Towards a corpus-based online dictionary of Italian Word Combinations

Castagnoli Sara[1], Lebani E. Gianluca[2], Lenci Alessandro[2], Masini Francesca[1], Nissim Malvina[3], Piunno Valentina[4]

[1] University of Bologna, Italy, [2] University of Pisa, Italy
[3] University of Groningen, Netherlands, [4] University of Roma Tre, Italy
E-mail: s.castagnoli@unibo.it, gianluca.lebani@for.unipi.it, alessandro.lenci@unipi.it, francesca.masini@unibo.it, m.nissim@rug.nl, valentina.piunno@uniroma3.it

Keywords: Word Combinations, Italian, dictionary, automatic extraction from corpora

1. Introducing CombiNet

It is widely acknowledged that lexicographers' introspection alone cannot provide comprehensive and accurate information about word meaning and usage, and that investigation of language in use is fundamental for any reliable lexicographic work (Atkins and Rundell 2008:47,53). This is even more true for dictionaries that record the combinatorial behaviour of words, where the lexicographic task is to detect the typical combinations a word participates in. In fact, it was hardly possible to study lexical combinatorics empirically before the advent of large corpora and the definition of statistical techniques for the analysis of word associations (Hanks 2012).

This paper introduces CombiNet, an ongoing national project aimed at studying Italian Word Combinations and at building an online, corpus-based combinatory lexicographic resource for the Italian language[1]. Our working definition of Word Combinations (WoCs) is provided in section 2. Section 3 presents the computational methods and tools we currently use to extract candidate WoCs from corpora, whereas section 4 describes how the automatically acquired information is processed and evaluated by the lexicographers in charge of compiling the dictionary entries. Finally, section 5 introduces current attempts to develop a fully automatic approach to WoC extraction, classification and representation in a combinatory resource.

[1] CombiNet (*Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, http://combinet.humnet.unipi.it/) is a joint project funded by the Italian Ministry of Education, University and Research. Project members: University of Roma Tre (Raffaele Simone, Lunella Mereu, Anna Pompei, Valentina Piunno), University of Pisa (Alessandro Lenci, Gianluca Lebani), University of Bologna (Francesca Masini, Sara Castagnoli, Malvina Nissim).

2. Word Combinations

Following a constructionist approach (Goldberg 2006, Hoffman and Trousdale 2013; Simone 2007; but see also Benson et al.'s definition of *combinatory dictionary*, 2010: vii), we use the term Word Combinations (WoCs) to refer to the whole range of combinatory possibilities typically associated with a word. On the one hand, the term thus encompasses so-called Multiword Expressions (MWEs), i.e. a variety of WoCs – such as phrasal lexemes, collocations and idioms – characterised by different degrees of fixedness and idiomaticity that act as a single unit at some level of linguistic analysis (Calzolari et al. 2002; Sag et al. 2002). On the other hand, we also take WoCs to include the preferred distributional properties of a word at a more abstract level, such as argument structure, subcategorization frames and selectional preferences.

3. WoC extraction/acquisition

Currently, apart from purely statistical approaches, the most common methods for the extraction of WoCs involve searching a corpus via sets of shallow morphosyntactic patterns and then ranking the extracted candidates according to various association measures (Villavicencio et al. 2007, Ramisch et al. 2010).

Our approach to WoCs implies that, in order to define the full combinatory potential of a lexeme, both the more constrained surface level and the level of syntactic dependencies should be considered. Accordingly, different extraction methods are used, i.e. a surface, POS pattern-based (P-based) method and a deeper, syntax-based (S-based) method. Their performance has been suggested to vary according to the different types of WoCs targeted (Sag et al., 2002; Evert and Krenn, 2005): while P-based methods yield satisfactory results for relatively fixed, short and adjacent WoCs, S-based methods should help capture discontinuous and more syntactically flexible WoCs (Seretan 2011).

3.1 Resources and Tools

Candidate WoCs are extracted, using the two above-mentioned methods separately, from a version of the *La Repubblica* corpus (approx. 380M tokens, Baroni et al. 2004) that was POS-tagged using the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency parsed with DeSR (Attardi and Dell'Orletta, 2009).

3.1.1 Resources for P-based extraction

As regards the P-based method, we prepared a comprehensive list of 122 POS sequences deemed representative of Italian WoCs including:

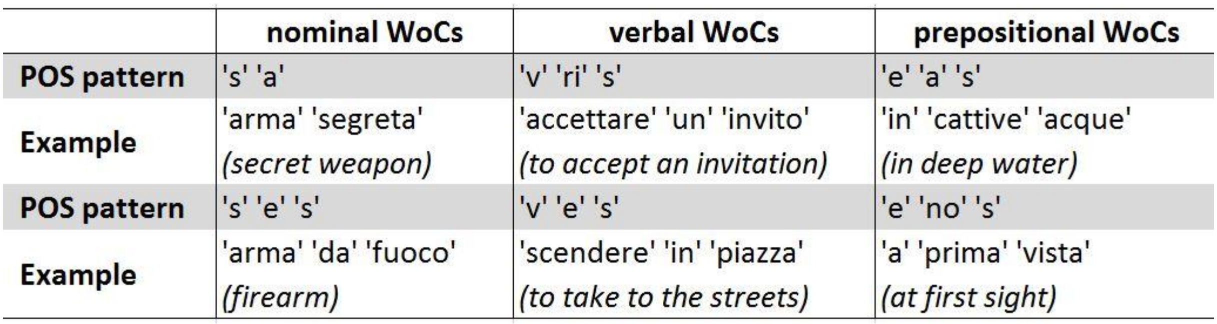
a. patterns mentioned in existing combinatory dictionaries (previously identified in Piunno et al. 2013) and relevant theoretical literature (e.g. Voghera, 2004;

Masini, 2012);

b. “new” patterns identified through corpus-based, statistical experiments (Nissim et al. 2014);

c. patterns added manually by elaborating on the previous lists.

POS patterns are divided in three subsets, broadly representing nominal, verbal and prepositional WoCs respectively (see examples in Table 1). The subsets are used in three independent extraction rounds performed using the *EXTra* tool (Passaro & Lenci, 2015-forthcoming). *EXTra* retrieves all occurrences of the specified patterns (contiguous sequences only, no optional slots can be included) and ranks them according to a variety of association measures, among which we chose Log Likelihood. We also set a minimum frequency threshold of >5.



	nominal WoCs	**verbal WoCs**	**prepositional WoCs**
POS pattern	's' 'a'	'v' 'ri' 's'	'e' 'a' 's'
Example	'arma' 'segreta' *(secret weapon)*	'accettare' 'un' 'invito' *(to accept an invitation)*	'in' 'cattive' 'acque' *(in deep water)*
POS pattern	's' 'e' 's'	'v' 'e' 's'	'e' 'no' 's'
Example	'arma' 'da' 'fuoco' *(firearm)*	'scendere' 'in' 'piazza' *(to take to the streets)*	'a' 'prima' 'vista' *(at first sight)*

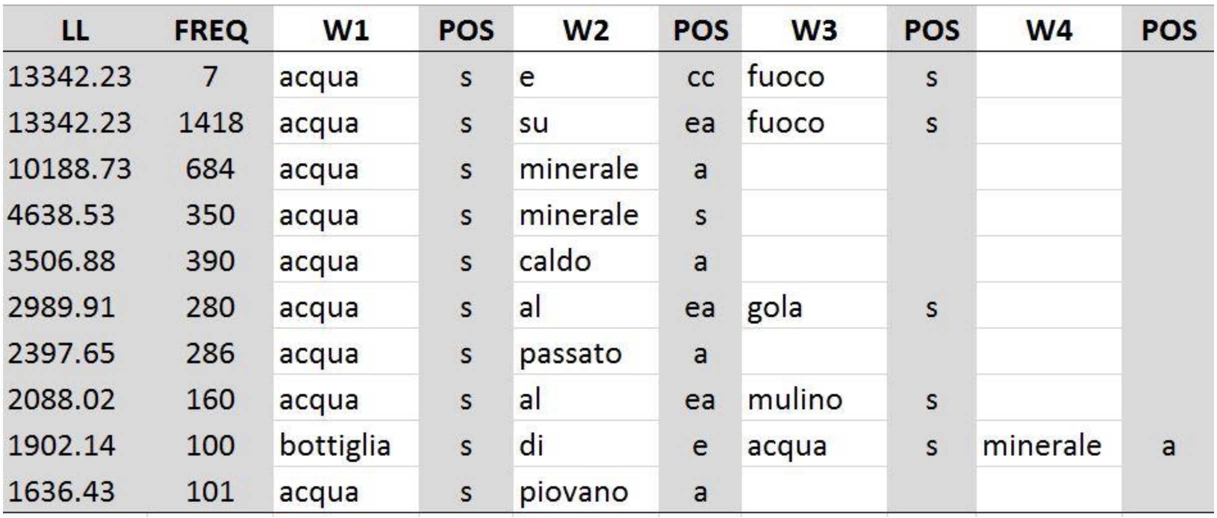
Table 1: Sample POS patterns and corresponding WoCs

3.1.2 Resources for S-based extraction

Information about syntactic dependencies is exploited by the LexIt tool (Lenci 2014), which extracts distributional profiles of Italian nouns, verbs and adjectives from the dependency-parsed corpus. The LexIt distributional profiles contain the syntactic slots (subject, complements, modifiers, etc.) and the combinations of slots (frames) with which words co-occur, abstracted away from their surface morphosyntactic patterns. For instance, *Gianni ha dato volentieri un libro a Maria* and *Gianni ha dato a Maria un libro* (lit. “John has willingly given a book to Mary” “John has given Mary a book”) are both mapped onto the syntactic frame subj#obj#comp_a, despite the different order of their slots and the presence of adverbial modifiers, Moreover, each slot is associated with lexical sets formed by its most prototypical fillers. The statistical salience of each element in the distributional profile is estimated with LL.

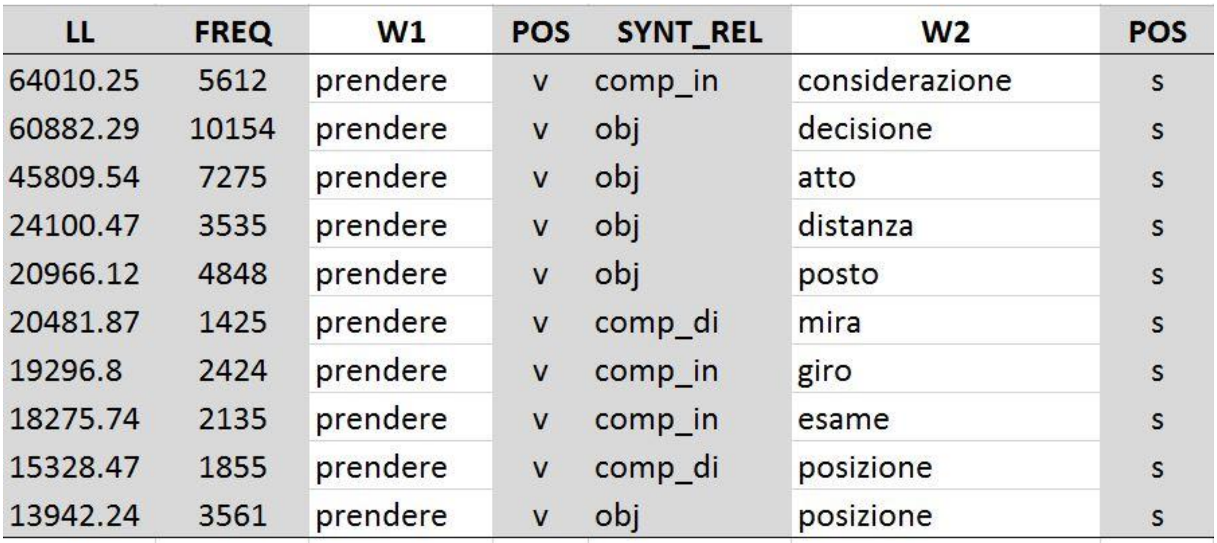
4. Lexicographic processing

In order to provide the lexicographers with manageable sets of data and favour processing, the lists of candidate WoCs obtained as described above are filtered to extract lines containing specific Target Lemmas (TLs)[2], i.e. future dictionary headwords. As shown in Tables 2-3, lexicographers are provided with structured lists in which lemmatised candidate WoCs for a given TL are ranked according to their LL score; information is also provided about the raw frequency of each combination in the corpus, and about the underlying POS pattern or syntactic relation.



LL	FREQ	W1	POS	W2	POS	W3	POS	W4	POS
13342.23	7	acqua	s	e	cc	fuoco	s		
13342.23	1418	acqua	s	su	ea	fuoco	s		
10188.73	684	acqua	s	minerale	a				
4638.53	350	acqua	s	minerale	s				
3506.88	390	acqua	s	caldo	a				
2989.91	280	acqua	s	al	ea	gola	s		
2397.65	286	acqua	s	passato	a				
2088.02	160	acqua	s	al	ea	mulino	s		
1902.14	100	bottiglia	s	di	e	acqua	s	minerale	a
1636.43	101	acqua	s	piovano	a				

Table 2: Top 10 candidates for the TL *acqua* (‘water’) – P-based extraction, nominal patterns



LL	FREQ	W1	POS	SYNT_REL	W2	POS
64010.25	5612	prendere	v	comp_in	considerazione	s
60882.29	10154	prendere	v	obj	decisione	s
45809.54	7275	prendere	v	obj	atto	s
24100.47	3535	prendere	v	obj	distanza	s
20966.12	4848	prendere	v	obj	posto	s
20481.87	1425	prendere	v	comp_di	mira	s
19296.8	2424	prendere	v	comp_in	giro	s
18275.74	2135	prendere	v	comp_in	esame	s
15328.47	1855	prendere	v	comp_di	posizione	s
13942.24	3561	prendere	v	obj	posizione	s

Table 3: Top 10 candidates for the TL *prendere* (‘to take’) – S-based extraction

[2] TLs – which include nouns, verbs and adjectives – are taken from the Senso Comune resource (http://www.sensocomune.it/), which contains 2,010 fundamental lemmas of the Italian lexicon.

As our current lexicographic layout groups WoCs on the basis of their syntactic configuration and function[3], lexicographers can scroll the lists or filter them so as to be able to observe and evaluate only candidate WoCs corresponding to specific POS patterns and/or syntactic relations. Candidates considered as valid WoCs are manually selected and recorded in the relevant part of the lexicographic entry.

The latter records WoCs showing different degrees of lexical specification. On the one hand, it includes fully lexically specified combinations showing a high degree of lexical and syntactic cohesion, e.g. *aprire le danze* (lit. to open the dance, ‘to start something’), *casa di riposo* ('rest home'), *di buona famiglia* ('coming from a good family'). On the other hand it includes sets of examples showing weaker cohesion and internal lexical variation: for instance, NOUN+*dell'anno* ('NOUN+of the year'), where the selection of the NOUN is restricted to specific semantic classes, such as HUMAN (*uomo* 'man') or ARTIFACT (*auto* 'car').

4.1 Evaluation

Although we have not completed any systematic empirical evaluation of the quality of extracted data yet, the study described in Castagnoli et al. (2015-forthcoming) – which was aimed at comparing the performance of the two above-mentioned extraction methods – seems to provide support to mostly impressionistic feedback by our lexicographic team:

- Lexicographers find that P-based data are more useful to compile the entries for nominal and adjectival headwords, whereas S-based data would provide more meaningful insights about verbal headwords. In Castagnoli et al. (ibid.) we calculated the recall of the two systems with respect to a gold standard represented by an existing combinatory dictionary, and found it to be indeed related to the headwords’ POS, thus confirming the lexicographers’ intuition.

- LL ranking is reported to be helpful overall, as most higher-ranking candidates represent (or contain, or suggest) proper WoCs which deserve inclusion in the dictionary. However, lexicographers report finding it difficult to set thresholds, since WoCs which they would intuitively include in the entry also appear in the middle and lower part of the ranking. Preliminary analyses in Castagnoli et al. (ibid.) suggest that recall for the P-based method may plateau at around 2,000 candidates, but need further investigation and refinement.

- Lexicographers report adding WoCs that “should intuitively be there” but are not extracted from the corpus. More research is needed a) to analyse the nature of these WoCs and b) to assess the impact of corpus type and size, as

[3] For instance, for each (sense of a) nominal TL, combinations corresponding to the POS pattern NOUN+ADJ are listed first, followed by combinations of the ADJ+NOUN type, NOUN+PREP+(DET)+NOUN and so on.

well as of extraction techniques and settings.

5. Further developments

Our current approach to WoC extraction follows the tendency to keep P-based and and S-based extraction techniques computationally separate. However, both approaches have limitations: fine-grained differences do not emerge with the S-method, while the P-based method fails to capture the higher-level generalizations one can obtain with the S-method. As a consequence, the lexicographer needs to analyse and evaluate several sets of data for each single lemma.

We believe that, in order to obtain a comprehensive picture of the combinatory potential of a word and enhance extracting efficacy for WoCs, the two approaches should be integrated. For this reason we are developing SYMPAThy (*SYntactically Marked PATterns*), a model of data representation that integrates both surface and deeper linguistic information usually targeted (separately) in S-based and P-based methods. For more details about SYMPAThy, see Lenci et al. 2014 and Lenci et al. 2015. We intend to exploit this combinatory base to model the gradient of schematicity/productivity and fixedness of combinations, and develop an index (or indexes) of fixedness in order to automatically classify the different types of WoCs on the basis of their distributional behaviour.

6. References

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Attardi, G. & Dell'Orletta, F. (2009). Reverse revision and linear tree combination for dependency parsing. *Proceedings of NAACL HLT 2009: Short Papers*, pp. 261–264.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. & Mazzoleni, M. (2004). Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. *Proceedings of LREC 2004*, pp. 1771–1774.

Benson, M., Benson, E. & R. Ilson (2010). *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam/Philadelphia: John Benjamins.

Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C. & Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. *Proceedings of LREC 2002*, pp. 1934–1940.

Castagnoli, S., Lebani, G.E., Lenci, A., Masini, F., Nissim, M. & Passaro, L.C. (2015-forthcoming). POS-patterns or Syntax? Comparing methods for extracting Word Combinations. *Proceedings of EUROPHRAS 2015 - Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*. Malaga, Spain, 29 June - 1 July 2015.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. *Proceedings of*

EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian.

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

Goldberg, A. (2006). *Constructions at work*. Oxford University Press, Oxford.

Hanks, P. (2012). Corpus evidence and Electronic Lexicography. In S. Granger & M. Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. pp. 57-82.

Hoffmann, T. & Trousdale G. eds. (2013). *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.

Lenci, A. (2014). Carving verb classes from corpora. In R. Simone & F. Masini (eds.), *Word Classes. Nature, typology and representations, Current Issues in Linguistic Theory*. Amsterdam/ Philadelphia: John Benjamins, pp. 17–36.

Lenci, A., Lebani, G.E., Senaldi, M.S.G., Castagnoli, S., Masini, F. & Nissim, M. (2015). Mapping the Constructicon with SYMPAThy. Italian Word Combinations between fixedness and productivity. In V. Pirrelli, C. Marzi & M. Ferro (eds.), *Word Structure and Word Usage – Proceedings of the NetWordS Final Conference, Pisa, March 30 - April 1, 2015*, pp. 144-149.

Lenci, A., Lebani, G.E., Castagnoli, S., Masini, F. & Nissim, M. (2014). *SYMPAThy: Towards a comprehensive approach to the extraction of Italian Word Combinations*. In R. Basili, A. Lenci & B. Magnini (eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014, 9-11 December 2014, Pisa. Volume I.* Pisa: Pisa University Press, pp. 234-238.

Masini, F. (2012). *Parole sintagmatiche in italiano*. Roma: Caissa Italia.

Nissim, M., Castagnoli, S. & Masini, F. (2014). Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology. *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014) – EACL 2014, Gothenburg, Sweden, April 26-27, 2014*, pp. 57-61.

Passaro, L.C. & Lenci, A. (2015-forthcoming). Extracting Terms with EXTra. *Proceedings of EUROPHRAS 2015 - Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*. Malaga, Spain, 29 June - 1 July 2015.

Piunno, V., Masini, S. & Castagnoli, S. (2013). *Studio comparativo dei dizionari combinatori dell'italiano e di altre lingue europee.* CombiNet Technical Report. Roma Tre University and University of Bologna.

Ramisch, C., Villavicencio A., Boitet C. (2010) mwetoolkit: a Framework for Multiword Expression Identification. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 662–669.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. *Proceedings of CICLing 2002*, pp. 1–15.

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

Simone, R. (2007). Constructions and categories in verbal and signed languages. In

P. Pietrandrea, R. Simone (eds.), *Verbal and signed languages. Comparing Structures, Constructs and Methodologies*. Berlin: Mouton de Gruyter, pp. 197-252.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M. & Ramisch, C. (2007). Validation and evaluation of automatically acquired multiword expressions for grammar engineering*. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1034–1043.

Voghera, M. (2004). Polirematiche. In M. Grossmann & R. Franz (eds.) *La formazione delle parole in italiano*. Tübingen: Max Niemeyer Verlag, pp. 56-69.