GLAW-IT: a Free Large Italian Dictionary Encoded in a Fine-Grained XML Format

Basilio Calderone, Franck Sajous, Nabil Hathout

GLAW-IT
A Free Large Italian Dictionary Encoded in a Fine-Grained XML Format

Basilio Calderone, Franck Sajous and Nabil Hathout
(CLLE-ERSS - CNRS & Université de Toulouse 2)

We present GLAW-IT[1] (*Un Grande Lessico Automaticamente estratto da Wiktionary Italia*), a free large dictionary extracted from Wikizionario, the Italian edition of the collaborative dictionary Wiktionary.

The rise and quick expansion of corpus linguistics and data-driven approaches made lexical resources essential for quantitative and qualitative language studies. Italian language suffers from data scarceness: most of the lexical resources are not freely available (Bel et al., 2000; Roventini et al., 2000) while others focus on particular aspects: morphology and syntax (Talamo et al., 2016; Zanchetta & Baroni, 2005; Bertinetto et al., 2005), phonology transcriptions (Goslin et al., 2014), argument structures (Lenci et al., 2012). To our knowledge no workable, free and large Italian dictionary combines different levels of linguistic information. This lack motivated the creation of GLAW-IT, produced by collecting Wikizionario's lexicographical material. As Sajous & Hathout (2015), we adopted a lexicographical perspective focused on the integration of different levels of linguistic information in a coherent, structured and homogeneous format, rather than resorting to automated translation, aggregation and alignment of heterogeneous resources, such as BabelNet (Navigli & Ponzetto, 2010).

As a result, GLAW-IT is a machine-readable dictionary structured in an XML format encoding the micro- and macrostructure of Wikizionario's articles (an example is given in Figure 1).

It contains almost 320,000 wordforms with sections reporting:

(1) Definitions (glosses) including linguistic labels (attitudinal, diatopic, diachronic information and specialized domain markers)
(2) Usage examples
(3) Parts of speech including morphosyntactic features
(4) Phonemic transcriptions in IPA
(5) Inflectional paradigms: the complete set of the inflected forms related to lemmas
(6) Morphologically or semantically related words
(7) Etymology
(8) Translations in different languages

GLAW-IT is the only free machine-readable dictionary for Italian. Coverage evaluation shows GLAW-IT's superiority and qualitative comparisons to other Italian lexicons confirm the overall quality and reliability of GLAW-IT, notably concerning the consistency of the phonemic transcriptions. To our knowledge, only the recent lexicon PhonItalia (Goslin et al., 2014) provides phonemic transcriptions for Italian words (smaller than GLAW-IT, it counts 120,000 wordforms).

Conceived as a general-purpose dictionary, GLAW-IT is intended to be directly workable or used to easily derive customized lexicons dedicated to specific uses including NLP, linguistic description and psycholinguistics. Filtering linguistic labels or other markups instantly permits on demand tailoring of sublexicons such as loanwords, regionalisms, dated and domain- specific words, etc. Regarding lexicographic analysis, an immediate application could be the exploitation of GLAW-IT for neologisms detection. We also are currently designing a large morphosyntactic and phonological based grounded on GLAW-IT that hopefully will benefit Italian linguistic studies and NLP applications.

[1] Freely available at: http://redac.univ-tlse2.fr/lexicons/

References

Bel, Núria, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Alessandro Lenci, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas & Antonio Zampolli. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000),* Athens, Greece.

Bertinetto, Pier Marco, Cristina Burani, Alessandro Laudanna, Daniela Ratti Lucia Marconi & Claudia Rolando. 2005. *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. http://linguistica.sns.it/CoLFIS/Home.htm.

Goslin, Jeremy, Claudia Galluzzi & Cristina Romani. 2014. PhonItalia: a phonological lexicon for Italian. *Behavior Research Methods* 46(3): 872–886.

Poletto, Cecilia. 2000. *The Higher Functional Field.* Oxford: Oxford University Press.
Wolfe, Sam. 2015. The nature of Old Spanish Verb Second reconsidered. *Lingua* 165: 132–155.

Lenci, Alessandro, Gabriella Lapesa & Giulia Bonansinga. 2012. LexIt: A Computational Resource on Italian Argument Structure. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2012)*, 3712–3718. Istanbul, Turkey.

Navigli, Roberto & Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'2010)*, 216–225. Uppsala, Sweden.

Roventini, Adriana, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini & Francesca Bertagna. 2000. ItalWordNet: a Large Semantic Database for Italian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2000)*, 783–790. Athens, Greece.

Sajous, Franck & Nabil Hathout. 2015. GLAWI, a free XML-encoded Machine-Readable Dictionary built from the French Wiktionary. In *Proceedings of the of the eLex 2015 conference*, 405–426. Herstmonceux, UK.

Talamo, Luigi, Chiara Celata & Pier Marco Bertinetto. 2016. derIvaTario: An Annotated Lexicon of Italian Derivatives. *Word Structure*: 9(1).

Zanchetta, Eros & Marco Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. *Corpus Linguistics*: 1(1).

```
<article>
  <title>danno</title>
  <pageId>70444</pageId>
  <text>
    <pos type="sost" lemma="1" m="1" s="1">
      <paradigm>
        <inflection gracePOS="Ncmp" form="danni"/>
        <inflection gracePOS="Ncms" form="danno"/>
      </paradigm>
      <definitions>
        <definition>
          <gloss>
            <labels>
              <label type="domain">diritto</label>
              <label type="domain">economia</label>
            </labels>
            <wiki>{{term|diritto|it}}{{term|economia|it}} conseguenza di un'[[azione]] negativa subita</wiki>
            <txt>conseguenza di un'azione negativa subita</txt>
          </gloss>
        </definition>
        <definition>
          <gloss>
            <labels>
              <label type="domain">medicina</label>
            </labels>
            <wiki>{{Term|medicina|it}} ogni fenomeno patologico che modifichi l'organismo
            o l'[[efficienza]] di una parte del corpo</wiki>
            <txt>ogni fenomeno patologico che modifichi l'organismo o l'efficienza di una parte del corpo</txt>
          </gloss>
        </definition>
      </definitions>
    </pos>
    <pos type="verb" lemma="0">
      <inflectionInfos>
        <inflectedForm gracePOS="Vmip3p-" lemma="dare"/>
      </inflectionInfos>
      <definitions>
        <definition>
          <gloss>
            <wiki>terza persona plurale, indicativo presente di [[dare]]</wiki>
            <txt>terza persona plurale, indicativo presente di dare</txt>
          </gloss>
        </definition>
      </definitions>
    </pos>
    <phonology>
      <sill>dàn | no</sill>
      <pron type="IPA">'da:nno</pron>
    </phonology>
    <section type="semRel">
      <item type="synonym">danneggiamento</item>
      <item precisions="medicina" type="synonym">lesione</item>
      <item type="synonym">svantaggio</item>
      <item precisions="medicina" type="synonym">alterazione</item>
      <item type="antonym">sistemazione</item>
      <item type="antonym">risarcimento</item>
    </section>
    <section type="morpho">
      <item type="derivative">dannare</item>
    </section>
  </text>
</article>
```

Figure 1. General structure of the article *danno* in GLAW-IT