Extracting and Analysing Italian Word Combinations with SYntactically Marked PATterns

Sara Castagnoli[1], **Francesca Masini**[1], **MalvinaNissim**[1,2]
Gianluca E. Lebani[3], **Marco S.G. Senaldi**[3], **Alessandro Lenci**[3]

[1]University of Bologna, [2]University of Groningen, [3]University of Pisa
{s.castagnoli|francesca.masini}@unibo.it, m.nissim@rug.nl, gianluca.lebani@for.unipi.it,
mrcsenaldi@gmail.com, alessandro.lenci@ling.unipi.it

Working Group: WG1

The term Word Combinations (WoCs) broadly refers to the range of combinatory possibilities typically associated with a word. On the one hand, it comprises MWEs: a large variety of recurrent word combinations that act as a single unit at some level of linguistic analysis (Calzolari et al. 2002; Sag et al. 2002; Gries 2008) and that display different degrees of fixedness and idiomaticity. On the other hand, WoCs also include the preferred distributional interactions of a word with other lexical entries at a more abstract level (argument structure patterns, subcategorization frames, and selectional preferences). These two aspects are regarded here not as separate phenomena, but as part of a very intricate continuum that links fixed and flexible combinations, compositional and totally idiomatic ones.

An immediate consequence of this unified view of lexical combinatorics is that the full combinatory potential of a lexical entry can be grasped both at the level of syntactic dependencies and at the more constrained surface level. These two levels, however, are often kept separate in both theory and practice. From a theoretical viewpoint, argument structure is often perceived as a "regular" syntactic affair, whereas MWEs are characterised by "surprising properties not predicted by their component words" (Baldwin & Kim 2010: 267). At the practical level, different extraction methods are used, according to the different types of WoCs/MWEs (Sag et al. 2002; Evert & Krenn 2005): parsers and syntactic dependencies (*S-based methods*) are obviously more suitable to capture more regular combinations, whereas POS-patterns (*P-based methods*) are typically used to extract MWEs (with the aid of Association Measures), although the use of parsers for MWE extraction is becoming more and more widespread (Seretan 2011).

Both methods have pros and cons. Overall, the P-based method – which requires a POS-tagged corpus and a predetermined list of meaningful POS-patterns – yields satisfactory results for relatively fixed, adjacent, and short (2-4 words) MWEs (e.g. Italian *alte sfere* 'high society'). However, some combinations, especially verbal ones, can be very complex and display a high degree of syntactic flexibility (e.g. passivization, dislocation, etc.), which makes it difficult to capture them with POS-patterns only. Syntactic flexibility is well addressed by the S-based method, which is based on dependency relations extracted from parsed corpora. It is therefore possible to extract co-occurrences of words in specific syntactic configurations (e.g. subject#verb, verb#object, etc.) irrespective of their superficial realizations, i.e. generalizing over syntactic flexibility and interrupting material. For this reason, the S-based method is particularly useful to extract "abstract" structures (such as most prototypical objects of a verb, subcategorization frames, etc.), but also more flexible MWEs. However, precisely because S-based methods abstract away from specific constructs and information (word order, morphosyntactic features, interrupting material, etc.), they do not consider how exactly words are combined: they are therefore less suitable to extract fixed MWEs, and can hardly distinguish frequent "regular" combinations (e.g. It. *gettare la sigaretta* 'throw the cigarette') from idiomatic combinations that have the very same syntactic structure (e.g. It. *gettare la spugna* lit. throw the sponge 'throw in the towel').

We argue that, in order to obtain a comprehensive picture of the combinatorial potential of a word, and enhance extracting efficacy for both MWEs and WoCs, the two methods (P-based and S-based) should be combined. The theoretical premises lie in a constructionist view of the language architecture. In Construction Grammar, the basic unit of analysis is the Construction, intended as a conventionalized association of a form and a meaning that can vary in both complexity and schematicity (Fillmore et al. 1988). Therefore, Constructions span from specific structures such as single words to complex, abstract structures such as argument patterns (Goldberg1995), in what is known as the lexicon-syntax continuum, which – in our terms – comprises both MWEs and other types of WoCs.

We implemented this view in a distributional knowledge base – called **SYMPAThy** (Syntactically Marked PATterns) – that is built by: i) extracting from a dependency-parsed corpus all the occurrences of a given lemma; and ii) processing them so as to obtain an integrated representation of the combinatorial information usually targeted (separately) in S-based and P-based methods. The ultimate goal is to filter and interpret the linguistic annotation provided by a pipeline of NLP tools and to represent it with a data format that allows for the simultaneous encoding of the following linguistic information, for any terminal node that depends on a given target lemma TL or on its direct governor (see (1)):

- its lemma;
- its POS tag;
- its morphosyntactic features;
- its linear distance from the TL (and any intervening element);
- the dependency path linking it to TL.

(1) (*La società*) *getta acqua sul fuoco*‘The company pours oil on troubled waters’
[TARGET **gettare-v|s3ip|0#H** [OBJ acqua-s|sf|1#H] [COMP_SU su-ea|sm|2 fuoco-s|sm|3#H]]

The data to be presented in the full poster are extracted from a version of the *la Repubblica* corpus that has been POS tagged with the Part-Of-Speech tagger described in Dell’Orletta (2009) and dependency parsed with DeSR (Attardi and Dell’Orletta 2009). Although SYMPAThy is being developed on Italian data (within the larger project CombiNet: http://combinet.humnet.unipi.it), it can in principle be adapted to other languages.

We intend to exploit this combinatory base to model the gradient of schematicity/productivity and fixedness of combinations, and develop an index of fixedness in order to classify the different types of WoCs on the basis of their distributional behavior.

References

Giuseppe Attardi and Felice Dell’Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In Proceedings of NAACL 2009, pages 261–264.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, Handbook of Natural Language Processing, 2nd Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In Proceedings of LREC 2002, pages 1934–1940.

Felice Dell’Orletta. 2009. Ensemble system for Part-of-Speech tagging. In Proceedings of EVALITA2009.

Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech &Language, 19(4):450–466. Special issue on Multiword Expression.

Charles J. Fillmore, Paul Kay, and Mary Catherine O’Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. Language, 64(3):501–538.

Adele Goldberg. 1995. Constructions. A Construction Grammar Approach to Argument Structures. The University of Chicago Press, Chicago.

Stefan Th. Gries. 2008. Phraseology and linguistic theory: a brief survey. In Sylviane Granger and Fanny Meunier, editors, Phraseology: an interdisciplinary perspective, pages 3–25. John Benjamins, Amsterdam & Philadelphia.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Proceedings of CICLing 2002, pages 1–15.

Seretan, V., 2011. *Syntax-based Collocation Extraction*, Dordrecht, Springer.