proceedings of the 5th International Conference on Language resources and Evaluation (LREC 2006)*, pages 233–236, Genova, Italy, 2006. European Language Resources Association (ELRA).

[9] Aldo Gangemi, Roberto Navigli, and Paola Velardi. The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet. In Robert Meersman, Zahir Tari, and DouglasC. Schmidt, editors, *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, pages 820–838. Springer Berlin Heidelberg, 2003.

[10] Yoshihiko Hayashi. Direct and Indirect Linking of Lexical Objects for Evolving Lexical Linked Data. In Elena Montiel-Ponsoda, John McCrae, Paul Buitelaar, and Philipp Cimiano, editors, *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web (MSW 2011), Bonn, Germany, October 23, 2011*, volume 775 of *CEUR Workshop Proceedings*, pages 62–67. CEUR-WS.org, 2011.

[11] Fahad Khan and Francesca Frontini. Publishing PAROLE SIMPLE CLIPS as Linguistic Linked Open Data. In *CLIC-IT, la Prima Conferenza di Linguistica Computazionale Italiana*, Pisa (Italy), December 2014.

[12] Alessandro Lenci, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. SIMPLE: A General Frameword for the Development of Multilingual Lexicons. *International Journal of Lexicography*, 13(4):249–263, 2000.

[13] Ernesto William De Luca, Martin Eul, and Andreas Nürnberger. Converting EuroWordNet in OWL and Extending it with Domain Ontologies. In C. Kunze, L. Lemnitzer, and R. Osswald, editors, *Proceedings of the Gesellschaft fũr linguistische Datenverarbeitung (GLDV 2007) Workshop on Lexical-Semantic and Ontological Resources*, pages 39–48, Tübingen, 2007. Gunter Narr Verlag.

[14] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with *lemon*. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011) on The Semantic Web: research and applications - Volume Part I*, pages 245–259, Berlin, Heidelberg, 2013. Springer-Verlag.

[15] James Pustejovsky. The generative lexicon. *Comput. Linguist.*, 17(4):409–441, December 1991.

[16] 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 (LREC 1998)*, pages 241–248. European Language Resources Association (ELRA), 1998.

[17] Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, and Antonio Zampolli. CLIPS, a Multi-level Italian Computational Lexicon: a Glimpse to Data. In *Proceedings of the 3rd International Conference on Language resources and Evaluation (LREC 2002)*, pages 792–799, Las Palmas, Canary Islands, Spain, 2002. European Language Resources Association (ELRA).

[18] Antonio Toral and Monica 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 (IA*AI)*, Rome (Italy), 2007.

[19] Marta Villegas and Núria Bel. PAROLE/SIMPLE Lemon ontology and lexicons. *Semantic Web Journal. Special issue on Multilingual Linked Open Data (MLOD) 2012 Data Post Proceedings*, 2013.