# One Language to rule them all: Modelling Morphological Patterns in a Large Scale Italian Lexicon with SWRL

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

**Abstract**

We present an application of Semantic Web Technologies to computational lexicography. More precisely we describe the publication of the *morphological layer* of the Italian Parole Simple Clips lexicon (PSC-M) as linked open data. The novelty of our work is in the use of the Semantic Web Rule Language (SWRL) to encode morphological patterns, thereby allowing the automatic derivation of the inflectional variants of the entries in the lexicon. By doing so we make these patterns available in a form that is human readable and that therefore gives a comprehensive morphological description of a large number of Italian words.

**Keywords:** Morphology, Linked Open Data, Italian Lexicon, SWRL, SQWRL

## 1. Introduction

The publication of lexical resources as Linked Data (LD) has recently become a important issue in computational lexicography. However in the rush to convert all kinds of lexicons and dictionaries into RDF it is perhaps the case that the technical limitations of this mode of publication, as well as the new potentialities which it offers, are not as thoroughly understood as they ought to be. What, then, are some advantages of using Semantic Web technologies to publish lexicons that might compensate for the loss in performance which is associated with LD with respect to other representation formats? Obviously the fact that LD makes it easier to link together datasets (interoperability), and to make them'open' and publicly available (as with Linked Open Data) is a crucial factor in its favour, but then there are also important technological reasons for publishing datasets as LD. For instance the Semantic Web gives us access to a whole ecosystem of standards, languages, and technologies which we can use to work with and to explore linked data. In this paper we look at one such language, the Semantic Web Rule Language (SWRL), and explore the extent to which it might potentially be able to play a useful role in the publication of lexicographic resources. We do this by detailing the conversion into RDF and publication of the **morphological layer (PSC-M)** of the Parole Simple Clips (PSC) Italian language lexicon. While the publication of this resource will make an important source of Italian morphological data freely and openly available to researchers and the wider public, the novelty of this work is in our use of SWRL to encode morphological patterns, something that allows the automatic derivation of the inflectional variants of the entries in the lexicon. By doing so we also make these patterns available in a form that is human readable and that therefore gives a comprehensive morphological description of a large lexicon's worth of Italian words. Note that we have already presented the first stage of the conversion of PSC-M, that of the nouns, in previous work (Khan et al., 2017). In the current article we will describe the complete conversion of the PSC-M into linked open data and focus on the challenges which arose in converting the other parts of speech in the lexicon.

## 2. Background

### 2.1. Why SWRL?

SWRL is, as its name suggests, a rule language[1]. It is based on a subset of Datalog with both unary and binary predicates and is probably the best known attempt at an implementation the 'Rules' layer of the Semantic Web stack. By providing an extension of the Web Ontology Language (OWL) with Horn-like clauses SWRL permits modelers to overcome some of OWL's expressive limitations as a formalism. Although there is a long tradition of using rule languages such as Prolog in computational linguistics, previous work on use of SWRL in this domain seems to be thin on the ground (see (Wilcock, 2007)) – we speculate that this is due in large part to SWRL's own limited expressivity, at least in comparison to most of the other rule languages used in the past, and which makes it inadequate to the task of representing more complex kinds of syntactic and semantic phenomena. And so one of the core aims behind this work was to understand the viability of using SWRL rules in the modeling of at least part of a language, and more precisely to see if SWRL could help us encode part of a medium-to-large scale lexicon. What we needed in order to do this was a resource that provided us with a large number of rules which we could encode using the restricted syntax offered by SWRL.

### 2.2. Why Parole Simple Clips?

Luckily the authors of the paper had access to just such a resource, namely Parole Simple Clips (PSC), a wide-coverage, multi-layered computational lexicon for Italian that was built up within the framework of three different national and international projects[2]. After studying the composition of the lexicon it became clear to us that the conversion of PSC's morphological (PSC-M) layer into LOD

[1] https://www.w3.org/Submission/SWRL/

[2] For more information about Parole Simple Clips see http://www.ilc.cnr.it/it/content/risorse.

would provide an excellent test case for the use of SWRL[3]. What made PSC-M so attractive in this regard was the fact it featured both extensional and intensional morphological data for each of its lexical entries, and in the latter case this was in the form of representations of morphological patterns.

### 2.2.1. The Make-Up of PSC-M

In terms of size, PSC-M contains over 53,000 lexical entries and over 380,000 inflected forms. As mentioned above it also contains **inflectional schemes** representing the derivation of inflected forms from the lemma of each lexical entry. Take for instance the lexical entry for the adjective "bello" (Figure 1), PSC-M lists its inflected form "bella" as well as registering the fact that it is a feminine singular adjective. But it also links the entry with a morphological rule that can be used to generate the feminine form (*remove 1 letter from the end of the lemma and add "a"*). While the explicitly stated inflected forms are unique to each lemma (taking homonyms into consideration of course), inflectional rules can apply to more than one lexical entry. This allows us to group lexical entries together in classes based on their morphological behaviour as encoded in their respective morphological patterns. We call such classes **inflectional classes**. Each morphological pattern in PSC-M belongs to a single one of these inflectional classes and each inflectional class is associated with two or more patterns describing the derivation of the morphological variants of each of the lexical entries belonging to the class.

PSC was originally stored as an relational database which made the morphological patterns difficult for human beings to read and which also meant that the patterns weren't immediately machine actionable either. Interestingly the Lexical Markup Framework standard (Francopoulo, 2013) has a morphology module that also allows for the representation of such rules; see for instance the LMF representation of the rule to generate the present first person plural of regular "**are**" verbs below.

```
<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>
```

Unfortunately no standard XML-technology exists to derive an inflected form from its lemma using these LMF encoded rules. In contrast to this however, the authors felt that if PSC-M's morphological patterns could be successfully encoded in SWRL then it would offer us the possibility of not only publishing the morphological information in PSC in a human readable format, but of doing so in a way that made these patterns immediately machine actionable using openly available and common standards and technologies. And so it was that after a first successful experiment in encoding the morphological patterns pertaining to the nouns into SWRL, as described in (Khan et al., 2017), we decided to go ahead and encode the rest of the morphological patterns (pertaining to the parts of speech verb and adjective) into SWRL. We describe this in the next section.

[3]The full conversion of PSC into LD is still ongoing. So far, aside from our work on the morphological layer, only part of the semantic layer of PSC has been converted into LOD (Del Gratta et al., 2015; Khan and Frontini, 2014).

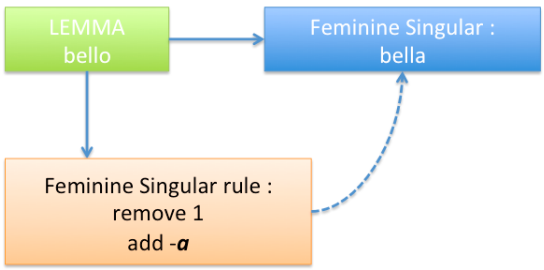


Figure 1: The content of the PSC-M in short

## 3. Modelling the PSC-M using SWRL

As in our previous experiment on converting nouns we decided that in the interests of efficiency it was better not to convert morphological patterns associated with inflectional classes containing a very small number of lexical entries, i.e., the most irregular entries, into SWRL rules. In such cases, we decided just to enumerate all the variants of an lexical entry without using rules. This meant that with respect to the nouns only the first 30 inflectional classes in the lexicon were converted into SWRL rules, offering a coverage of 96.6%, i.e., almost 97% of the nouns in the lexicon have their morphology captured by SWRL rules. The same strategy was applied to verbs and adjectives and the resulting coverage per part of speech can be seen in Table 1 where as expected the majority of the number of inflectional forms per each verb is very much larger than that for lexical entries for other parts of speech. In summary then, the vast majority of the inflected forms in the lexicon can be generated by SWRL rules and only a small number of irregular lexical entries need to have their morphological variants listed explicitly. Note also that due to the complexity of the verbal inflectional paradigms in Italian, the first 11 verbal inflectional classes have 588 rules that generate 311,543 inflected forms. In the following two subsections we look at how lexical entries and rules were encoded into linked data.

| | **Nouns** | **Adjectives** | **Verbs** |
|---|---|---|---|
| lexical forms in PSC | 76416 | 45723 | 345320 |
| lexical forms with SWRL | 73829 | 45503 | 311543 |
| nr. of SWRL rules | 79 | 64 | 588 |
| nr. of classes | 30 | 16 | 11 |
| **lexical coverage (%)** | **96.6** | **99.5** | **90.2** |

Table 1: Lexical coverage of part of speech categories.

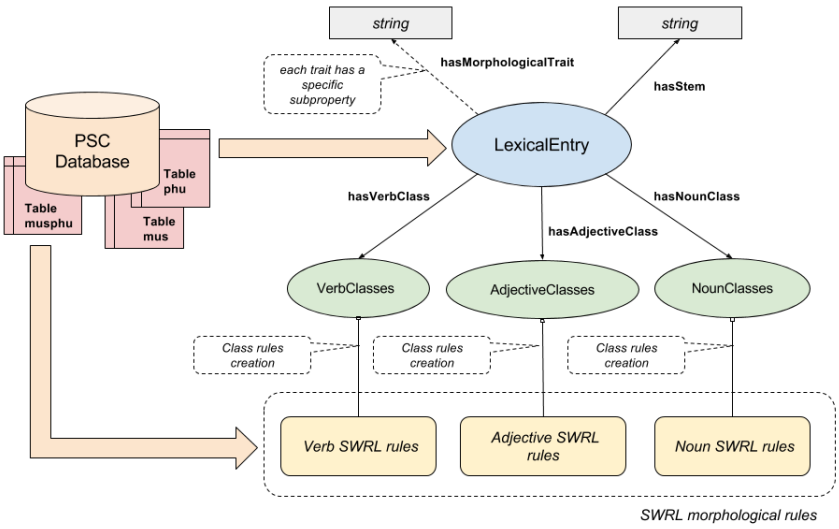


Figure 2: PSC-M to linked data conversion.

### 3.1. Lexical Entries

In this section we describe the extraction and encoding of the entries in the dataset; in the next section we look at the design and encoding of the rules. The overall lexicon schema is depicted in Figure 2(b).

For each lexical entry in the PSC-M database, amongst those covered by the classes[4], we extract its lemma form, its part of speech and the inflectional class to which it belongs. Each lexical entry is encoded as a member of the lemon[5] class *LexicalEntry*. For the three parts of speech covered by PSC-M we define individuals corresponding to each inflectional class; these are subclasses of the respective classes *NounClass*, *VerbClass*, and *AdjectiveClass*. Each lexical entry is linked to the corresponding inflection class by means of the respective properties *hasNounClass*, *hasVerbClass*, *hasAdjectiveClass*.

```
:VerbClass a owl:Class .
:VClass399 a :VerbClass,
                owl:NamedIndividual .
:hasVerbClass a owl:ObjectProperty ;
   rdfs:domain lemon:LexicalEntry ;
   rdfs:range :VerbClass
```

Although the morphological patterns defined in the original DB version of PSC-M were based on adding and removing suffixes from the lemma form, we decided that encoding this using SWRL rules would make the resulting rules too unwieldy and that the ruleset would end up being inefficient and unusable. Instead we preprocessed the lemmas by removing the suffixes given in the *remove* part of each pattern in the database in order to define a *stem* for each entry. In essence this means that our SWRL rules work by adding string suffixes to stem versions of each lemma. Accordingly we defined the datatype property *hasStem* to associate each lexical entry with its stem. In certain cases it was necessary to define more than one stem and here we use the properties *hasStem1* and *hasStem2*. So for example for the Italian verb *accendere* we have the following.

```
:accendere a lemon:LexicalEntry ;
   hasVerbClass :VClass399 ;
   hasStem1 "accend" ;
   hasStem2 "acce" .
```

PCS-M uses a specific code in order to refer to the morphological variants of a word. For instance the morphological code "S3IP" refers to the singular, third person, indicative, present form of a verb, whereas 'G' refers to the gerund. For each morphological code in PSC-M we created a corresponding datatype property, subproperty of *hasMorphologicalTrait*, in order to relate a lexical entry with its morphological variants represented as strings. The use of strings here was, once again, to simplify the SWRL rules for each inflectional class; instead of directly linking a lexical entry to its variants using a string data property, however, we could have gone via the lemon class *Form* and its property *writtenRep* but this would have made the resulting rules too complex. In the case of *accendere* the rules give us the following:

```
:abbaiare a lemon:LexicalEntry ;
   hasG "accendendo" ;
   hasS3IP "accende" .
```

### 3.2. The Rules

Once we had identified which inflectional classes we wanted to encode, we created SWRL rules for the corresponding morphological patterns. These rules work by generating new strings from an initial stem and are of the general form:

$$hasStemI(?x, ?y) \wedge hasXClass(?x, \alpha) \wedge stringConcat(?z, ?y, s)$$
$$\rightarrow hasMorphVar(?x, ?z)$$

where $hasStemI$ can either be $hasStem$, $hasStem1$ or $hasStem2$; $hasXC$ represents the appropriate data property for the part of speech to which the rule applies, e.g., $hasVerbClass$; $\alpha$ is the name of a inflectional class;

[4]For those not covered by the classes we extract all the different variant forms from the PSC database.

[5]http://lemon-model.net/

$s$ is a string; $stringConcat$ is a built in property; and $hasMorphVar$ in the head of the rule represents a data property that associates lexical entries with specific morphological variants such as e.g., $hasS3IP$. As the general form shows, the premise of each rule is composed of 3 atoms: the first identifies the stem of an entry, the second its inflectional class and the last concatenates the right suffix for the inflected form to the right stem.In the following we give one of the rules for Class 399, for generating the singular, third person, indicative, present form:

```
hasVerbClass(?x, Class399)
^ hasStem1(?x, ?y)
^ swrlb:stringConcat(?z, ?y, "E")
-> hasS3IP(?x, ?z)
```

Class 399, like all the verbal inflectional classes is associated with around 50 SWRL rules.By running all the rules for class 399 on "accendere" our lexicon is populated as follows:

```
:accendere a lemon:LexicalEntry ;
 hasVerbClass :VClass399 ;
 hasStem1 "accend" ; hasStem2 "acce" ;
 hasF "accendERE" ;
 hasFP_PP "accendENTI" ;hasFP_PR "acceSE" ;
 hasFS_PP "accendENTE" ;hasFS_PR "acceSA" ;
 hasG "accendENDO" ;
 hasMP_PP "accendENTI" ;hasMP_PR "acceSI" ;
 hasMS_PP "accendENTE" ;hasMS_PR "acceSO" ;
 hasP1CI "accendESSIMO" ;hasP1CP "accendIAMO" ;
 hasP1DP "accendEREMMO" ;hasP1IF "accendEREMO" ;
 hasP1II "accendEVAMO" ;hasP1IP "accendIAMO" ;
 hasP1IR "accendEMMO" ;
 hasP2CI "accendESTE" ;hasP2CP "accendIATE" ;
 hasP2DP "accendERESTE" ;hasP2IF "accendERETE" ;
 hasP2II "accendEVATE" ; hasP2IP "accendETE" ;
 hasP2IR "accendESTE" ;hasP2MP "accendETE" ;
 hasP3CI "accendESSERO" ; hasP3CP "accendANO" ;
 hasP3DP "accendEREBBERO" ; hasP3IF "accendERANNO" ;
 hasP3II "accendEVANO" ; hasP3IP "accendONO" ;
 hasP3IR "acceSERO" ;   hasS1CI "accendESSI" ;
 hasS1CP "accendA" ; hasS1DP "accendEREI" ;
 hasS1IF "accendERO'" ; hasS1II "accendEVO" ;
 hasS1IP "accendO" ;hasS1IR "acceSI" ;
 hasS2CI "accendESSI" ; hasS2CP "accendA" ;
 hasS2DP "accendERESTI" ;hasS2IF "accendERAI" ;
 hasS2II "accendEVI" ; hasS2IP "accendI" ;
 hasS2IR "accendESTI" ; hasS2MP "accendI" ;
 hasS3CI "accendESSE" ; hasS3CP "accendA" ;
 hasS3DP "accendEREBBE" ;hasS3IF "accendERA'" ;
 hasS3II "accendEVA" ; hasS3IP "accendE" ;
 hasS3IR "acceSE .
```

## 4. Evaluation

In order to evaluate our approach and the resulting resource we decided to consider two different aspects of the dataset. On the one hand, it was important, in order to test the effectiveness of using SWRL rules as an integral part of a medium to large sized computational lexicon, to look at time and resource consumption issues. On the other hand, we wanted to check if the rules were able to generate all and only the correct variants for each lexical entry. With regard to the first point, our lexicon currently comes in two varieties: an *empty* variety that, in the case of regular lexical entries (i,e., those entries belonging to one of the inflection classes associated with SWRL rules), does not contain any of the morphological variants associated with the entry, but only indicates the inflection class to which the entry belongs; and a *post-generation* variety that for each entry includes all its morphological variants. We decided to run the rules on the empty version of the lexicon to see how it takes to generate the lexicon. We used a PC with an Intel®Core™i7 @3.4 GHZ with 16GB of RAM. Table 2 shows the generation time for each grammatical category. As the table shows the maximum generation time is 25 seconds.

| | **Nouns** | **Adjectives** | **Verbs** |
|---|---|---|---|
| SWRL generation time (sec) | 15.8 | 6.4 | 25 |

Table 2: Time for generating all the inflected forms.

Finally, and as a sort of informal evaluation of the lexicon we ran a series of test SPARQL queries on it, of varying degrees of complexity. For instance the following query returns all the inflected forms of the lemmas (adjectives) that start with "ESP"[6].

```
SELECT ?wr ?p ?infl
WHERE {
  ?le lemon:writtenRep ?wr .
  ?le lexinfo:PartOfSpeech lexinfo:adjective .
  ?le ?p ?infl .
  ?p rdfs:subPropertyOf psc:hasAMorphologicalTrait .
  FILTER (regex(str(?wr),"^esp"))
}
```

We believe these sorts of queries reveal the usefulness of our resource for answering reasonably complicated questions about Italian morphology.

## 5. Access

We have made a post generation version of the lexicon available, containing both the rules used to generate the lexicon and the axioms 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`. By the time of the conference itself, Spring 2018, we plan to have released a number of versions of the lexicon and to have developed an interface that includes a description of the different classes, the SPARQL endpoint and a number of example queries. One of the versions of the lexicon which we plan to publish will be a post-generation version in which the morphological data is structured using both the data properties mentioned above along with the linked data morphological vocabulary MMooNN (Klimek, 2017). Another version will contain the rules and the lexicon without the generated variants allowing users to generate them for themselves as and if they require.

## 6. Conclusion

One of the main aims of the present work has been to study the viability of encoding linguistic information using SWRL in a lexical linked data resource. In this case it seems that the answer is a positive one; SWRL rules allow us to present morphological patterns in both a human readable and machine actionable form — although we will have to wait for user feedback on our resource for a more authoritative confirmation of the former. Our particular case

[6] A number of the queries can be found on the web page of the endpoint.

study was Italian inflectional morphology, but the implications of our work go beyond this limited domain. A similar approach could be applied to similar phenomena in Italian and other languages such as derivational morphology and syntactic pattern transformations. Indeed numerous derivation rules can be extracted from the information contained in PSC. One further advantage of using rules is that we are able to quickly derive inflectional paradigms for new forms by associating them to an existing rule. In the future we plan to provide online facilities to allow users to enter morphological information for new words by associating them with pre-existing inflectional classes.

## 7. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

## 8. Bibliographical References

Del Gratta, R., Frontini, F., Khan, F., and Monachini, M. (2015). Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with lemon. *Semantic web (Print)*, 6:387–392.

Francopoulo, G. (2013). *LMF Lexical Markup Framework*. John Wiley & Sons.

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

Khan, F., Bellandi, A., Frontini, F., and Monachini, M., (2017). *Using SWRL Rules to Model Noun Behaviour in Italian*, pages 134–142. Springer International Publishing, Cham.

Klimek, B. (2017). Proposing an ontolex-mmoon alignment: Towards an interconnection of two linguistic domain models. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017.*, pages 68–73.

Wilcock, G. (2007). An owl ontology for hpsg. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 169–172, Stroudsburg, PA, USA. Association for Computational Linguistics.