Modeling Subcategorization through Co-occurrence: a Computational Lexical Resource for Italian Verbs

Gabriella Lapesa, Alessandro Lenci
University of Osnabrück, University of Pisa

1. Goals and Methodology

The aim of this abstract is to introduce *LexIt*,[1] a freely available lexical resource to characterize Italian verb argument properties in terms of distributional information automatically extracted from large corpora with state-of-the-art computational linguistics methods. Research on automatic extraction of subcategorization frames from corpora has a long tradition in computational linguistics, but to the best of our knowledge this is the first large-scale resource of such type for Italian, aiming at characterizing the predicate valence properties fully on distributional ground.

Theoretically grounded on Levin's assumption that distributional data can be used as "a probe into the elements entering into the lexical representations of word meaning" (Levin, 1993: 14) and methodologically based on the Distributional Semantics framework (Miller and Charles, 1991), our approach proposes to model subcategorization and semantic selection properties through co-occurrence data. Co-occurrence turns out to be a powerful instrument for the study of subcategorization, for more than one reason. First of all, co-occurrences can be automatically extracted from large corpora. Moreover, co-occurrences can used to model the association between verbs and syntactic constructions, arguments and semantic classes as a gradient preference instead of categorical selection. Last but not least, the basic notion of collocation as surface co-occurrence can be integrated with more abstract syntactic or semantic information. We used stochastic association measures (Evert, 2008) traditionally applied to the study of word collocations to evaluate the strength of the correlation between: verbs and syntactic frames, argument slots and the words filling them; argument slots and the semantic classes (or polysemies) selected by them. Currently, *LexIt* contains more than 3,900 Italian verbs associated with a syntactic and a semantic profile, automatically extracted from *La Repubblica Corpus* (Baroni *et al*. 2004). The syntactic profile contains the syntactic frames that best characterize the target verb. The semantic profile is further articulated in two subgroups: the prototypical fillers of each argument slot and the semantic classes abstracted over these fillers.

2. Syntactic Profiles

In order to assign syntactic profiles to the target verbs, we implemented the following procedure:

1. we automatically extracted from the parsed corpus[2] the syntactic patterns (subcategorization frames, henceforth SCF) associated to the target verbs. 96 SCF (including up to four slots) were then selected among the 100 most frequent SCF in the corpus;
2. we computed the joint frequency between each verb and the 96 SCF;
3. we estimated the statistical salience of each SCFs with the target verb, by calculating a statistical association measure (Local Mutual Information, LMI). LMI proved to be particularly useful for the identification of the most prototypical SCFs for a verb; moreover, the application of LMI allowed us to discard frames that had been mistakenly associated to the target verb because of parsing errors.

3. Semantic Profiles

Semantic profiles are further articulated into *lexical sets* and *selectional preferences* over semantic classes. The notion of lexical set (Hanks and Pustejovsky, 2005) defines the set of the words that typically occur with a target verb in a given syntactic position. In *LexIt*, the lexical set assigned to each argument slot is composed by the best value (LMI > 0), prototypical examples of the words filling it.

Lexical sets were then used to gain more insight about the selectional preferences of the target verbs over semantic classes. We implemented the following variation of the algorithm described in Schulte-Im Walde

[1] http://sesia.humnet.unipi.it/lexit

[2] The text was dependency-parsed with DeSR, a stochastic dependency parser which constructs dependency trees employing a deterministic bottom-up algorithm, without relying on a subcategorization lexicon. (Attardi and Dell'Orletta, 2009).

(2006):

1. the co-occurrence frequency of each noun as a slot filler in the lexical set associated to an argument slot was divided among the different senses assigned to the noun in the Italian section of MultiWordNet (Pianta *et al.*, 2002).
2. the sense frequency was then propagated up the hierarchy, to 24 mutually exclusive top-nodes, representing broad semantic classes with which verb selectional preferences are represented: ANIMAL, ARTIFACT, ACT, ATTRIBUTE, FOOD, COMMUNICATION, KNOWLEDGE, BODY, EVENT, NATURAL PHENOMENON, SHAPE, GROUP, LOCATION, MOTIVATION, NATURAL OBJECT, PERSON, PLANT, POSSESSION, PROCESS, QUANTITY, FEELING, SUBSTANCE, STATE, TIME. As a result, we obtained the joint frequency between each argument slot and the semantic classes.
3. as an element of novelty with respect to Schulte Im Walde (2006), we calculated the LMI association between each argument slot and the 24 semantic classes.

The following table reports the results for the direct object slot of the verb "leggere" (to read) and "mangiare" (to eat), in the transitive frame (fillers and classes are ordered by decreasing values of LMI):

Verb	*Lexical Set*	*Semantic Classes*
Leggere (to read)	libro (book), giornale (journal), testo (text), articolo (article), lettera (letter), dichiarazione (declaration), racconto (novel), pagina (page).	Communication Artifact Time Substance
Mangiare (to eat)	carne (meat), panino (sandwich), pizza (pizza), pane (bread), pesce (fish), cibo (food), minestra (soup), gelato (ice cream), pasta (pasta)	Food, Substance Plant Natural Object Animal

As a further piece of information to be represented in *LexIt,* we tried to model argument polysemies. Relying on the information concerning selectional preferences over single classes (i.e: the preference of the direct object of the verb "to read" for the types COMMUNICATION and ARTIFACT) we applied LMI to construct corpus-based complex types (Pustejovsky, 1995) eventually associated to frame slots (i.e: the COMMUNICATION-ARTIFACT type for the direct object of "to read"). We achieved interesting results: we will report about this section of *LexIt* (currently still prototypical) in the final presentation.

4. Conclusion

In this abstract, we have briefly sketched the distributional methodology we applied to build *LexIt*, and to show the different levels at which the quantitative analysis of co-occurrences can contribute to a better understanding of argument realization properties. A case-study of lexical analysis based on the *LexIt* data will be presented in the final presentation. Specifically, corpus-driven co-occurrences will be used to identify Levin-style classes of Italian verbs sharing similar argument structure properties.

5. References

Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (NAACL-Short '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 261-264.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI (XML)-compliant corpus of newspaper Italian. In *Proceedings of LREC 2004*, pages 1771-1774, Lisboa, Portugal.

Stefan Evert. 2008. Corpora and collocations. In A. Ludeling and M. Kyto, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Beth Levin. 1993 English Verb Classes and Alternations. A Preliminary Investigation. The University of Chicago Press, Chicago.

George Miller and Walter Charles 1991. Contextual Correlates of Semantic Similarity. Language and Cognitive Processes, 6(1):1–28.

Emanuele Pianta, Luisa Bentivogli, Cristian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In: *Proceedings of the first International Conference on Global WordNet*. Mysore, India.

James Pustejovsky. 1995 The Generative Lexicon. Cambridge, MA: MIT Press.

Sabine Schulte Im Walde. 2006 Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*; 32(2):159-194.