Towards the establishment of a linguistic linked data network for Italian

Roberto Bartolini
ILC - CNR
Via Moruzzi 1 - Pisa, Italy

Riccardo Del Gratta
ILC - CNR
Via Moruzzi 1 - Pisa, Italy

Francesca Frontini
ILC - CNR
Via Moruzzi 1 - Pisa, Italy

name.surname@ilc.cnr.it

Abstract

This paper describes the conversion of ItalwordNet and of a domain WordNet into RDF and their linking to the (L)LOD cloud and to other existing resources. A brief presentation of the resources is given, and the conversion and resulting datasets are described.

1 Introduction

Lexical Resources, both manually and automatically created, are an indispensable component to many NLP applications. In order to make lexical resources more accessible, the importance of adhering to common models has always been underlined, and in the course of time standards and best practices for the representation of such resources have emerged.

With the rise of the Semantic Web, efforts that aimed to provide common annotation and sharing formats to make resources more interoperable have found a new ally in the linked data paradigm (Berners-Lee, 2006), which generally pairs with the adoption of the RDF formalism (Lassila and Swick, 1999).

Indeed a new trend in the publication of linguistic resources as linked open data seems to be establishing itself: a survey on the formats and frameworks used in the last 20 years to exchange linguistic resources, (Lezcano et al., 2013) found “an increase in recent years in approaches adopting the Linked Data initiative”.

Although still quantitatively a minority within the linked data cloud, (Linguistic) Linked Open Data ((L)LOD)[1], (Chiarcos et al., 2011; Chiarcos, 2012), is growing and becoming a central modality for linguistic data and especially for lexical data publication. Lexicographic data may not always be big in number of triples, but they are significant in specific weight - especially the resources manually developed/checked, as they contain complex semantic information that has been encoded by humans.

Following the path of this movement, the publication of lexical resources in the Italian language has also started.

In this paper a description of the conversion of ItalwordNet and of a WordNet in the geographic domain is given.

2 Resource used for establishing a linguistic linked data network for Italian

2.1 PAROLE SIMPLE CLIPS

PAROLE SIMPLE CLIPS is a multi-layered Italian language lexicon that was the outcome of three major lexical resource projects: PAROLE (Ruimy et al., 1998) and SIMPLE (Lenci et al., 2000), two consecutive European projects, and CLIPS , an Italian national project which enlarged and refined the Italian PAROLE-SIMPLE lexicon.

The lexical information in PAROLE SIMPLE CLIPS is encoded at different descriptive levels; these are the phonetic, morphological, syntactic and semantic layers. The semantic layer of PAROLE SIMPLE CLIPS (PSC), SIMPLE, is largely based on Pustejovsky’s Generative Lexicon (GL) theory (Pustejovsky, 1991; Bel et al., 2000). This level 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: more specifically, these encode the *extended qualia structure* (Ruimy et al., 2002) and provide useful information on the semantic type of a concept (formal quale), its constituent parts (constitutive quale), on how it came into being (agentive quale) and on its purpose (telic quale). SIMPLE lexicons exist for several languages and *Usems* are consistently

[1] http://linguistics.okfn.org/resources/llod/.

linked to a common Simple Interlingual Ontology (SIO) of generic concepts labeled in English.

Recently a partial publication of the Italian PSC lexicon as RDF linked data has been carried out (del Gratta et al., 2013) and provided to the community. Other SIMPLE lexicons such as the Spanish one (Villegas and Bel, 2013) are currently also publicly available in RDF. The Simple Interlingual Ontology has been formalized into OWL by (Toral and Monachini, 2007) and it is also publicly available.

2.2 ItalwordNet

ItalwordNet (IWN) (Roventini et al., 2003) is a semantical lexical database developed along the lines of Princeton WordNet, (Fellbaum, 2010). IWN started within the EuroWordNet[2] project as the "Italian WordNet" and then subsequently refined in different Italian projects such as SI-TAL.

The ItalwordNet resource increased thanks to manually-developed mapping, known as Inter-Lingual Index (ILI), between its synsets and synsets in different WordNets (WNs). As the name suggests, the ILI is a connection among concepts in different languages. In ItalwordNet the ILI between Italian and English synsets in WordNet 1.5, (WN1.5), has been used to connect Italian to English concepts; successively, exploiting the WN1.5 to WN3.0 mapping[3], IWN and WN3.0 have been semi-automatically linked. The formalization of IWN into RDF is finalized with a partial mapping to PSC, see figure 1.

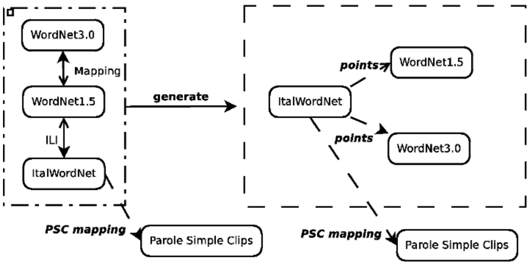


Figure 1: ILI and WordNet1.5-WN3.0 mapping expand the IWN; the mapping to PSC adds more dimensions.

3 ItalwordNet schema and dataset description

The conversion of ItalwordNet into RDF was carried out following the strategy used to convert WN into RDF, whose rules and philosophy are reported in http://www.w3.org/TR/wordnet-rdf. This schema[4] is still the reference schema for any other WN[5] and contains all objects we need to perform the conversion.

As a consequence, the proposed schema for ItalwordNet complements the one adopted for WN2.0: the main classes (*Synset*, *WordSense* and *Word*) and subclasses[6] of WordNet have been extended to address specificities of ItalwordNet. For example, the proposed schema contains additional subclasses for both *Synset* and *WordSense* to address the *ProperNoun (NP)* part of speech which is present in the ItalwordNet only, see figure 2.

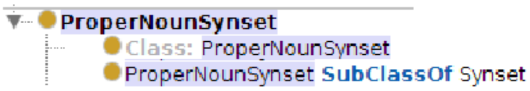

ProperNounSynset
Class: ProperNounSynset
ProperNounSynset **SubClassOf** Synset

Figure 2: IWN schema is an extension of WN2.0

Similarly the set of relations in ItalwordNet is different from the one of WordNet.

Due to the specificity of the Italian language, IWN contains relations that are not defined in WN. Relations among synsets such as "involved_location" and "be_in_state" do not exist in WN2.0 but are strongly used in IWN: as a consequence, they have been defined in the IWN schema, enforcing the concept of IWN schema as a complementing schema.

Finally, the ItalwordNet schema defines relations for managing interlingual "pointers" to WNs and links to PSC. Such relations can be both *objectProperty*, used to manage the pointer(s) between IWN and the corresponding WN3.0 synset(s) and *dataProperty*, used to managed the pointer(s) between IWN and the static value of the corresponding synset(s) in WN1.5, since this resource in not available in RDF.

[2] http://www.illc.uva.nl/EuroWordNet/.

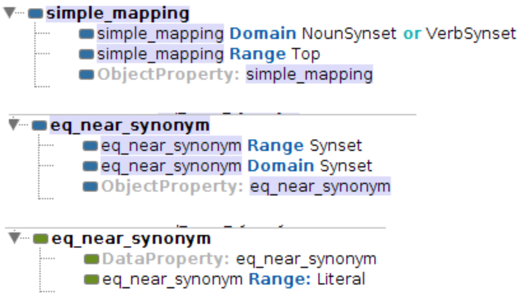
[3] The static mapping between WordNets 1.5 and 3.0 have been downloaded from http://nlp.lsi.upc.edu/tools/download-map.php.

[4] The complete schema for WN2.0 is available at http://www.w3.org/2006/03/wn/wn20/schemas/wnfull.rdfs.

[5] Cf. http://purl.org/vocabularies/princeton/wn30/, for example.

[6] Subclasses of *Synset* and *WordSense* are related to parts of speech: *Noun (N)* part of speech generates *NounSynset* and *NounWordSense* subclasses.

The PSC mapping is managed by as *objectProperty* as well, see figure 3.



```
simple_mapping
    simple_mapping Domain NounSynset or VerbSynset
    simple_mapping Range Top
    ObjectProperty: simple_mapping

eq_near_synonym
    eq_near_synonym Range Synset
    eq_near_synonym Domain Synset
    ObjectProperty: eq_near_synonym

eq_near_synonym
    DataProperty: eq_near_synonym
    eq_near_synonym Range: Literal
```

Figure 3: Object and Data properties

3.1 ItalwordNet Naming Convention

The unique identifiers for instances of *Synset*, *WordSense* and *Word* follow the syntactic pattern defined for WN2.0:

```
synset|wordsense-lexicalentry-pos-sense

word-lexicalentry
```

For example, `synset-casa-noun-1` identifies the synset whose list of members contains the sense 1 of the word *casa* (home).[7] Therefore, the Uniform Resource Identifiers (URIs) for such instances are generated by combining the basic namespace (hereafter, `base`): `www.languagelibrary.eu/owl/italWordNet15` with the keyword `instances` and the corresponding class identifiers. For example:

```
base/instances/synset-casa-noun-1
```

is the URI where the synset (identified by `synset-casa-noun-1`) is accessible. To refine the ItalwordNet resource we have defined a second namespace for its official schema[8], `iwn15schema = base/schema`, and a set of files which group the synsets according to a given relation.[9]

3.2 ItalwordNet in RDF: triples

Table 1 gives the number of effective *subject-predicate-object* triples, table 2 reports some example data in terms of obtained triples for some relations and table 3 sums all triples obtained from the relations among IWN and WN synsets as well as the one from the mapping to PSC:

Table 1: Files, units and triples

File	**Original Units**	**Triples**
synset	46, 769	148, 050
wordsenseandwords	68, 548 (wordsenses) 46, 769 (words)	367, 766

Table 2: A sample of files and obtained triples

Namespace	**File**	**Triples**
iwn15schema	has_hyponym	44, 603
iwn15schema	has_meronym	323
iwn15schema	eq_synonym	35, 653

Table 3: Internal and external relations (`iwn15schema` namespace)

Source resource	**Triples**	**Target Resource**
IWN	132, 212	IWN
	56, 074	WN1.5
	54, 717	WN3.0
	19, 896	PSC

IWN → IWN Triples as *objectProperty* encoding all internal synset-synset relations in ItalwordNet;

IWN → WN1.5 Triples as *dataProperty* encoding ILI relations;

IWN → WN3.0 Triples as *objectProperty* encoding ILI relations. The domain of the relation is a IWN synset, the range is a valid WN3.0 URI;[10]

IWN → PSC Triples as *objectProperty* encoding the IWN PSC mapping. The domain of the relation is a IWN synse, the range is a valid PSC URI.[11]

[7] The synset identified above contains 6 senses, including the one related to "casa", (home), that to "abitazione", (habitation) etc.. We selected "casa" (and its sense) to be the part of the human readable synset identifier.

[8] The schema is available at `base/schema/iwn`.

[9] For example, the file "has_hyponym" contains all couples of synsets which are connected by the "hyponym" relation.

[10] Such as http://purl.org/vocabularies/princeton/wn30/synset-chair-noun-1.

[11] such as http://www.languagelibrary.eu/owl/simple/inds/2/299/USem1450limone.

4 Geodomain resources

The Geodomain WNs were created within the framework of the GLOSS project (Frontini et al., 2012) in order to initialize a parallel terminology for the semantic annotation and mining of documents in the public security domain. The English resource was created by using the Geonames ontology[12], transforming each English label into a lexical entry, and then manually linking them to corresponding synsets.

Subsequently the English labels and glosses have been translated into Italian to produce an equivalent Italian resource.

4.1 Building a Domain WordNet

In this section we describe the strategy used to create a Domain WordNet from an human made list of domain lexical entries. The strategy follows the following steps: (i) a sense number is added to a lexical entry: in principle, we have to take care of the fact that the same lexical entry can belong to different concepts, such as for example for the lexical entry "hill" which can be both an underwater hill and a small mountain; (ii) then a referent (identifier) of the synset must be created; (iii) WordNet2.0 relations among synset are established; finally (iv) the synset previously created is connected to the concept into the Geonames ontology through the *owl:sameAs* property.

4.2 GeoDomainWN schema and dataset description

The conversion of GeoDomainWN into RDF was carried out following the steps described in section 3 but, at the moment, there is no need to create a dedicated schema, so that the provided resource will use the standard WN2.0 schema.

4.3 GeoDomainWN Naming Convention

The unique identifiers for instances of *Synset*, *WordSense* and *Word* follow the syntactic pattern defined for IWN, see section 3.1, but we have prefixed each identifier with `geo` to avoid confusion:

```
geosynset-lexicalentry-pos-sense

geowordsenselexicalentry-pos-sense

geoword-lexicalentry
```

For example, `geosynset-lago-n-1` identifies the synset whose list of members contains the sense 1 of the word *lago* (lake). Therefore, the Uniform Resource Identifiers of the resources corresponding to the main classes are obtained by combining the basic namespace (hereafter, `base`):[13]
`www.languagelibrary.eu/owl/geodomainWN/`

with the keyword `instances` and the corresponding class identifiers. For example:

```
base/instances/geosynset-lago-n-1
```

is the URI where the geosynset (identified by geosynset-lago-n-1) is accessible.

4.4 GeoDomainWN dataset description

Table 4 gives the number of effective *subject-predicate-object* triples.

Table 4: Files, units and triples

File	Original Units	Triples
synset	657	1, 971
wordsenseandwords	657 (wordsenses) 632 (words)	4, 781

Since the GeodomainWN synsets are 1 : 1 mapped onto the geonames ontology, the final resource also contains 657 relations which connect the concepts using the `owl:sameAs` property.

4.5 GeodomainWN in *lemon*

lemon (LExicon Model for ONtologies)[14] (McCrae et al., 2011) 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 according to the (L)LOD paradigm.

Following the work performed in the Monnet project[15] for creating a WordNet in *lemon*[16] we decided to transform the GeodomainWN into *lemon*. The resulting resource is a collection of *lemon* lexical entries. *lemon* lexical entries are formally equivalent to the *word* in WordNet but contain more details such as the part of speech and the explicit "narrower/broader" relations among *lemon* senses. In view of 632 units, the resulting resource contains 6, 373 triples.

[12] http://www.geonames.org/ontology/.

[13] Actually there are two different namespaces, one for Italian: `base/ita`, and one for English `base/eng`.

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

[15] www.monnet-project.eu/.

[16] http://monnetproject.deri.ie/lemonsource/wordnet.

5 Data Distribution

The lexical resources described in this paper are freely available from the *datahub* portal[17] which is synchronized with the *languagelibrary* initiative website.[18]

More specifically, interested people can directly access/download the resources from the following endpoints:

ItalwordNet from
http://datahub.io/dataset/iwn

PAROLE SIMPLE CLIPS from
http://datahub.io/dataset/simple

GeodomainWN from
http://datahub.io/dataset/geodomainwn

6 General picture

The figure 4 sums up the connections between the datasets described in this paper and the rest of the (L)LOD cloud.

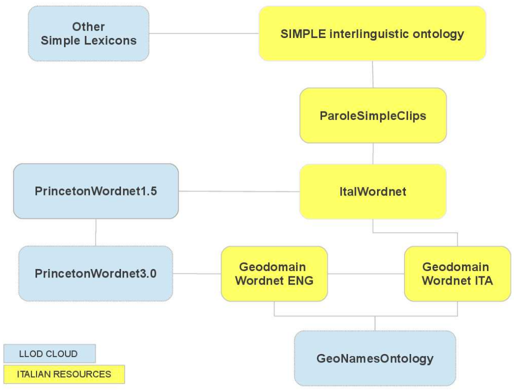


Figure 4: The linguistic linked data network for Italian

The mapping between the PAROLE SIMPLE CLIPS *Usems* and ItalwordNet synsets enriches the synset with semantic information coming from the *Usems*. The depth of information provided by the qualia structure surpasses the one available through (Ital)WordNet, and can be accessed both from Italian and from English, thanks to the IWN - WN$x.y$ mapping.

Although a direct linking between SIMPLE *Usems* in different languages is not currently available, it is imaginable that it might be automatically attempted by combining an automatic translation of the corresponding lexical form and the disambiguation that is provided by the common ontological concepts.

Finally the linking to Geonames connects the presented resources to the non linguistic linked data cloud, for example the word “lago” (lake) is connected to the geonames ontology concept “H.LKS”.

7 Conclusion and future work

In this paper we have presented three different types of Resource Description Framework (RDF) rendering.

The first one is the conversion of ItalwordNet in RDF according to the rules of the W3C consortium. The second conversion is twofold: a list of domain specific terms has been transformed into a WordNet equivalent resource and then rendered as RDF. This resource has been published also using the *lemon* model (which is the third type of rendering). This exercise will help us to serialize in *lemon* also the complete ItalwordNet resource.

Having mapped the ItalwordNet synsets into the Simple Interlingual Ontology via PSC is fundamental because it provides landscapes for interesting future works and it maps WordNet synsets onto an interesting ontological resource.

Finally the linking to Geonames offers possible applications for Named Entity Recognition and data mining: for example to solve the (Italian) (unambiguous) query such as: “Trova tutti i laghi in Toscana” (Select all lakes in Tuscany), the system uses the “lago” (lake) - H.LKS mapping to perform a query in the Geonames dataset retrieving all instances of that feature concept, namely all lakes, that are located within a specific geographic area, Toscana.

References

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 *Proceedings of LREC 2000*, Athens, Greece.

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

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open

[17] http://www.datahub.io/

[18] http://www.languagelibrary.eu

data cloud: The open linguistics working group. *TAL*, 52(3):245–275.

Christian Chiarcos. 2012. *Linked Data in Linguistics*. Springer.

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

Christiane Fellbaum. 2010. Wordnet. In Roberto Poli, Michael Healy, and Achilles Kameas, editors, *Theory and Applications of Ontology: Computer Applications*, pages 231–243. Springer Netherlands.

Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno, and Tiziana Soru. 2012. Gloss, an infrastructure for the semantic annotation and mining of documents in the public security domain. In *EEOP2012: Exploring and Exploiting Official Publications Workshop Programme*, page 21.

Ora Lassila and Ralph R. Swick. 1999. Resource description framework (RDF). model and syntax specification. Technical report, W3C, 2.

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

Leonardo Lezcano, Salvador Sanchez, and Antonio J Roa-Valverde. 2013. A survey on the exchange of linguistic resources: Publishing linguistic linked open data on the web. *Program: electronic library and information systems*, 47(3):3–3.

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.

Adriana Roventini, Antonietta Alonge, Francesca Bertagna, Nicoletta Calzolari, Christian Girardi, Bernardo Magnini, Rita Marinelli, and Antonio Zampolli. 2003. Italwordnet: building a large semantic database for the automatic treatment of italian. *Computational Linguistics in Pisa, Special Issue, XVIII-XIX, Pisa-Roma, IEPI*, 2:745–791.

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.

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, and Antonio Zampolli. 2002. Clips, a multi-level italian computational lexicon: a glimpse to data. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain.

Antonio Toral and Monica Monachini. 2007. 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*.

Marta Villegas and Nuria Bel. 2013. Parole/simple lexinfo ontology and lexicons. *Semantic Web Journal (Submitted)*.