Extracting and Analyzing Italian Word Combinations with SYntactically Marked PATterns

Sara Castagnoli[1], Francesca Masini[1], Malvina Nissim[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

PARSEME — WG 1: Lexicon-Grammar Interface

CombiNet

1. What's a WoC

We use the term ***Word Combinations*** (WoCs) to refer to **the range of combinatory possibilities associated with a word**, including:

- MWEs of various kinds (e.g., idioms, phrasal lexemes, collocations, etc.)
- more abstract combinations (e.g., semi-productive patterns, subcategorization frames, selectional preferences, etc.)

WoCs are usually extracted from corpora using either POS-patterns (**P-based** methods) or dependency relations (**S-based** methods).

1.1 P-based extraction

- Strong points
  - good results for relatively fixed | adjacent | short WoCs
- Weak points
  - patterns need to be specified a priori
  - not every extracted string is a WoC, even using AMs
  - some WoCs (especially Vs) can be very complex and flexible: difficult to be captured without syntactic information
  - dismissing more abstract combinatory information (e.g. argument structure)

1.2 S-based extraction

- Strong points
  - extracting syntactically related words irrespective of their superficial realizations, thus handling the complexity and syntactic variability of some WoCs
- Weak points
  - abstracting away from specific constructs and information (e.g., linear order, morphosyntactic features, etc.)
  - cannot distinguish frequent, *regular* combinations from idiomatic ones with the very same syntactic structure

2. Our proposal: A unified approach to WoC extraction

- P-based and S-based methods are in fact highly complementary, as their performance varies according to the different types of combinations that we want to track.
- We propose a new approach to WoC extraction which combines P-based and S-based approaches in order to obtain **a unified and integrated view of a lexeme's *combinatory potential***, i.e. to extract both fixed, lexically specified combinations, such as MWEs of various types, and more abstract, productive aspects of the lexeme's distributional behaviour (such as argument structure patterns, subcategorization frames, and selectional preferences).
  - Our approach is theoretically grounded in a *constructionist* view of the language architecture: Constructions (Cxns) are conventionalized form-meaning pairings that can vary in both complexity and schematicity in what is known as the *lexicon-syntax continuum*, therefore they virtually include all kinds of WoCs.

3. SYMPATHY: SYntactically Marked PATterns

- SYntactically Marked PATterns are obtained from a dependency-parsed corpus by retrieving all the occurrences of a Target Lexeme (TL) and saving into a distributional knowledge base the part of the sentence that is relevant to characterize the combinatorial behavior of TL.
  - Meaningful chunks are formed by all the constituents that govern or are governed by TL, including any intervening elements (e.g., determiners, quantifiers, modifiers).
- This data representation model allows for the *simultaneous* encoding of linguistic and combinatorial information separately targeted in P-based and S-based methods: POS tags, morphosyntactic features, linear order, distance from TL, dependency path linking constituents.
  - For instance, from *La società getta acqua sul fuoco* 'The company pours oil on troubled waters', the following SYMPAThy pattern is extracted:

[TARGET **gettare-v**|**s3ip**|**0#H** [OBJ acqua-s|sf|1#H] [COMP_SU su-ea|sm|2 fuoco-s|sm|3#H]]

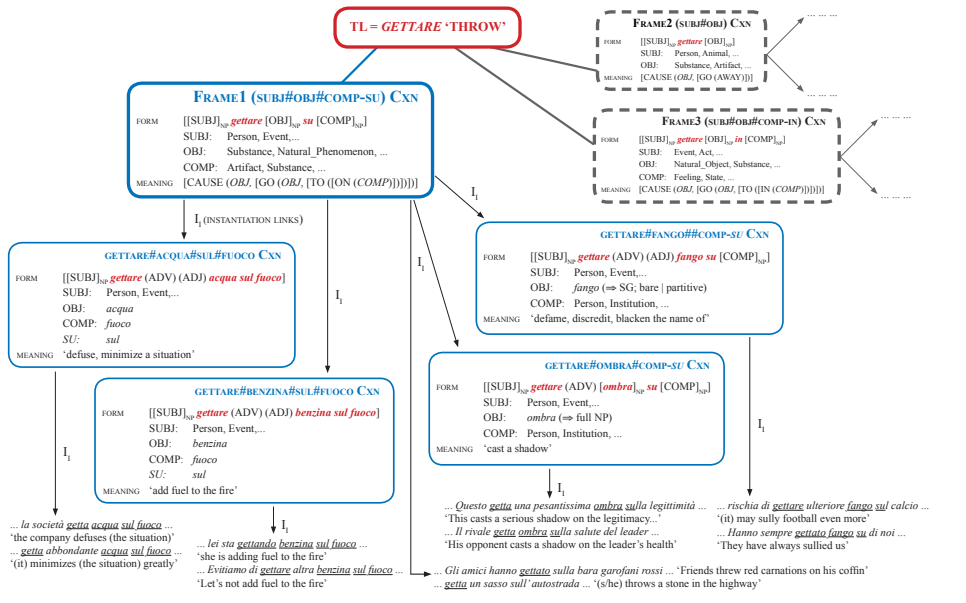
3.1 Method

- In the SYMPAThy model, the combinatory space of a TL is assumed to be formed by a network of Cxns, varying for their degree of fixedness/productivity.
- For any given TL such a representation is built by means of the following procedure:
  - its SYMPAThy patterns are extracted from a corpus;
  - the set of single and multiple slot Cxns (Frames and Fillers) that TL combines with are identified;
  - each Cxn is associated with a *variational profile* (formed by a number of statistics based on SYMPAThy), which is then used to measure their lexical | morphological | syntactic *degrees of freedom*, providing a multidimensional quantitative characterization of their level of fixedness (*work-in-progress*).
- In sum:
  - we first exploit **S-based** information to capture the fact that TL typically occurs with some syntactic Frames, and that for each Frame we have typical Fillers (lexical items) instantiating the Frame slots;
  - then, in order to analyse the degree of fixedness of these Cxns, we turn to **P-based** aspects: presence/absence of intervening materials, morphological variability, surface realization, linear order, etc.

3.2 A sympathetic example

- The verb *gettare* 'throw' combines with a number of schematic Frames/Cxns, among which:
  - subj#obj#comp-su
    - OBJ Filler: {*acqua*, *ombra*, *benzina*, ...}; {Substance, Natural_Phenomenon, ...}
    - COMP-su Filler: {*fuoco*, *tavolo*, *bilancia*, *lastrico*, *istituzione*, ...}; {Artifact, Substance, ...}
  - subj#obj#comp-in
    - OBJ Filler: {*scompiglio*, *sasso*, *corpo*, *fumo*, *cadavere*, ...}; {Natural_Object, Substance, ...}
    - COMP-in Filler: {*panico*, *caos*, *sconforto*, *mare*, *stagno*, *cestino*, ...}; {Feeling, State, ...}
  - subj#obj
    - OBJ Filler: {*spugna*, *base*, *ombra*, *acqua*, *luce*, *ponte*, ...}; {Substance, Artifact, ...}
- Frame1, SUBJ#OBJ#COMP-SU, is schematic and its slots can freely vary with respect to linear order, presence of determiners, modifiers, etc.
- A semi-productive instance of this construction is the SUBJ#OMBRA#COMP-SU Cxn, with a fixed object slot (*ombra* 'shadow') and a partially variable oblique slot, which can appear with a semantically limited range of arguments.
- A fully lexically specified instance of the same Cxn is instead the SUBJ#ACQUA#SUL#FUOCO Cxn, where both slots are fully lexically specified (*acqua* 'water' and *fuoco* 'fire') and show limited degree of variability.

Figure 1



Future work and Acknowledgements

- We aim to exploit this combinatory base to model the gradient of schematicity/productivity and fixedness of combinations, in order to develop a sort of **WoC-hood** indicator and to classify the different types of WoCs on the basis of their distributional behavior (see Lenci et al. Forthcoming).
- Research for this paper was funded by the Italian Ministry of Education, University and Research under the PRIN project **CombiNet**: *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary* (n. 20105B3HE8 003). Partner universities: Roma Tre, Pisa, Bologna. Website: `combinet.humnet.unipi.it/`.

Main references

~ *Baldwin / Kim* 2010. Multiword expressions. In Indurkhya / Damerau (eds), Handbook of Natural Language Processing, CRC Press. ~ *Calzolari et al.* 2002. Towards best practice for multiword expressions in computational lexicons. In Proceedings of LREC 2002, 1934-1940. ~ *Fillmore / Kay / O'Connor* 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. Language 64(3):501-538. ~ *Goldberg* 1995. Constructions. The University of Chicago Press. ~ *Gries* 2008. Phraseology and linguistic theory: a brief survey. In Granger / Meunier (eds), Phraseology: an interdisciplinary perspective, 3-25. Benjamins. ~ *Hanks* 2013. Lexical Analysis: Norms and Exploitations. MIT Press. ~ *Lenci* 2014. Carving verb classes from corpora. In Simone / Masini (eds), Word Classes, 17-36. Benjamins. ~ *Lenci et al.* 2014. SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations. In Proceedings of CLiC-it 2014, Pisa University Press, 234-238. ~ *Lenci et al.* Forthcoming. Mapping the Constructicon with SYMPAThy. ItalianWord Combinations between fixedness and productivity. In Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015. ~ *Nissim / Castagnoli / Masini* 2014. Extracting MWEs from italian corpora: A case study for refining the pos-pattern methodology. In Proceedings of the 10th MWE Workshop, 57-61. ~ *Nissim / Zaninello* 2011. A quantitative study on the morphology of Italian multiword expressions. Lingue e Linguaggio X:283-300. ~ *Ramisch et al.* 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In Proceedings of the LREC Workshop on MWE 2008, 50-53. ~ *Sag et al.* 2002. Multiword expressions: A pain in the neck for NLP. In Proceedings of CICLing 2002, 1-15. ~ *Seretan / Nerima / Wehrli* 2003. Extraction of multi-word collocations using syntactic bigram composition. In Proceedings of RANLP-03, 424-431. ~ *Villavicencio et al.* 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Proceedings of EMNLP-CoNLL 2007, 1034-1043.