10.12871/CLICIT2014143

Publishing PAROLE SIMPLE CLIPS as Linguistic Linked Open Data

Fahad Khan
ILC CNR Pisa
fahad.khan@ilc.cnr.it

Francesca Frontini
ILC CNR Pisa
francesca.frontini@ilc.cnr.it

Abstract

English. This paper presents the ongoing project for the conversion and publication of the Italian lexicon Parole Simple Clips in linked open data, illustrating the chosen model, with a particular focus on the translation of the syntactic and semantic information pertaining verbs and their predicates.

Italiano. *Questo paper presenta il progetto in corso per la conversione e pubblicazione del lessico italiano Parole Simple Clips nel formato linked open data, descrivendo il modello adottato con particolare riferimento alla traduzione delle informazioni sintattico semantiche dei verbi e dei loro predicati.*

1 Introduction

The aim of the present paper is to describe the ongoing conversion of the semantic layer of the Parole Simple Clips (PSC) lexical resource into linked open data. We have previously presented the conversion of the nouns in PSC in (Del Gratta et al., 2013). In this paper we will continue this work by presenting the model we intend to use for converting the verbs.

In the next section we shall give a general background on the linguistic linked open data (LLOD) cloud and discuss the importance of putting lexical resources on the cloud. We also discuss the *lemon* model which we have chosen as the basis of the conversion of the PSC resource. In the following section we discuss PSC itself and give a brief overview of its structure.

Finally in the last section we will outline how we intend to proceed with the conversion of the PSC verbs, illustrating the proposed schema with an example.

2 Linguistic Linked Open Data

The term linked open data refers to the practice of publishing data online in a standardised format that makes the interlinking of distributed datasets more straightforward and so much more commonplace. Furthermore the modifier "open" in this context refers to the idea that the datasets in question should be free to be downloaded and used by the public.

Over the last few years data about the growing number of datasets published as linked open data, the so called linked open data cloud, has been presented in the form of a diagram in which each dataset is represented by a node and the links between each dataset by edges between the corresponding nodes.

The publishing of data as linked open data is based on principles first elucidated by Tim Berners Lee (Berners-Lee, 2006). These principles recommend the use of the resource description framework (RDF), a language that models data in terms of triples of resources. Each of the resources in a triple is named using a unique resource identifier (URI). An RDF triple can be regarded as representing data in the form of a subject-predicate-object statement.

The many advantages and benefits of the emerging linked open data paradigm are obvious from a scientific standpoint. By putting different resources on the linked open data cloud it becomes far easier to link them together with each other in ways which render single resources much more useful than before, it also makes them more accessible and usable, facilitaing their reuse in an open ended variety of contexts (as is the case with the linked data version of Wikipedia). Indeed, this fact has not been lost on the language resources community, and the specific part of the linked open data cloud diagram dealing with language resources and datasets now includes a wide array

of linguistic resources including translations of the current version of the Princeton wordnet and ItalWordnet into RDF (Assem et al., 2006; Gangemi et al., 2003), (Bartolini et al., 2013), as well as a number of important vocabularies for language resources.

The *lemon* model (McCrae et al., 2011) is currently one of the most popular rdf based models for enabling the publishing of lexical resources as linked open data on the web. Its original focus was on the addition of linguistic information to ontologies, but it has by now been used to translate numerous different kinds of lexical resources into RDF, including many different wordnets.

Because the intial focus was on enriching already existing ontologies with lingustic information, the lemon model makes a clear distinction between a lexicon and an ontology. The pairing of a lexicon and an ontology as a combined lexico-semantic resouce takes place via the interlinking of a RDF based lexicon with an RDF based ontology. This is done using so called sense objects which are pointed to by lexical entries and which then in turn point to the vocabulary items in an ontology.

3 Parole Simple Clips and the Generative Lexicon

Parole Simple Clips (PSC) is a large, multilayered Italian language lexicon, the result of work carried out within the framework of three successive European/national projects. The first two of these were the European projects PAROLE (Ruimy et al., 1998) and SIMPLE (Lenci et al., 2000a) which produced wide coverage lexicons for a number of different European languages, including Italian, and all of which were designed to a common set of guidelines. These lexicons are arranged into phonetic, morphological, syntactic and semantic layers (the semantic layers were actually added during the SIMPLE project, the other layers during the earlier PAROLE project). The last of the projects instrumental in the creation of PSC was CLIPS, an Italian national project, which had the aim of expanding upon the Italian Parole-Simple lexicon. In this paper we focus on the translation of the syntactic and semantic layers of PSC into RDF using the *lemon* model.

The construction of the semantic layer of PSC was heavily influenced by Generative Lexicon (GL) theory (Pustejovsky, 1991; Bel et al., 2000). GL theory posits a complex multi part structure for individual word senses, making provision for the encoding of information related to different, salient, dimensions of a lexical entry's meaning[1].

In GL a lexical entry contains information on the position of the lexical entry in a language wide type system its so called lexical type structure; a predicative argument structure; information on the event type of the entry, the event structure; as well as a data structure known as a qualia structure. This qualia structure presents four distinct, orthogonal aspects of a word's meaning in terms of which polysemy as well as the creative and the novel uses of words based on established meanings can be straightforwardly explained.

These four aspects or qualia roles contained in each lexical entry's qualia structure, can be defined as follows. **The formal quale**: this corresponds to the ontological isA relation; **the constitutive quale**: this encodes meronymic or partOf relationships between an entity and the entities of which it is composed; **the telic quale**: this encodes the purpose for which an entity is used; **the agentive quale**: this encodes the factors that were involved in an entity's coming into being.

3.1 The Structure of the Semantic Layer of PSC

The semantic layer of PSC builds upon this theoretical foundation by introducing the notion of an Extended Qualia Structure (Lenci et al., 2000a) according to which each of the four qualia roles are further elaborated by being broken down into more specific relations. This means that for example the constitutive relation is further elaborated by relations specifying whether a constitutive relation holds between two elements on the basis of location, group membership, etc; telic relations are specified in terms of purpose, classified with respect to direct and indirect telicity, etc.

In PSC this Extended Qualia Structure is represented as a relation that holds between semantic units or USems in the terminology of PSC. In addition to the extended qualia relations there are also a number of so called lexical relations organised into the following five clases SYNONYMY, POLYSEMY, ANTONYMY,

[1]This information is used to construct larger units of meaning through a compositional process in which, to use the slogan common in GL theory literature, the semantic load is more equally spread over all of the consituents of an utterance, rather than being largely focused on the verbs.

DERIVATION, METAPHOR.

PSC makes use of a language independent, 'upper' ontology that is also common to the PAROLE-SIMPLE lexicons for other European languages; this has been converted into an OWL ontology (Toral and Monachini, 2007) which we make use of in our translation. This ontology contains 153 so called semantic types which provide a higher level structuring of the circa 60k Italian language specific USems contained in PSC. In order to illustrate the levels of information available in PSC, we use the example of the verb "dare", to give.

The verbal lexical entry *dare* maps onto 3 different semantic units (USem): (i) USem7149dare as in "to give something to someone"; (ii) USem79492dare as in "to give the medicine to the patient" (make ingest); (iii) USem79493dare as in "the window faces the square".

In PSC, the first two USems map onto the same syntactic frame with 3 positions, representing subject, object and indirect object, all of which are noun phrases and the latter of which is introduced by the preposition "a". We also know that this frame selects the auxiliary "avere". The other USem uses a bivalent frame instead.

In the semantic layer, a mapping is defined between each USem and a predicate. This mapping is not one-to-one, as in some cases two senses may map onto one predicate[2].

The predicates are then linked to their argument structures, so for instance the predicate structure of USem7149dare has three arguments, the first has the role of **Agent** and selects ontological type *Human*, the second has the role of **Patient** and selects the ontological type *Concrete_entity*, the third has role **Beneficiary** and selects the ontological type *Human*. A linking is also available between the semantic and the syntactic structure; in this case the predicative structure and the syntactic frame of USem7149dare are linked by an isomorphic trivalent relation, which means that Position1 maps onto Argument1, Position2 maps onto Argument2, and Position3 maps onto Argument3.

Finally, each of the USems of *dare* linked to other USems in the lexicon by means of the complex network of relations of the Extended Qualia Structure, and is also linked to the Interlingual upper level SIMPLE ontology. So for instance USem7149dare has ontological type *Change_of_Possession* and is linked to USem3939cambiare ("to change") on the formal axis and to USemD6219privo ("deprived of") on the constitutive axis, the USem USemD6219privo being the resulting state of the USem7149dare. Lexical relations such as polysemy or derivation are also possible for verbs.

4 Converting PSC into linked Data with *lemon*

A detailed account of the challenges brought about by the translation of the PSC resource into RDF is presented in (Del Gratta et al., 2013). Here we will summarize that work and thus lay the ground for further discussion on the translation of the PSC verbs in the next section.

The main challenge that arose during the conversion of the PSC nouns related to how best to understand the status of the Usems, namely whether these were better viewed as *lemon* senses which could then in turn be understood as reified pairings of lexical entries with ontological vocabulary items; or whether PSC USems should instead be seen as elements in an ontological layer.

As mentioned above USems take part in lexical relations such as synonymy, polysemy and antonymy which in standard works are treated as relations between lexical senses[3]. On the other hand PSC USems also take part in (Extended Qualia Structure) relations that are arguably better classed as ontological relations holding between the referents of words rather than between their senses, e.g., produces, produced-by, used-for, is_a_follower_of, is_the_habit_of: at the very least it seems odd to say that the relation of synonymy and a relation specifying whether relations of one class "produce" members of another hold between the same kind of element.

In the end the considerations given above along with the fact that the lemon model makes such a clear distinction between lexicon and ontology led to the decision to duplicate the USems: once as *lemon* lexical senses, with lexical relations like synonymy holding between them, and in the second instance as ontological entities. These are then to be seen as an lower level of the already

[2]This is especially the case for reflexive verbs such as *incolonnarsi* ("to line up") vs their transitive counterparts ("to line something/one up"), that are represented as different senses and different syntactic frames, but have the same underlying argument structure.

[3]Although the aforementioned lexical relations can themselves bedefined differently in different sources.

existing SIMPLE OWL ontology.

4.1 The Verbs

The modelling of the PSC verbs in linked open data involves a number of challenges over and above those that arose during the modelling of the nouns. In particular it is important to represent information about both the syntactic frames and semantic predicates associated with verb senses[4]. In addition it is also desirable to have some kind of mapping between these two kinds of representation, so that the syntactic arguments of a verb frame can be mapped to the semantic arguments of the verb's semantic predicative representation.

One of the considerations that we have been most keenly aware of throughout the process of developing a model for the PSC verbs is that we are attempting to convert a legacy resource with a relatively long history and a well documented design that was developed through the collaboratio of a number of experts in the field.

We have therefore tried to remain as faithful as possible to the original intentions of the designers of PSC, while at the same time exploiting the advantages and opportunities offered up by the linked data model.

We present our proposal for verbs below. Once more we are working with the Italian verb *dare*.

```
:dare_1 a lemon:sense ;
        lemon:reference :USem7149dare ;
        psc:synBehavior frames:t-ind-xa ;

lmf:hasSemanticPredicate :PREDdare#1 ;
        psc:hasSynSemMapping
                    ssm:Isotrivalent .

:PREDdare#1 a lmf:SemanticPredicate ;
        lmf:hasArgument ARG0dare#1 ;
        lmf:hasArgument ARG1dare#1 ;
        lmf:hasArgument ARG2dare#1 .

:ARG0dare#1 a lmf:Argument ;
        a simple:ArgHuman .

:ARG1dare#1 a lmf:Argument ;
        a simple:Concrete_Entity .

:ARG2dare#1 a lmf:Argument ;
        a simple:ArgHuman .

:ARG2dare#1 lemon:marker :a .
```

The lexical entries point to their reified sense objects. In the example these are named dare_1, dare_2, whereas the USem ID is used to name the ontological counterpart of the original PSC USem, the reference object[5].

We use the psc prefix to refer to the name space main file containing the definitions of concepts and properties in the example.

Each lexical sense points to a lemon:frame by means of the psc:synBehavior property[6]. These frames are stored in a separate file, each frame in this file is an abstraction over many syntactic frames. So in the example the verb sense dare_1 is mapped to a frame t-ind-xa. This represents a transitive frame for a verb with both a direct and indirect object and which takes *avere* as an auxilliary verb.

The sense is also linked to a predicate object, which provides descriptions of the argument structure. We use the lmf property hasSemanticPredicate to link to an lmf SemanticPredicate PREDdare#1. The type selected by each argument of the predicatepoints back to the SIMPLE Ontology.

Finally the sense dare_1 is linked to ssm:Isotrivalent an object representing the mapping between the syntactic frame and the semantic predicate via the hasSynSemMapping property. We have created a file ssm that contains a number of these mappings as represented in the PSC specifications. The particular mapping object in question, Isotrivalent, represents the isomorphic trivalent relation mentioned above. Details on the best way of representing these mappings using OWL will be provided in the final paper.

5 Conclusion

In this paper we have presented our model for representing the PSC verbs using the lemon model. As we have stated above this is currently work in progress. In the final paper the link to the public dataset will be provided.

References

Mark Van Assem, Aldo Gangemi, and Guus Schreiber. 2006. Conversion of wordnet to a standard rdf/owl representation. In *Proceedings of the Fifth International Conference on Language Resources and Eval-*

[4]It is also true that PSC nouns have predicative structure but this was ignored during the intial translation of PSC into linked data.

[5]Although for space reasons we are unable to show this here, in our model lexical relations are kept between senses, while ontological relations are implemented between usems qua ontological objects, as for nouns.

[6]In our model, this is a property of Senses rather than LexicalEntries.This is in keeping with the PSC specifications.

uation (LREC-2006), Genoa, Italy, May. European Language Resources Association (ELRA).

Roberto Bartolini, Riccardo Del Gratta, and Francesca Frontini. 2013. Towards the establishment of a linguistic linked data network for italian. In *2nd Workshop on Linked Data in Linguistics*, page 76.

Núria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas, and Antonio Zampolli. 2000. Simple: A general framework for the development of multilingual lexicons. In *LREC* (DBL, 2000).

Tim Berners-Lee. 2006. Linked data. *W3C Design Issues*.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. Towards linguistically grounded ontologies. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 111–125, Berlin, Heidelberg. Springer-Verlag.

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

Philipp Cimiano, John McCrae, Paul Buitelaar, and Elena Montiel-Ponsoda, 2012. *On the Role of Senses in the Ontology-Lexicon*.

2000. *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece*. European Language Resources Association.

Riccardo Del Gratta, Francesca Frontini, Fahad Khan, and Monica Monachini. 2013. Converting the parole simple clips lexicon into rdf with lemon. *Semantic Web Journal (Under Review)*.

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

Gil Francopulo, Romary Laurent, Monica Monachini, and Nicoletta Calzolari. 2006. Lexical markup framework (lmf iso-24613). In European Language Resources Association (ELRA), editor, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Genova, IT.

Aldo Gangemi, Roberto Navigli, and Paola Velardi. 2003. The ontowordnet project: extension and axiomatization of conceptual relations in wordnet. In *in WordNet, Meersman*, pages 3–7. Springer.

Yoshihiko Hayashi. 2011. 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.

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. 2000a. Simple: A general framework for the development of multilingual lexicons. *International Journal of Lexicography*, 13(4):249–263.

Alessandro Lenci, F. Busa, Nilda Ruimy, E. Gola, Monica Monachini, Nicoletta Calzolari, and Antonio Zampolli. 2000b. Simple linguistic specifications. Deliverable. In: LE-SIMPLE (LE4-8346), Deliverable D2.1 & D2.2. ILC and University of Pisa, Pisa, 404 pp. 2000.

Ernesto William De Luca, Martin Eul, and Andreas Nrnberger. 2007. 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*, page 3948.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. 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. Springer-Verlag.

James Pustejovsky. 1991. The generative lexicon. *Comput. Linguist.*, 17(4):409–441, dec.

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

Antonio Toral and Monica Monachini. 2007. Simpleowl: 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.*