derIvaTario: an interactive tool for the study of Italian derivational morphology

Luigi Talamo (University of Bergamo and University of Pavia),
Chiara Celata (Scuola Normale Superiore, Pisa) &
Pier Marco Bertinetto (Scuola Normale Superiore, Pisa)

In the last fifteen years, a number of Natural Language Processing (NLP) tools have been developed to provide Italian corpora with morphological annotation. Although these tools can effectively analyze and generate Italian inflectional paradigms, none of them is able to handle the whole set of Italian derivatives (for an overview, see Tamburini 2013). This is unsurprising: while inflectional paradigms can be computationally formalized on the basis of a finite set of elements (e.g., Pirrelli & Battista 2000 on Italian verbal inflection), combinations between derivational affixes and lexical roots are hardly predictable.

An alternative way to approach the issue consists in the manual annotation of a large list of derivatives, to be further exploited for corpus tagging. *derIvaTario* is a manually annotated lexicon developed at Scuola Normale Superiore, Pisa, consisting of over 10,000 Italian derivatives extracted from the CoLFIS corpus (Bertinetto *et al.* 2005).

The items were morphologically parsed and annotated according to two schemata, respectively describing the nature of the base, with the relevant lexical and morphological information, and the word-formation cycles, with the relevant derivational morpheme and the specific allomorphs implemented.

In addition, two types of transparency values were added, respectively based on a morphotactic and a morphosemantic scale. The former was inspired by the transparency scale proposed by Dressler (1985); e.g., the action noun *unzione* 'unction' is less transparent than the action noun *compilazione* 'compilation', for its stem undegoes a morphophonological change (*unt-ione un[ts]ione*), while in *compilazione* the suffix *-(z)ione* is directly attached to the stem *compilare* 'to compile'. The morphosemantic scale was instead inspired by suggestions in Dressler (2005), dealing with English compounds; e.g., the derivative *spaghetto* 'spaghetti(.SG)' has a lower degree of transparency as compared with *vasetto* 'small jar', since its relation tho the lexical base *spago* 'cord' is purely metaphorical.

derIvaTario can be queried through a graphical interface available at: http://derivatario.sns.it
It can also be downloaded as a text file to be employed as a resource to other NLP tools. The presentation at SLE will illustrate the salient features of *derIvaTario*, with selected examples of possible queries to the database.

References

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., and Thornton, A. M. (2005). *Corpus e Lessico di Frequenza dell'Italiano Scritto*.

Dressler, W. U. (1985). On the Predictiveness of Natural Morphology. *Journal of Linguistics*, **21**(2), 321–337.

Dressler, W. U. (2005). Word-Formation in Natural Morphology. In P. Štichauer and R. Lieber, editors, *Handbook of Word-Formation (Studies in Natural Language and Linguistics Theory, volume 64)*, pages 267–284. Springer.

Pirrelli, V. and Battista, M. (2000). The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics / Rivista di Linguistica*, **12**(2), 307–379.

Tamburini, F. (2013). The AnIta-Lemmatiser: a tool for accurate lemmatisation of Italian texts. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 266–273. Berlin Heidelberg, Springer Verlag.