SWRL your lexicon: adding inflectional rules to a LOD

Fahad Khan[1], Andrea Bellandi[1], Francesca Frontini[2], Monica Monachini[1]
Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR) Pisa, Italy
PRAXILING UMR 5267 Univ Paul Valéry Montpellier 3 & CNRS - Montpellier, France
name.surname@ilc.cnr.it, francesca.frontini@univ.montp3.fr

Abstract

Over the past few years the publication of lexical resources as Linked Data (LD) has taken on ever greater significance within the field of computational lexicography. So far the efforts of the community have been largely directed towards the definition of standards[1] and the conversion of single resources (see McCrae et al 2012, Khan et al 2016), but with less of a focus on the technical possibilities afforded by this new mode of publishing lexical data. However, the fact is that the Semantic Web gives us access to a whole ecosystem of standards, languages, and technologies. In this paper we will look at one of these languages, the Semantic Web Rule Language[2] (SWRL) and explore whether it might potentially play a useful role in the publication of lexical resources.

SWRL provides an extension of the Web Ontology Language (OWL) with Horn-like clauses that allows users to overcome some of OWL's expressive limitations. Even though there is a long tradition of using rule languages in computational linguistics, previous work on the use of SWRL in this domain seems to be thin on the ground (although see Wilcock 2007). And so one of the main aims of the work presented here is a better understanding of the value of including SWRL rules in an LD lexicon. Our case study concerns the morphological layer of the wide-coverage Italian lexicon Parole Simple Clips[3] (PSC-M), which provides inflectional information and morphological analysis of thousands of Italian words. What is important for our purposes here is that in addition to listing all the possible inflected forms for each word, PSC-M also contains inflectional rules which can be applied to derive these forms from the lemma. In Figure 1(a), we can see one such rule represented in LMF (Francopoulo ed. 2013).

```
<TransformSet>
  <Process>
    <feat att="operator" val="remove"/>
    <feat att="string" val="4"/>
  </Process>
  <Process>
    <feat att="operator" val="add"/>
    <feat att="string" val="IAMO"/>
  </Process>
  <GrammaticalFeatures>
    <feat att="morphofeat" val="P1IP"/>
  </GrammaticalFeatures>
</TransformSet>
```

(a)

```
hasVerbClass(?x, Class300) ^
hasStem1(?x, ?y) ^
swrlb:stringConcat(?z, ?y, "IAMO") -> hasP1IP(?x, ?z)
```

(b)

Figure 1: A rule deriving the first person plural of the present indicative for verbs such as "mangiare". (a) LMF version (b) SWRL version. In the SWRL version the premise of each rule is composed of 3 atoms: the first identifies the inflectional class of an entry, the second the stem of the entry, and the last concatenates the correct suffix for the inflected form with the correct stem.

Each rule applies to a subset of the verbs in a lexicon, those which form an inflectional class with the same behaviour; the same is true in Italian of other the parts of speech. While LMF allows such rules to be represented and associated with lexical entries it would be necessary to write specialised software to make them operational. On the other hand with SWRL rules we can use commonly available Semantic Web technologies to make patterns such as Figure 1(b) actionable. So that it is possible to enter a new word in the dataset, associate it to an inflectional class and be able to retrieve all its inflected forms.

[1] See the Ontolex Lemon final specifications (ONTOLEX - W3C. 2016 <https://www.w3.org/community/ontolex/>)
[2] https://www.w3.org/Submission/SWRL/
[3] AA. VV., 2016, PAROLE-SIMPLE-CLIPS, Digital Repository for the CLARIN Research Infrastructure provided by ILC-CNR, (http://hdl.handle.net/20.500.11752/ILC-88)

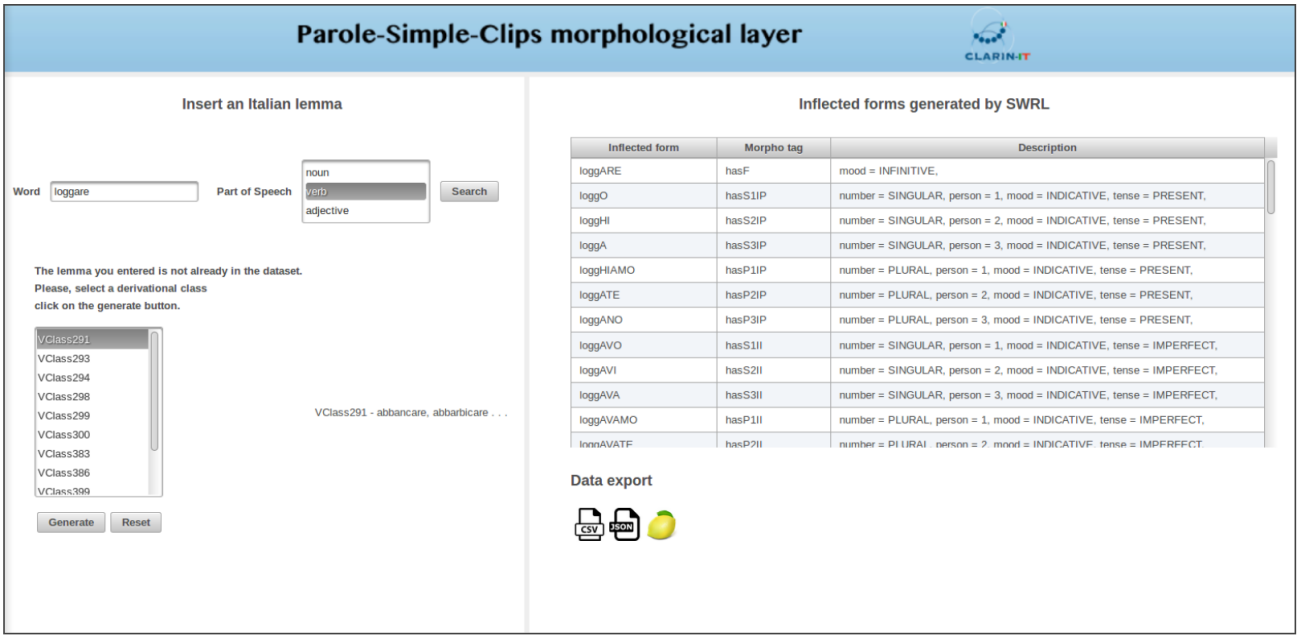

Figure 2: Inflected forms generator. The user enters the lemma "loggare" (to log in) which is not already in the dataset.

In order to demonstrate both the usefulness of the rule-based part of the lexicon as well as to make it easier for users to work with, we have developed an interface[4] for querying the dataset by entering the lemma forms of words. If the lemma exists in the dataset then the interface will show all of its inflected forms. If it doesn't however, then thanks to the fact that our morphological patterns are represented by SWRL rules, it is easy to add new lemmas to the dataset if they belong to one of the pre-existing morphological classes (see Figure 2). After inserting a lemma and its POS, the system queries a sparql endpoint and a list of relevant inflectional rules is produced[5]. For each rule example lemmas are provided in order allow the user to identify the correct inflectional class. On hitting enter, the SWRL rule generates the full inflectional paradigm on the fly.

We believe that SWRL could play an important role in overcoming some of the limitations of previous RDF-based lexicons, and so we would like to start a discussion on the implementation of rules in LD lexicons for the representation of morphology and other lexical features. In the final work we shall provide detailed information on the conversion of the PSC-M morphological rules into SWRL and on the validation that we carried out using the full inflected forms contained in the original PSC DB, as well as a number of technical details that may be of use to anyone wishing to implement SWRL rules as part of a LD dataset[6].

Acknowledgments: This work in part was funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Keywords: SWRL, inflectional morphology, Italian

References

Del Gratta, R., Frontini, F., Khan, F., Monachini, M. 2015. Converting the parole simple clips lexicon into rdf with lemon. *Semantic Web Journal* 6(4). 387–392.

Francopoulo, Gil. 2013. *LMF Lexical Markup Framework*. John Wiley & Sons.

Khan, F., Bellandi, A., Frontini, F., Monachini, M. 2017. Using SWRL Rules to Model Noun Behaviour in Italian. *LDK 2017: Language, Data, and Knowledge*, 134–142. (Lecture Notes in Computer Science). Springer, Cham.

Khan, F., Frontini, F.. 2014. Publishing PAROLE SIMPLE CLIPS as Linguistic Linked Open Data. *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, 224–228. Pisa University Press.

Khan, F., Frontini, F., Boschetti, F., Monachini, M. 2016. Converting the Liddell Scott Greek-English Lexicon into Linked Open Data using lemon. *Digital Humanities 2016: Conference Abstracts*, 593–596. Kraków: Jagiellonian

[4] A demo of the application of rules to neologism is available at http://lari-lsj.ilc.cnr.it/pscMorphoRules.

[5] Currently the system shows an unordered list of all classes to the user, but in the future we are planning to order the list in terms of relevance and remove classes which cannot be candidates.

[6] We have made a version of the lexicon available containing both the rules used to generate the lexicon and the properties that result as an RDF dump at http://lari-datasets.ilc.cnr.it/pscMorph#; a SPARQL endpoint is available at http://lari-datasets.ilc.cnr.it/pscMorph/queryForm.html.

University & Pedagogical University.
McCrae, J., Montiel-Ponsoda, E., Cimiano, P. 2012. Integrating WordNet and Wiktionary with lemon. *Linked Data in Linguistics*, 25–34. Springer.
ONTOLEX - W3C. 2016. Final Model Specification - Ontology-Lexica Community Group. https://www.w3.org/community/ontolex/wiki/Final_Model_Specification (8 June, 2017).
Ruimy, N., Corazzieri, O., Gola, E., Spanu, A., Calzolari, N., Zampolli, A.. 1998. LE-PAROLE Project: The Italian Syntactic Lexicon. *EURALEX '98*.
Wilcock, G. 2007. An OWL Ontology for HPSG. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 169–172. (ACL '07). Stroudsburg, PA, USA: Association for Computational Linguistics.Linguistics, 1–6.