10.12871/CLICIT2014145

SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations

Alessandro Lenci[1], Gianluca E. Lebani[1], Sara Castagnoli[2], Francesca Masini[2], Malvina Nissim[3]
[1]University of Pisa, Department of Philology, Literature, and Linguistics
[2]Alma Mater Studiorum, University of Bologna, LILEC
[3]Alma Mater Studiorum, University of Bologna, FICLIT
alessandro.lenci@ling.unipi.it,gianluca.lebani@for.unipi.it,
{s.castagnoli|francesca.masini|malvina.nissim}@unibo.it

Abstract

English. The paper presents SYMPAThy, a new approach to the extraction of Word Combinations. The approach is new in that it combines pattern-based (P-based) and syntax-based (S-based) methods in order to obtain an integrated and unified view of a lexeme's combinatory potential.

Italiano. *L'articolo presenta SYMPAThy, un nuovo metodo per l'estrazione di Combinazioni di Parole. L'originalità dell'approccio consiste nel combinare il metodo basato su sequenze di parti del discorso (P-based) e quello basato sulle dipendenze sintattiche (S-based) per arrivare a una visione integrata e unitaria del potenziale combinatorio di un lessema.*

1 Introduction: Word Combinations

The term Word Combinations (WOCs), as used here, broadly refers to the range of combinatory possibilities typically associated with a word.

On the one hand, it comprises so-called Multiword Expressions (MWEs), intended as a 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): they include phrasal lexemes, idioms, collocations, etc.

On the other hand, WOCs also include the preferred distributional interactions of a word (be it a verb, a noun or an adjective) with other lexical entries at a more abstract level, namely that of argument structure patterns, subcategorization frames, and selectional preferences. Therefore, WOCs include both the *normal* combinations of a word and their idiosyncratic *exploitations* (Hanks, 2013).

The *full combinatory potential* of a lexical entry can therefore be defined and observed at the level of syntactic dependencies and at the more constrained surface level. In both theory and practice, though, these two levels are often kept separate. Theoretically, argument structure is often perceived as a "regular" syntactic affair, whereas MWEs are characterised by "surprising properties not predicted by their component words" (Baldwin and Kim, 2010, 267). At the pratical level, in order to detect potentially different aspects of the combinatorics of a lexeme, different extraction methods are used – i.e. either a surface, pattern-based (**P-based**) method or a deeper, syntax-based (**S-based**) method – as their performance varies according to the different types of WOCs/MWEs (Sag et al., 2002; Evert and Krenn, 2005).

We argue that, in order to obtain a comprehensive picture of the combinatorial potential of a word and enhance extracting efficacy for WOCs, the P-based and S-based approaches should be combined. Thus, we extracted corpus data into a database where both P-based and S-based information is stored together and accessible at the same time. In this contribution we show its advantages. This methodology has been developed on Italian data, within the CombiNet[1] project, aimed at building an online resource for Italian WOCs.

2 Existing extraction methods

The automatic extraction of combinatory information at both the P-level and the S-level is usually carried out in a similar fashion: first, dependency or surface structures are automatically extracted from corpus data, and second, the extracted structures are ranked according to frequency and/or one or more association measures, in order to distinguish meaningful combinations from sequences of words that do not form any kind of relevant unit (Evert and Krenn, 2005; Ramisch et al., 2008; Villavicencio et al., 2007). Let us summarize pros and cons of both methods.

[1]http://combinet.humnet.unipi.it

2.1 P-based approach

P-based methods exploit shallow (POS-)patterns, and are often employed for extracting WOCs. The specification of POS-patterns is a necessary step to obtain a better set of candidate structures with respect to (adjacent) unspecified n-grams. However, despite any attempt to obtain a comprehensive list of language-appropriate patterns (Nissim et al., 2014), not every extracted combination is a WOC, even after association measures are applied. The string may be part of a larger WOC (see *stesso tempo* 'same time', which is a very frequent bigram in itself, but is in fact part of the larger *allo stesso tempo* 'at the same time'), or it may contain a WOC plus some extra element (e.g. *anno/i di crisi economica* 'year(s) of economic crisis', containing *crisi economica* 'economic crisis'). Overall, however, the P-based method yields satisfactory results for relatively fixed, adjacent, and short (2-4 words) WOCs (e.g. *alte sfere* 'high society').

Some WOCs, however, especially verbal ones[2], allow for higher degrees of syntactic flexibility (e.g. passivization, dislocation, variation/addition/dropping of a determiner, internal modification by means of adjectives/adverbs, etc.) (Villavicencio et al., 2007) and/or display a complexity which is difficult to capture without resorting to syntactic information. A collocation like *aprire una discussione* 'start a discussion', for instance, is syntagmatically non-fixed in a number of ways: the determiner can vary (*aprire una/la discussione* 'start a/the discussion'), the object can be modified (*aprire una lunga e difficile discussione* 'start a long and difficult discussion'), and passivization is allowed (*la discussione è stata aperta* 'the discussion was started'). This would require taking into account and specifying all possible variations a priori. Similarly, some idioms can be very difficult to capture with POS-patterns because of their length and complexity, which is hardly "generalizable" into meaningful POS sequences (e.g.: *dare un colpo al cerchio e uno alla botte* lit. give a blow to the ring and one to the barrel 'run with the hare and hunt with the hounds'). Last but not least, P-based approaches are not able to address more abstract combinatory information (e.g. argument structures) and are thus typically limited to MWEs.

[2]In Italian, verbal MWEs are less fixed than nominal ones (Voghera, 2004), even though variability is a thorny issue for nominal MWEs, too (Nissim and Zaninello, 2013).

2.2 S-based approach

S-based methods are based on dependency relations extracted from parsed corpora. They offer the possibility 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. S-based extraction methods thus have two major advantages. First, by moving away from surface forms, they can help account for the complexity and the syntactic variability that some WOCs – like the V+N combination *aprire una discussione* above – might exhibit. Second, by taking into account the dependency between elements, they minimise the risk of extracting unrelated words (Seretan et al., 2003). As a consequence, they are particularly useful to extract "abstract" structures such as lexical sets, i.e. lists of fillers in given slots (e.g. the most prototypical objects of a verb), argument structure patterns and subcategorization frames.

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. Thus, the regular phrase *gettare acqua su un fuoco* 'throw water on a fire' and the structurally similar idiom *gettare acqua sul fuoco* 'defuse' would be treated equally, on the basis of the combination of throw-water-fire.

Also, S-based approaches cannot distinguish frequent "regular" combinations (e.g. *gettare la sigaretta* 'throw the sigarette') from idiomatic combinations that have the very same syntactic structure (e.g. *gettare la spugna* lit. throw the sponge 'throw in the towel'). Statistical association measures alone are not able to discriminate between them as both *sigaretta* and *spugna* are likely to appear among the preferred fillers of the object slot of *gettare*.

3 SYMPAThy: A unified approach

P-based and S-based methods for WOC analysis are in fact highly complementary. In our view, the existing dualism does not reflect the fact that all these combinatory phenomena are interconnected with one another, and that there is a very intricate continuum that links fixed and flexible combinations, compositional and totally idiomatic ones.

In order to represent the full combinatory potential of lexemes, and in an attempt to disentan-

gle this continuum of WOCs, we propose to adopt a unified approach, whose 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; Goldberg, 2006; Hoffmann and Trousdale, 2013). Therefore, Constructions span from specific structures such as single words (Booij, 2010) to complex, abstract structures such as argument patterns (Goldberg, 1995), in what is known as the lexicon-syntax continuum, which comprises MWEs and other types of WOCs.

3.1 SYntactically Marked PATterns

We implemented this view in a distributional knowledge base, SYMPAThy (SYntactically Marked PATterns), built by extracting from a dependency-parsed corpus all the occurrences of a set of lemmas and processing them so as to obtain an integrated representation of the kinds of combinatorial information usually targeted in S-based and P-based methods, albeit separately.

The ultimate goal of our extraction algorithm 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:

- its lemma;
- its POS tag;
- its morphosyntactic features;
- its linear distance from the TL;
- the dependency path linking it to TL.

By building on an automatically annotated corpus, the actual implementation of the SYMPAThy extraction algorithm is largely dependent on the properties of the specific linguistic annotation tools exploited. Here we report examples extracted from a version of the "la Repubblica" corpus (Baroni et al., 2004) 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).

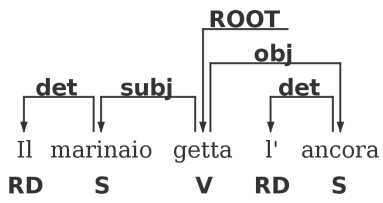
Figure 1 shows the different patterns that can be extracted from the sentence *il marinaio getta l'ancora* 'the sailor throws the anchor', for two different TLs: *gettare* 'throw' and *ancora* 'anchor'. In this representation, the terminal nodes are labeled with patterns of the form `lemma-pos|morphological features|distance_from_target`. For instance, the label `il-r|sm|-2` should be read as an instance of the singular masculine form (`sm`) of the lemma *il* 'the', that is an article (`r`) linearly placed two tokens on the left of TL[3].

The structural information encoded by our patterns, moreover, abstracts from the one-to-one dependency relations identified by the parser in order to build macro-constituents somehow reminiscent of the tree structure typical of phrase structure grammars. Such macro-constituents represent meaningful chunks of linguistic units, in which one element (the 'head', marked by a superscript ^{H}) is prominent with respect to the others. Non-head elements include intervening elements, like determiners, auxiliaries and quantifiers, whose presence is crucial to determine how fixed a linguistic construction is (and that is usually neglected in S-based approaches), and whose linear placement should be known a priori in a P-based perspective. This information is vital in distinguishing idioms, like *gettare acqua sul fuoco* (see Section 2.2), from otherwise identical compositional expressions like *gettare acqua su quel grande fuoco* ('throw water on that big fire').

Finally, the contrast between the two patterns reported in Figure 1 gives a measure of how much the SYMPAThy data representation format is target-dependent. On the one hand, both the syntactic annotation and the linear order are represented with respect to the TL: see the inverse OBJ-1 dependency in the *ancora*-based pattern, as well as the rationale of the indexing encoding the linear positions of terminal elements.

On the other hand, only the part of the sentence that is relevant to characterize the combinatorial behavior of the TL is extracted. In the preliminary work presented here, such a relevant portion includes all the constituents that are directly or indirectly governed by TL (e.g. the object of a verb together with the prepositional phrases modifying its nominal head), and the constituent that governs TL, thus encoding inverse relations like the OBJ-1 dependency in the lower pattern of Figure 1.

[3]For a description of the tagsets used to annotate the corpus, see: `http://medialab.di.unipi.it/wiki/Tanl_Tagsets`.



$[_{TGT}$ $[_{SUBJ}$ il-r|sm|-2 marinaio-s|sm|-1^{H}] **gettare-v|s3ip|$\mathbf{0}^{H}$**
$[_{OBJ}$ il-r|sn|1 ancora-s|sn|2^{H}]]

$[_{OBJ-1}$ gettare-v|s3ip|-2^{H} $[_{TGT}$ il-r|sn|-1 **ancora-s|sn|$\mathbf{0}^{H}$**]]

Figure 1: LEFT: dependency tree[4] for the sentence *il marinaio getta l'ancora* 'the sailor throws the anchor'; RIGHT: SYMPAThy patterns for the TLs *gettare* 'throw' (above) and *ancora* 'anchor' (below).

3.2 A sympathetic example

Here follows a small example showing how such a representation can be used to integrate S-based and P-based approaches. We extracted from our parsed version of the "la Repubblica" corpus all the SYMPAThy patterns featuring a transitive construction governed by the TL *gettare* 'throw'. In a S-based fashion, we ranked the nominal heads filling the object position and found that the most frequent object fillers of *gettare* are *spugna* 'sponge', *acqua* 'water' and *ombra* 'shadow'.

By taking into account the whole subcategorization frame in which these $<TL, obj>$ pairings occur, other interesting patterns emerge. When occurring with *acqua*, TL is often associated with a complement introduced by the preposition *su* and headed by the noun *fuoco* 'fire'. Another salient pattern displays TL with the object *ombra* and an indirect complement introduced by *su* and filled by a nominal head other than *fuoco*.

At the S-level only, it is difficult to guess what the status of these constructions is. Are they compositional or somehow fixed? If the latter, in which way and to what extent is their variation limited? The P-based side of the SYMPAThy data format comes in handy to address such issues. Here crucial pieces of information are the presence/absence of intervening material between TL and the heads of the governed constituents, how variable is the morphological behavior of the relevant lexical elements and to what extent they are free to be superficially realized with respect to TL.

By looking at this information, we can see that the strong association *gettare* + *obj:spugna* is due to the high frequency of the idiomatic expression *TL_la_spugna* 'throw in the towel'. Indeed, 98% of the patterns are linearly and morphologically fixed, with most of the remaining cases (1.7%) being superficial variations due to the presence of interrupting material, typically adverbs.

Cases with *acqua* in the object position present a more articulated picture. Half of them (53.5%) are instances of the rigid idiomatic expression *TL_acqua_sul_fuoco* 'defuse'. As for the remaining cases, even if there is a strong preference for realizing TL and the object one next to the other, with no morphological variation (84%), there is substantial variability in the number, type and filler of the indirect complement (36% of the remaining cases are instances of a subcategorization frame different from the simple transitive one).

When the object slot is filled by *ombra*, finally, the constructions appear to be freer. Even if there is a strong preference (40% of the cases) for the idiom *TL_(una|la)_ombra_su*, roughly meaning 'cast a shadow on', dimensions of variability include the presence/absence of a determiner, its type, and the optional presence of intervening tokens (e.g. adverbs/adjectives) between TL and the object.

Overall this brief example shows how P-based and S-based ideas can be used together to obtain a better description of the combinatoric behavior of lexemes, thus advocating for the usefulness of a resource like SYMPAThy that is able to bridge between the aforementioned approaches.

4 Conclusions

In this paper we presented SYMPAThy, a new method for the extraction of WOCs that exploits a variety of information typical of both P-based and S-based approaches. Although SYMPAThy was developed on Italian data, it can be adapted to other languages. In the future, we intend to exploit this combinatory base to model the gradient of schematicity/productivity and fixedness of combinations, in order to develop an "WOC-hood" indicator to classify the different types of WOCs on the basis of their distributional behavior.

[4]Plotted with DgAnnotator: `http://medialab.di.unipi.it/Project/QA/Parser/DgAnnotator/`

Acknowledgments

This research was carried out within the CombiNet project (PRIN 2010-2011 *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, n. 20105B3HE8), coordinated by Raffaele Simone (Roma Tre University) and funded by the Italian Ministry of Education, University and Research (MIUR).

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, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL.

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.

Geert Booij. 2010. *Construction morphology*. Oxford University Press, Oxford.

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 EVALITA 2009*.

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.

Adele Goldberg. 2006. *Constructions at work*. Oxford University Press, Oxford.

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.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press, Cambridge, MA.

Thomas Hoffmann and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

Malvina Nissim and Andrea Zaninello. 2013. Modeling the internal variability of multiword expressions through a pattern-based method. *ACM Trans. Speech Lang. Process.*, 10(2):1–26.

Malvina Nissim, Sara Castagnoli, and Francesca Masini. 2014. Extracting mwes from italian corpora: A case study for refining the pos-pattern methodology. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 57–61.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the LREC Workshop MWE 2008*, pages 50–53.

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.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of RANLP-03*, pages 424–431.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of EMNLP-CoNLL 2007*, pages 1034–1043.

Miriam Voghera. 2004. Polirematiche. *Linguistica Pragensia*, 67(2):100–108.