Using SWRL Rules to Model Noun Behaviour in Italian

Anas Fahad Khan, Andrea Bellandi, Francesca Frontini, Monica Monachini

Using SWRL Rules to Model Noun Behaviour in Italian

Fahad Khan[1], Andrea Bellandi[1], Francesca Frontini[2], and Monica Monachini[1]

[1] Istituto di Linguistica Computazionale "A. Zampolli" - CNR,
Pisa, Italy
firstname.secondname@ilc.cnr.it,
[2] Université Paul-Valry Montpellier 3, Praxiling UMR 5267 CNRS - UPVM3
France
francesca.frontini@univ.montp3.fr

Abstract. In this article we describe our ongoing attempts to use the Semantic Web Rule Language (SWRL) to model the morphological layer of a wide-coverage Italian lexical resource, Parole-Simple-Clips (PSC); in this case that subset of PSC dealing with Italian noun morphology. After giving a brief introduction to SWRL and to Italian noun morphology we go onto describe the actual transformation itself. Finally we describe an experiment on our dataset using SWRL rules and queries written in the Semantic Query-Enhanced Rule Web Language (SQWRL).

Keywords: SWRL, Italian Nouns, Morphology, Linked Open Data, SQWRL

1 Introduction

The publication of language resources as Linked (Open) Data is by now a fairly well established practice, with such large-scale resources as WordNet (in English and other languages), Wiktionary, and the Brown Corpus, already available as LOD datasets[3]. The most commonly referenced, and most commonly used, model for the conversion of lexical language resources is the Lexicon Model for Ontologies, lemon [6]. Lemon, however, does not go very far in addressing the complex problem of representing natural language morphology in RDF. Instead, this task has been taken up by the MMoOn (Multilingual Morpheme Ontology) model for encoding morphemic data in RDF, first presented in [5] where it was applied in the case of a Hebrew morpheme inventory. If we broaden our scope to take in lexical representation models that aren't native to RDF, however, then the Lexical Markup Framework (LMF), [3], in particular, boasts a highly comprehensive morphology module taking in both extensional and intensional descriptions of word morphology[4]. LMF allows users to encode morphological

[3] http://linguistic-lod.org/llod-cloud

[4] Note that here we intend 'extensional description' to refer to cases in which the inflected forms of a lexeme are explicitly given in a lexicon, and 'intensional' to cases where such forms are represented implicitly through morphological patterns that can be used to generate them.

rules using a specific formalism[5]; with the limitation that since these patterns are given as strings one would have to use a specialised parser to generate the inflected forms of each lexical entry. We believe, however, that the Semantic Web offers up significant opportunities for using already existing, freely available, and well integrated standards and technologies, in particular the Semantic Web Rule Language (SWRL) and the Semantic Query-enhanced Web Rule Language (SQWRL), to represent such morphological data in a directly machine actionable form: without having to take recourse to specialised parsers or technologies. In this paper we will look at how to use SWRL rules to represent Italian noun morphology, more specifically inflectional morphology, and show how generalisations about Italian nouns based on their inflectional behaviour can be encoded with such rules and used to generate variant noun forms using general purpose rule engines and Semantic Web reasoners[6].

2 The Semantic Web Rule Language

SWRL is a rule language that extends OWL with Horn-like clauses and is based on a subset of Datalog with unary and binary predicates[7]. It was expressly developed as a rule language for the Semantic Web – that is, to realise the "rule" segment of the semantic web stack – thus allowing users to overcome some of the expressive limitations of OWL as a language for Knowledge Engineering. Tool support for the creation of OWL knowledge bases with SWRL rules has become more readily available of late, and the popular ontology design/visualisation tool Protege now comes with SWRL and SQWRL tabs already pre-installed. SWRL was used by Wilcock to create a context free grammar parser [8], as well as in work on the modelling of rhetoric [7], but aside from these two cases it doesnt seem to have been utilised all that often in the creation of computational lexical resources. It is our hope that this article will demonstrate the potential usefulness of SWRL in constructing RDF-based lexical resources.

3 Italian Noun Morphology

Italian nouns tend to be assigned, on most morphological treatments, to classes that determine their inflectional behaviour[8]. However, word endings by themselves usually do not suffice to determine which inflectional class a noun belongs to, and this has to be inferred by observing the inflectional behaviour of the

[5] See for instance the morphological pattern for the inflection of adjectives at http://www.tagmatica.fr/lmf/FrenchLMFTestSuites1.xml

[6] Note that as we are looking at the use of SWRL explicitly as a Semantic Web-based rule language we will not, in this article, make comparisons between our work and the existing literature on modelling natural language morphology using other logic programming languages like Prolog. Our emphasis here is on making morphological data accessible using Semantic Web technologies.

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

[8] For a good introduction to Italian noun morphology see [4].

noun itself (e.g., its plural form) as well its agreement with other words (e.g., the inflectional behaviour of adjectives or determiners that are controlled by the noun in a phrase). To begin with, the word-final morpheme is a weak predictor of a word's morphological gender: although there is a general tendency for masculine words to end in '-o' and feminine words to end in '-a', this rule has several exceptions and a large number of nouns ending in '-e' and other vowels or consonants exist, for which there is an even weaker association with a given gender. So, for instance, words like *crema*, *siepe*, *crisi*, *mano*, are all feminine, but this can only be determined by observing their agreement behaviour (e,g., *la crema*, *la crisi*, *la mano*, *la siepe*). In certain rare cases, a word may have one gender in the singular and another in the plural (e.g., *un bell'uovo*, *delle belle uova*). Moreover, the morphological class to which a word is assigned determines the inflection of the noun for number [1]. So the words *poeta* and *casa*, both ending in '-a', are not only distinguished by the fact that they belong to different morphological genders (masculine and feminine) but also because they form the plural in a different way (i.e., *poeti*, *case*).

If we consider only the formation of the plural then six main inflection classes are often recognized. They have a weak association with gender, for instance the 'a/e' class contains feminine nouns. But in other cases the same behaviour in plural formation may be associated with both masculine and feminine nouns. If we create separate classes for each gender, the six classes become more numerous. Moreover, for nouns referring to humans, the formation of the feminine (singular and plural) is also to be considered. Take the words *pittore* and *cantante*; they form the plural in the same way but form the feminine in different ways. *Cantante/i* is invariable for gender, and whereas *pittore/i* becomes *pittrice/i*. In fact all human-referring nouns ending in '-tore' have the same behaviour, thus constituting a separate class. The combination of all these dimensions gives rise to a large set of inflectional classes of Italian that determine both the inflection of the noun itself and that of its accord targets.

4 The Parole-Simple-Clips Morphological Layer

The preceding section will hopefully have served to convince the reader of the complexities inherent in any morphologically driven classification of Italian nouns – as well of the usefulness of lexical resources that render this kind of knowledge accessible to researchers and to language learners. This, then, leads us onto the next topic which we wish to touch upon in this article, and which relates to the publication of legacy lexical resources as Linked Open Data. The legacy resource in this instance, Parole-Simple-Clips (PSC), constitutes a large, wide-coverage, multi-layered computational lexicon for Italian that was built up, and then extended upon, in successive stages through the course of three major, international and national, projects. In previous work we have described the publication of part of the semantic layer of PSC as LOD[2]. However, PSC also contains a rich morphological layer that contains a substantial amount of structured morphological data for 72,001 lemmas, amongst which there are 48,735

nouns, and where 33,362 of these noun lemmas belong to a specific morphological pattern class. We were inspired by the structure and make up of PSCs morphological layer, and in particular the fact that it contained both extensional and intensional morphological data – the latter in the form of representations of morphological patterns – to attempt to publish it, or at least that part of it pertaining to Italian noun morphology, using SWRL rules.

Each lexeme in the original PSC dataset is classified according to its morphological behaviour; in the case of nouns this means that each lexeme is associated with an abstract class, which is described in terms of a set of operations on strings. So that, to take an example, one of these classes with the id number 279, encompasses two different transformation operations, which are as follows:

- (279a) Remove the last two characters from the string; add IO to the string; assign the feature Masculine Singular to the result;
- (279b) Remove the last two characters from the string; add II to the string assign the feature Masculine Plural to the result.

One of the lexical entries belonging to class 279 is the beautiful Italian word *scombussolio*, meaning 'muddle'. In this case the first operation (279a) is redundant when performed on the lemma, in the sense of that it doesnt alter the original string – although it does give us important information about the endings of the masculine singular forms of this class; the second operation, instead gives us the plural *scombussolii*. In other cases, these class descriptions include rules that enable the derivation of strings giving us both male and female forms of nouns. So for example class 226 allows us, given the lemma form *bischero* of the popular Tuscan word meaning 'fool' or 'prick', to derive the masculine plural *bischeri* as well as the feminine singular *bischera* and the feminine plural *bischere*.

There exist 171 such classes covering the totality of the nouns in PSC; 86 of these classes only have one member and 124 of them contain 10 members or less: instead the top 30 most productive classes together cover around 98% of the nouns in the PSC dataset with a morphological class assigned.

5 Transforming the PSC Nouns Morphological Layer into Linked Open Data

Although we had originally intended to model the ostensibly more complex morphology of Latin nouns using SWRL rules, we eventually settled on Italian nouns as our first case study, given that we had the PSC morphological data already to hand – a comparable Latin dataset being seemingly much harder to come by – and given the non-trivial challenges that modeling Italian nouns offers in terms of the potential number of rules and exceptions that one has to deal with. The basic idea of the case study was to take the nouns in PSC, arrange them in classes, based in this instance on the classes already given in PSC, and then to create SWRL rules referencing these classes that would then allow us to derive the various different inflected forms of each individual noun. A number of

practical issues became immediately apparent on first working with the PSC dataset, the first of which was the fact that, as SWRL doesnt allow the 'creation' of new OWL individuals representing abstract forms. We had to instead focus on generating new strings to represent the inflected forms of the original lemmas. Consequently, we devised a series of rules where the head of each rule was a clause associating an individual lexical entry with a string instantiating one of the different forms of the noun using an appropriate datatype property. We created a family of datatype properties that related variant forms as strings with lexical entries: such properties as `hasForm`, `hasStem`, `hasLemma`, `hasPlural` etc. This idea can in principle be extended to other lexical categories in Italian, so that for example we can create subproperties of verbal form for different combinations of person, aspect, tense and mood, and indeed we intend to pursue this line of research in future work.

In order to make the noun morphology rules as concise as possible, we decided to assume that the stem for each lexical entry was already in the lexicon. We created these stems by pre-processing the lemma strings contained in the original dataset according to the original class transformation relations, and then associating the resulting strings with members of the class LexicalEntry using the newly created datatype property `hasStem`. This enabled us to create SWRL rules that derived both the lemma and the plural as well as, in appropriate instances, the feminine singular and/or plural of a given lexical entry: for this purpose we created the properties `hasFemaleLemma` and `hasFemalePlural`, siblings of `hasLemma` and `hasPlural`, all of which are collectively children of `hasForm`. For instance the rules 279a and 279b, associated with Class 25 in our classification, are represented respectively by the following two rules in the knowledge base:

```
hasStem(?x, ?y)^swrlb:stringConcat(?z, ?
y,"IO"^^xsd:string)^hasNounClass(?x,Class25)->hasLemma(?x, ?z)

hasStem(?x, ?y)^swrlb:stringConcat(?z, ?y,
"II"^^xsd:string)^hasNounClass(?x,Class25)->hasPlural(?x, ?z)
```

Note here our use of the built-in SWRL string method `stringConcat` which allows for the generation of lemma and plural forms through the classification of these forms as the concatentation of the stem with a string ending. So that given that the lexical entry for *scombussolio* belongs to Class 25 we are able to derive both lemma (singular) and plural forms from the stem *scumbussol*.

In cases where we have an additional female form, such as Class 8 we can have four rules:

```
hasStem(?x, ?y)^swrlb:stringConcat(?z, ?y,
"I"^^xsd:string)^hasNounClass(?x,Class8)->hasPlural(?x, ?z)

hasStem(?x, ?y)^swrlb:stringConcat(?z, ?y,
"O"^^xsd:string)^hasNounClass(?x,Class8)->hasLemma(?x, ?z)

hasStem(?x, ?y)^swrlb:stringConcat(?z, ?y,
"A"^^xsd:string)^hasNounClass(?x,Class8)->hasFemaleLemma(?x, ?z)
```

```
hasStem(?x, ?y)^swrlb:stringConcat(?z, ?y,
"E"^^xsd:string)^hasNounClass(?x,Class8)->hasFemalePlural(?x, ?z)
```

In the interests of efficiency, and from a desire to generalise over the levels of detail that we found in the PSC, we decided to encode only the first 30 classes – that is the first 30 classes in terms of number of members – as rules; the rest of the 125 classes have a very low productivity, with 86 of them having only 1 member. Indeed the former 30 classes give us a coverage of over 98% of the nouns in the lexicon belonging to a morphological class (which is 67% of the total number of nouns); in the remaining strongly irregular cases, individual inflectional forms are explicitly associated with the given lexical entries, i.e., we make use of extensional definitions. We believe that these 30 classes offer a comprehensive description of the noun morphology of Italian; each set of rules embodying both a description of the behaviour of Italian morphology and a means of deriving them: a means of deriving these inflectional forms, moreover, that uses general purpose Semantic Web technologies and that therefore renders this description much more accessible than would be otherwise possible. Using a rule engine it is possible to derive new OWL axioms from these rules and to subsequently add these to the morphological knowledge base resulting in a greatly expanded knowledge base. Another possibility would be to not generate new OWL axioms but to instead run SQWRL queries over our hybrid SWRL/OWL knowledge. SQWRL is a powerful query language whose syntax is based on SWRL itself. It can be used to run simple queries such as the following query, Query 1, which counts all the lexical entries in the lexical knowledge base:

```
LexicalEntry(?l) → sqwrl:count(?l).
```

we can also write more complicated queries such as the following query, Query 2, which finds all male Nouns with lemmas ending in 'O':

```
hasGender(?l,Male)^hasLemma(?l,?a) ^swrlb:endsWith(?a,"O") → sqwrl:select(?l).
```

The next query, Query 3, instead finds all male Nouns with a Plural ending in 'A'

```
hasGender(?l, Male)^hasPlural(?l, ?a)^swrlb:endsWith(?a, "A") → sqwrl:select(?l).
```

Query 4 finds all male nouns with a female form and give their female plurals:

```
hasFemalePlural(?l,?f) → sqwrl:select(?l,?f).
```

We can also find all nouns with female forms and give their female plurals, such as in the following query, Query 5:

```
hasFemalePlural(?l,?f) → sqwrl:select(?l,?f).
```

5.1 Practicalities

In order to understand the viability of this approach it was necessary to have some idea of the practicalities of using SWRL and SQWRL on realistically sized lexical knowledge bases such as the PSC morphological layer, especially with regards to time and resource consumption. To this end we used our SWRL rules along with a dataset consisting of the nominal lexical entries in the top 30 noun classes to test rule engine execution using the OWL API and the SWRL Rule Engine API's[9]; in addition we used the same lexical knowledge base to look at the time taken to respond to the SQWRL queries given above using the SQWRL API[10]. We carried out the experiment using two different configurations: the first, Configuration A, consisted of a Mac laptop with an Intel®Core™M3 @1.1GHz with 8GB RAM; the second, a PC with an Intel®Core™i7 @3.4 GHZ with 16GB of RAM. Table 1 gives the results for each of the two configurations on different numbers of nouns where the percentage coverage[11] is also given in each case. The second table, Table 2 gives the execution time for the five queries which

		Configuration A	**Configuration B**
#Nouns	**Coverage**	**Generation time (sec.)**	**Generation time (sec.)**
3,000	9%	6,19	2,92
10,607	31%	15,98	5,21
16,388	49%	19,41	8,94
29,789	89%	34,49	13,47
32,605	98%	53,86	14,18

Table 1. Generation time of SWRL rules.

we listed above, but only for Configuration B. We tried running the queries on Configuration A but this led to a time out. Time constraints prevented us from carrying out other tests, but the results so far seem to be hopeful, at least to some extent.

6 Further Work

In this article we have presented a method for encoding morphological patterns using the SWRL rule language. We hope that we have been able to demonstrate something of the viability of this approach as a means of providing a directly machine actionable classification and description of the morphology of a language,

[9] These API's can be found respectively at http://owlapi.sourceforge.net/ and https://github.com/protegeproject/swrlapi

[10] https://github.com/protegeproject/swrlapi/wiki

[11] This percentage is taken over the total number of nouns assigned to a morphological class

Query ID	Execution time (sec.)
Query 1	14.6
Query 2	14.6
Query 3	14.3
Query 4	14.6
Query 5	14.6

Table 2. Response time of SQWRL queries on Configuration B.

at least insofar as it pertains to the modeling of Italian noun morphology, and consequently of other languages with a similar nominal morphology. In future work we plan to use SWRL to represent the morphology of other parts of speech in Italian, again using PSC as our foundational dataset, with a view to publishing the whole PSC morphological layer as LOD. We also plan to apply our rule-based approach to Latin, a language whose complex nominal morphology would seem, on first sight, to offer a much greater challenge for this approach, in order to see how well it can carry over. A further challenge would be to see whether we can further generalise the approach to languages with markedly different morphological systems, such as for instance languages with root and pattern morphologies such as Arabic and Hebrew.

References

1. D'Achille, P., Thornton, A.M.: La flessione del nome dallitaliano antico allitaliano contemporaneo (2003)
2. Del Gratta, R., Frontini, F., Khan, F., Monachini, M.: Converting the PAROLE SIMPLE CLIPS Lexicon into RDF with lemon. Semantic web (Print) 6, 387–392 (2015)
3. Francopoulo, G.: LMF Lexical Markup Framework. John Wiley & Sons (2013)
4. Iacobini, C., Thornton, A.M.: Morfologia e formazione delle parole. Manuale di linguistica italiana 13, 190 (2016)
5. Klimek, B., Arndt, N., Krause, S., Arndt, T.: Creating linked data morphological language resources with mmoon - the hebrew morpheme inventory. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016) (2016)
6. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. In: The semantic web: research and applications, pp. 245–259. Springer (2011)
7. O'Reilly, C., Paurobally, S.: Lassoing rhetoric with owl and swrl. Unpublished MSc dissertation. Available: http://computationalrhetoricworkshop. uwaterloo. ca/wpcontent/uploads/2016/06/LassoingRhetoricWithOWLAndSWR L. pdf (2010)
8. Wilcock, G.: An owl ontology for hpsg. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 169–172. ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007)