Automatized Merging of Italian Lexical Resources

Thierry Declerck[1,2], Stefania Racioppa[1], Karlheinz Mörth[2]
[1] DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken, Germany
[2] Institute for Corpus Linguistics and Text Technology (ICLTT), Austrian Academy of Science
Sonnenfelsgasse 19/8, 1010 Vienna, Austria
E-mail: declerck@dfki.de, stefania.racioppa@dfki.de, karlheinz.moerth@oeaw.ac.at

Abstract

In the context of a recently started European project, TrendMiner, there is a need for a large lexical coverage of various languages, among those the Italian language. The lexicon should include morphological, syntactic and semantic information, but also features for representing the level of opinion or sentiment that can be expressed by the lexical entries. Since there is no yet ready to use such lexicon, we investigated the possibility to access and merge various Italian lexical resources. A departure point was the freely available Morph-it! lexicon, which is containing inflected forms with their lemma and morphological features. We transformed the textual format of Morph-it! onto a database schema, in order to support integration process with other resources. We then considered Italian lexicon entries available in various versions of Wiktionary for adding further information, like origin, uses and senses of the entries. We explore the need to have a standardized representation of lexical resources in order to better integrate the various lexical information from the distinct sources, and we also describe a first conversion of the lexical information onto a computational lexicon.

Keywords: Lexical Resources, Standards, Computational Lexicon

1. Introduction

In the context of a recently launched European R&D project, TrendMiner[1], there is a need for a large lexical coverage of Italian language. The lexical resources should include information about morphology, syntax and semantics, but also about opinions or sentiments that can be carried by the entries. The lexicon should also be easily extendable to new types of expressions, like those occurring in micro-blogs, twitter etc. We are therefore experimenting with integration issues of existing lexical resources, starting for now with good quality lexical data available from both language specialists and collaborative efforts. In a next step we will investigate how to integrate in a lexical framework "lower quality" or "noisy" lexical data, as these are typically used in short messaging frameworks or other forms of social media.

2. The First Set of Resources

A starting point for our work was a set of Italian resources made available to the NooJ community. Those relatively limited resources[2] gave us in first line the representation format for the NooJ resources, both for lexical entries and inflexion paradigms, against which we could start the porting of a larger available Italian lexicon, Morph-it![3], which contains more than 35.000 lemmas.

The textual format of the Morph-it! lexicon consists in a list of triples displaying a full-form, its corresponding lemma and the associated morpho-syntactic information, as can be seen in the examples in Table 1.

```
casco casco NOUN-M:s
caschi casco NOUN-M:p
…
casellari casellario ADJ:pos+m+p
casellari casellario NOUN-M:p
casellaria casellario ADJ:pos+f+s
casellarie casellario ADJ:pos+f+p
casellario casellario ADJ:pos+m+s
casellario casellario NOUN-M:s
casellarissima casellario ADJ:sup+f+s
casellarissime casellario ADJ:sup+f+p
casellarissimi casellario ADJ:sup+m+p
casellarissimo casellario ADJ:sup+m+s
```

Table 1: Examples of lexical entries in Morph-it!

We wrote a script for transforming the Morphit-it! textual representation into a hash table, with the lemmas used as the keys. This representation is more compact, since lemmas are not repeated as often as they have distinct full-form realizations, and our intermediate format is also giving a basic linguistic interpretation for the listed language data, allowing also marking explicitly ambiguities. An example of this intermediate format is given in Table 2.

```
"casco"  => {
    "NOUN"  => {
        "Infl_1"  => {
             "caschi" => "m+p",
        },
        "Infl_2"  => {
             "casco" => "m+s",
...
"casellario"  => {
    "NOUN"  => {
        "Infl_1"  => {
             "casellari" => "m+p",
        },
```

[1] http://www.trendminer-project.eu
[2] http://www.nooj4nlp.net/pages/italian.html. In the meantime, the author of the Italian resources for NooJ uploaded a much larger resource, which is for the time being available only in the compiled format, and therefore not usable for our experiment.
[3] http://dev.sslmit.unibo.it/linguistics/morph-it.php. See also (Zanchetta & Baroni, 2005)

```
        "Infl_2"  => {
             "casellario" => "m+s",
        },
    },
    "ADJ"  => {
        "Infl_1"  => {
             "casellari" => "pos+m+p",
        },
        "Infl_2"  => {
             "casellaria" => "pos+f+s",
    …
```

Table 2: Intermediate representation resulting from a transformation of Morph-it! Basic linguistic information is marked-up, contrary to the original format.

A second step consisted in computing the string differences between the lemma and the set of associated full-forms. This information is important in the case we want to use the lexicon in the context of a finite state machine (FST) platform, like this is the case in NooJ. The computed string differences are encoded in the form of morphological operations, that are performed by the FST engine in order to generate full-forms, as can be seen in Table 3. There, the value of the "fst" element in the first case tells that the engine processing the lemma "casco" has to go back one character (starting from the end of the lemma), delete the character that has been consumed, and add the letters "h" and "i" to the remaining of the string, and to mark the new word form with the inflectional values "m" and "p".

The "<E>" symbol in the second case specifies that no string operation is defined, and that the lemma and the full-form are thus identical, the latter being morphologically marked as "m" and "s".

```
"casco"  => {
        "NOUN"  => {
            "Infl_1"  => {
                 "fst" => "<B1>hi/+m+p",
                 "caschi" => "m+p",
            },
            "Infl_2"  => {
                 "fst" => "<E>/+m+s",
                 "casco" => "m+s",
            },
        ...
```

Table 3: Adding to the intermediate representation procedural information for the generation of full-forms.

At this level, we included thus some "operational" information to the lexicon, but this in a modular way. To use the LMF[4] terminology: we can consider this module describing operational information as being an extension of the core lexicon.

In NooJ, all those operational information can be encoded in inflectional paradigms, so that all the lemmas generating the same type of full-forms can share a unique paradigm, like for example the nominal lemmas "casco" and "carico" (and many other lemmas) are sharing the inflectional paradigm "NOUN_132", while the paradigm is specifying the concrete string operation (see Table 4)

[4] http://www.lexicalmarkupframework.org

```
carico,NOUN+FLX=NOUN_132

casco,NOUN+FLX=NOUN_132

------------

NOUN_132 = <B1>hi/+m+p + <E>/+m+s ;
```

Table 4: Sharing of a inflectional paradigm (NOUN_132) over various entries

The actual NooJ version of Morph-it contains all the main classes, and more specifically 6072 verbs, 17443 nouns and 9385 adjectives. The compiled inflected dictionary has 657062/12155 states and recognizes 442629 forms.

3. The Second Set of Resources: Entries in Wiktionary

As one could see from its description above, semantic information is not encoded in Morph-it! In order to palliate this lack of information, we searched for other freely available lexical sources, and we drove our attention to Witkionary. We didn't take Witkionary as our first source, assuming that the morpho-syntactic information encoded in Morph-it! is of a higher quality.

And in general, a drawback of the Wiktionary project is that the content of its lexical databases is formatted in a lightweight mark-up system commonly used in Wiki applications. This mark-up system is neither standardized nor very structure-oriented. To acerbate the situation, it is often applied in a considerably inconsistent manner, which makes extracting structured lexical information a really challenging task. But we consider Wiktionary still as a good source, also improving and in constant extension: We also discovered that the Italian Wiktionary[5] is one of the largest Wiktionary resources at all. Therefore we went into the task of porting the XML dump of this resource into our internal format. We extracted 29639 purely Italian entries; all encoded as lemma, and did not consider the full-form entries. An example of an entry we extract from the XML dump:

```
<page>
    <title>casco</title>
    <id>162499</id>
    <revision>
<id>1112205</id>

<timestamp>2011-10-29T05:27:52Z</timestamp>
      <contributor>
        <username>Ulisse</username>
        <id>18921</id>
      </contributor>
      <text
xml:space="preserve">{{in|it|noun}}
{{pn | w}} ''m sing '' {{linkp|caschi}}
# {{term|abbigliamento|it}} [[copricapo]]
difensivo atto a proteggere la [[testa]] da
urti
# particolare tipo di [[assicurazione]] che
copre anche i danni causati dal [[conducente]]
di un [[autoveicolo]] nei confronti del
medesimo
```

[5] http://it.wiktionary.org/wiki/Pagina_principale. The Italian Wiktionary (like other Wiktionaries) contains entries for many languages, but with all the associated information written in Italian: Therefore the use of the name "Italian Wiktionary".

```
# tipo di pettinatura femminile a forma di
{{pn}}
{{-hyph-}}
; cÃ  | sco
{{-etim-}}
dallo spagnolo [[casco]] di etimo incerto
{{-rel-}}
```

Table 5: An example of an entry in the Italian Wiktionary. Entries also have information about etymology, semantics, translation, etc., all of which can not be displayed here

We also transform this data representation onto a hash table, in order to allow comparisons with the data we already got from NooJ and Morph-it! Our main attention in this case is given to the acquisition of semantic information. An example of the transformation from the XML dump onto the machine readable hash table is given in Table 6.

```
"28033"
    => "casco" :: pos = noun {
              pl => caschi
              morph => m sing
             semantic[1] => [[copricapo]]
difensivo atto a proteggere la [[testa]] da
urti
             semantic[2] => particolare tipo
di [[assicurazione]] che copre anche i danni
causati dal    [[conducente]] di un
[[autoveicolo]] nei confronti del medesimo
             semantic[3] => tipo di
pettinatura femminile a forma di {{pn}}
              term[1] => abbigliamento
       synonym[1] => [[elmo]],
[[copricapo]], [[asciugacapelli]] ….
```

Table 6: Transformation of the XML dump of Wiktionary. Marking explicitly certain properties and re-organizing the distribution of information.

From this hash it is then easy to attach the semantic information to the already ported lexical entries from Morph-it! and to encode it also in the NooJ format[6], just extending slightly our script.

We mentioned above that encoding in Wiktionary is not always consistent. An example is given by the entries “blu”, which is associated to the semantic term “colore”, and “bianco” which is associated to the semantic term “colori”. There is a need for harmonization of the naming of the semantic categories. And further it would be better to use an Interlingua for naming related semantic categories.

[6] A reviewer of our submission very correctly noticed: things are not so easy, when one has to integrate semantic information in a lexicon that has already such information available. Decisions have to be taken, and it is not obvious how to deal with this aspect in an automated fashion. We will very soon attack this problem, also along the lines of very recently announced lexical resources, for English and German, which are integrating semantic information from various sources, like FrameNet and Wiktionary: UBY 1.0 - a large-scale lexical-semantic resource for natural language processing. See http://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/ or (Gurevych et al., 2012)

Fortunately the Wikimedia foundation has foreseen such a system, so that all the language specific Wiktionaries can point to a unique set of descriptors (in English) for semantic categories[7], while keeping the origin of the pointing with the use of standardized language codes. Nevertheless not a lot of contributors do this.

We further decided then to test the extraction of Italian entries from the English Wiktionary. Since the representation format of the English lexicon is different from the Italian one, we had to adapt our extraction and transformation script. We can extract the high number of 463480 Italian entries, and we are in the process of reducing this number to the entries being in fact lemmas.

An example in our intermediate hash format of an Italian entry we extract from the XML dump of the English Wiktionary is shown in Table 7

```
"13837"
    => “spumante” :: pos = Adjective {
             morph = {{it-adj|spumant|e|i}}
             transl = EN =foaming
    }
    =>  “spumante” :: pos = Noun {
             morph =
{{it-noun|spumant|m|e|i}}
             transl = EN = sparkling wine
    }
    =>  “spumante” :: pos = Verb
             morph = {{present participle
of|[[spumare#Italian|spumare]]|lang=it}}
    }
    => Related Topics: * [[frizzante]]
    => Category: [[Category:en:Wines]]
```

Table 7: An example of an Italian entry in the English Wiktionary, in our intermediate harmonized format

The reader can get an idea of the disparity of information encodings using in different editions of the Wiktionary dictionaries, when looking at the entry in the Italian lexicon (Table 8).

```
"2141"
    => "spumante" :: pos = agg {
              morph => m
    }
    => "spumante" :: pos = noun {
              morph => m
              pl => spumanti
    }
```

Table 8: The entry “spumante” in the Italian Wiktionary to be compared to the entry in the English version in Table 7

Our actual work consists in mapping the tagset from the Italian Wiktionary to the tagset of the English Wiktionary, as the basis for merging both lexicons. At the same time we will add a link to the ISO Data Categories (http://www.isocat.org/) for ensuring the re-usability of the tagset.

On the basis of the semantic categorization proposed by Wiktionary and the mapping of these category descriptors to the categories suggested in the language specific

[7] http://en.wiktionary.org/wiki/Category:All_topics

Wiktionaries, we also started to extract a multilingual Wiktionary-Net, which could be combined with WordNet (http://wordnet.princeton.edu/)[8]. And last but not least we are establishing a machine readable translation dictionary (IT <-> EN).

4. Standardization

In this submission we stressed our need to get relatively quickly a large Italian lexicon running on the platform used in the project. And although can report on successful and promising work, we are aware that some solutions are still ad-hoc, since the approach we described was motivated first by pragmatic needs. We identified clearly the need to propose, beyond the actual implementation in the context of a specific platform, more standardized representations. We mentioned already LMF and we are in the process of porting the basic lexical information of our merged lexicon onto the LMF model. Additionally we will map the used tagset onto the ISO Data Categories, and include this information into the LMF representation.

An additional plan consist in making the extracted and integrated lexical information in the context of the Linked Open Data initiatives active in the field of language resources. Some works in this direction have been presented at the recent Workshop "Linked Data in Linguistics"[9]. In this context a main effort consists in publishing linguistic data using W3C standards like RDF and SKOS[10]. An example of such work is given in (McCrae et. al, 2012).

But we first started with the porting of our lexical information onto TEI (P5), since some work as already been done in this respect at ICLTT, also in order to make our work easily available to the Digital Humanities community, which is making an heaving use of text annotation properties introduced in TEI. For now, the German Wiktionary has been converted into TEI (P5)[11], also making use of standardized feature structures (a joint work by ISO and TEI standardization bodies), especially for the representation of morpho-syntactic features, following the recommendation of ISO-MAF, which is not yet an established standard. The user can access the data both via a GUI and via a XML download[12]. We plan to achieve the same results for the merged Italian lexicon as the next step of our work, after we merged the entries from both the English and the Italian Wiktionary resources.

[8] As mentioned in footnote 6, we will have detailed look at the recent developments described in the work of (Gurevych et al., 2012)

[9] http://ldl2012.lod2.eu/

[10] See both http://www.w3.org/RDF/ and http://www.w3.org/2004/02/skos/

[11] Result of this work can be seen at: http://corpus3.aac.ac.at/showcase/index.php/wiktionaryconvertor

[12] See http://www.tei-c.org/Guidelines/P5/

5. Conclusion

We presented an approach for integrating various Italian resources, in the context of concrete needs. Beyond this we identified ways for publishing results of our work in standardized representations that can be used by the NLP community at large. We will establish concrete cooperation with initiatives like UBY (in the context of ISO standards) or LDL (in the context of W3C standards).

6. Acknowledgements

This work has been partly supported by R&D project TrendMiner, which is co-funded by the European Commission under the contract nr. 287863.

7. References

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C.M., Wirth, C. (2012). A Large-Scale Unified Lexical-Semantic Resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Avignon.

Krizhanovsky, A. (2010). The comparison of Wiktionary thesauri transformed into the machine-readable format. (http://arxiv.org/abs/1006.5040)

Krizhanovsky, A., Lin F (2009). Related terms search based on WordNet / Wiktionary and its application in ontology matching. In: Proceedings of the 11th Russian conference on Digital Libraries (RCDL 2009).

McCrae, J., Montiel-Ponsoda, E., Cimiano, P. (2012). IntegratingWordNet andWiktionary with Lemon. In Proceedings of the Workshop **"***Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata***"** (LDL).

Meyer, C.M., Gurevych, I. (2010): Worth its Weight in Gold or yet another resource – a comparative study of Wiktionary, OpenThesaurus and Germanet. In: *Proceedings of the 11th International conference on Intelligent Text Processing and Computational Linguistics.* Iasi (Romania) 2010: pp. 38-49

Moerth, K., Declerck, T., Lendvai, P., Váradi, T. (2011)**:** Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts. In: *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web* (Bonn 2011): 80-85.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S.-K., Kuo, T.-Y., Magistry, P., Huang, C.-R. (2009). Wiktionary and NLP: Improving synonymy networks. In: *Proceedings of the 2009 Workshop on Peoples's Web Meets NLP, ACL-IJCNLP*. Singapore: pp. 19-27.

Zanchetta, E., Baroni, M. (2005). Morph-it! A free corpus-based morphological resource for the Italian language. In *Proceedings of Corpus Linguistics 2005*, University of Birmingham, Birmingham, UK.

Zesch T., Mueller C., Gurevych I. (2008a). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In: *Proceedings of the Conference on Language Resources and Evaluation*. LREC 2008.

Zesch T., Mueller C., Gurevych I. (2008b). Using Wiktionary for computing semantic relatedness. In: *Proceedings of 23rd AAAI conference on Artificial Intelligenc*