SSL

How to harvest Word Combinations from corpora. Methods, evaluation and perspectives

Alessandro Lenci, Francesca Masini, Malvina Nissim,
Sara Castagnoli, Gianluca E. Lebani,
Lucia C. Passaro, Marco S. G. Senaldi

Abstract

This paper reports on work, carried out in the framework of the CombiNet project, focusing on the automatic extraction of word combinations from large corpora, with a view to represent the full distributional profile of selected lemmas. We describe two extraction methods, based on part-of-speech sequences (P-method) and syntactic patterns (S-method), respectively, evaluating their performance – contrastively, and with reference to external benchmarks – and discussing the relevance of automatic knowledge acquisition for lexicographic purposes. Our results indicate that both approaches provide valuable data and confirm previous claims that P-methods and S-methods are largely complementary, as they tend to retrieve different types of word combinations. In the second part of the paper, we present SYMPAThy, a data representation format devised to fruitfully merge the two methods by leveraging their respective points of strength. In order to explore SYMPAThy's potentialities, a preliminary investigation on a small set of Italian idioms, and specifically their degree of fixedness/productivity, is also described.

Keywords: word combinations, computational methods, idiomatic expressions.

1. *Introduction*

This paper reports on work carried out in the framework of the CombiNet project[1], focusing in particular on the computational task of extracting Word Combinations (WoCs) from corpora to support the creation of an online, corpus-based lexicographic resource for Italian WoCs (cf. Simone and Piunno, 2017). The term Word Combinations is used to encompass both various types of Multiword Expressions (MWEs) – namely WoCs characterised by (different degrees of) fixedness and idiomaticity that act as a single

[1] PRIN Project 2010-2011 *Word Combinations in Italian* (n. 20105B3HE8) funded by the Italian Ministry of Education, University and Research (MIUR). URL: *http://combinet.humnet.unipi.it*.

Received: March 2017
Accepted: July 2017

unit at some level of linguistic analysis, such as idioms, phrasal lexemes, collocations, preferred combinations (cf. Calzolari *et al.*, 2002; Sag *et al.*, 2002; Gries, 2008; Baldwin and Kim, 2010) – and the distributional properties of a word at a more abstract level (i.e., argument structure, subcategorization frames, selectional preferences). Our ultimate goal is to represent the full combinatory profile of selected lemmas using distributional data automatically extracted from very large corpora.

Our approach to WoC extraction is inspired by a constructionist view of language (Fillmore *et al.*, 1988; Goldberg, 2006; Hoffmann and Trousdale, 2013). This model conceives the grammar and the lexicon as a network of Constructions (Cxns), i.e. conventionalized form-meaning correspondences that differ in complexity and schematicity and span from fully specified structures like single words or fixed idioms (e.g., *kick the bucket*) to partially specified Cxns (e.g., *take* Obj *for granted*) and complex, productive abstract structures such as argument patterns (e.g., the Passive Cxn).

This paper describes two WoC extraction methods tested within the CombiNet project, evaluating their performance – contrastively, and with reference to external benchmarks – and discussing the relevance of automatic knowledge acquisition for lexicographic purposes. It also reports on SYMPATHY, a data representation format devised to extract WoCs from corpora which merges the two approaches. A preliminary investigation on a small set of Italian idioms that aimed at exploring SYMPATHY's potentialities is also described.

2. *Two computational methods to 'harvest' Word Combinations from corpora*

Apart from purely statistical approaches, the most common methods currently available for WoC extraction involve searching a corpus with sets of patterns and then ranking the extracted candidates according to various association measures, in order to distinguish relevant combinations from sequences of words that do not form any kind of combinatory unit. The level of linguistic information employed in candidate extraction depends on factors such as the language and the type of WOCs that is targeted. The search is generally performed for either shallow part-of-speech (POS) sequences or syntactic relations: whereas the former (henceforth, the *P-method*) was found to yield satisfactory results for fixed, short and adjacent WoCs

(e.g., *house of cards*), such as multiword terms (e.g., Daille, 2003) and noun compounds (e.g., Vincze *et al.*, 2011), the latter (*S-method*) has proven useful to target discontinuous and syntactically flexible WoCs (e.g., *take* Obj *on board*) (cf., among others, Seretan, 2011; for a more thorough review of the two methods, see Ramisch, 2015: 70-74).

In order to assess which method would provide better data for the CombiNet lexicographic project, both were tested on a lemmatized, POS-tagged and dependency parsed version of *la Repubblica* corpus (Baroni *et al.*, 2004)[2]. P-based and S-based combinatory information for a sample of 25 target lemmas (henceforth TLs)[3] – including high-frequency nouns, verbs and adjectives contained in the *Senso Comune* resource[4] – was extracted from the corpus using the EXTra tool (Passaro and Lenci, 2016) and the LexIt tool (Lenci, 2014) respectively.

The EXTra term extractor takes into account the linguistic structure of multiword terms by implementing a candidate selection step that uses manually-defined structured POS-patterns. Moreover, in order to tackle the complexity of term phrases, EXTra adopts a new association measure that promotes terms composed by one or more sub-terms. The intuition is that the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. The last step of EXTra applies a filtering function to separate real terms from wrong candidates. EXTra includes various parameters that allow users to optimize the extracted terms with respect to the target corpus and domain. In particular, users can specify the set of structured patterns that guide the extraction process, a list of stopwords, the association measure to be used by the weighting algorithm, as well as the thresholds for the association measure and the n-gram frequency.

In the present case, EXTra was fed with a list of 122 fully-specified POS

[2] *La Repubblica* corpus is a collection of newspaper texts from the homonymous Italian daily. While it is arguably not ideal as a reference corpus – being mono-genre and mono-source – compared to others such as *CORIS* (Rossini Favretti *et al.*, 2002), we chose it because it was already parsed and fully available for computational elaboration, besides being more controlled than larger corpora like *itWaC* (Baroni *et al.*, 2009). The version we used was POS-tagged with the tool described in Dell'Orletta (2009) and dependency-parsed with DeSR (Attardi and Dell'Orletta, 2009).

[3] Here follows the list of TLs. Nouns (10): *anno* "year", *governo* "government", *casa* "house", *fine* "end / goal", *guerra* "war", *famiglia* "family", *mano* "hand", *situazione* "situation", *morte* "death", *stagione* "season". Verbs (10): *parlare* "talk / speak", *prendere* "take", *tenere* "keep / hold", *vivere* "live", *perdere* "lose / miss", *uscire* "go out", *lavorare* "work", *costruire* "build", *pagare* "pay", *leggere* "read". Adjectives (5): *economico* "economic", *giovane* "young", *basso* "low / short", *facile* "easy", *rosso* "red".

[4] Cfr. *http://www.sensocomune.it*.

patterns deemed representative of Italian WoCs (up to five slots), which comprises: a) POS sequences mentioned in existing combinatory dictionaries (cf. the survey in Piunno *et al.*, 2013) and relevant theoretical literature (e.g., Voghera, 2004; Masini, 2012); b) sequences identified through corpus-based, statistical experiments (Nissim *et al.*, 2014); c) sequences that were added by elaborating on the previous lists, also as a result of constant interaction with the lexicographic team. The full list of POS patterns used for extraction is given in Appendix 1. In order to ease EXTra's processing, POS patterns were divided into three groups, exemplified in Table 1: round 1 thus contains adjectival and nominal patterns, round 2 verbal patterns, round 3 prepositional patterns. A fourth round was eventually added to include various patterns that emerged incrementally (see point 'c' above).

Adjectival and nominal patterns		Verbal patterns		Prepositional patterns	
N+PREP+N	*amico di famiglia*	V+ART+N	*alzare il gomito*	PREP+A	*a caldo*
A+N	*vecchia volpe*	V+N	*perdere tempo*	PREP+N+A	*a senso unico*
N+A	*piatto forte*	V+PREP+N	*finire in carcere*	PREPART+A+N	*all'ultimo momento*
		V+PREPART+N	*credere sulla parola*		
A+PREP+V	*difficile da credere*	V+A	*restare immobile*	PREPART+N+PREP	*ai fini di*

Table 1. *Examples of POS patterns used for P-based WoC extraction.*

LexIt is a computational framework whose aim is to automatically extract distributional information about the argument structure of predicates. LexIt processes linguistic information from a dependency-parsed corpus and then stores the results into a database where each predicate is associated with a distributional profile, i.e., a data structure that combines several statistical information about the combinatorial behavior of the lemma. This profile is articulated into a *syntactic profile*, specifying the syntactic slots (e.g. subject, complements, modifiers, etc.) and the subcategorization frames associated with the predicate; a *semantic profile*, composed of the *lexical set* of the most typical lexical items that occur in each syntactic slot, and the *semantic classes* characterizing the selectional preferences of the different syntactic slots. The most salient subcategorization frames can be identified directly from cor-

pora in an unsupervised manner, without resorting to *a priori* lists. Besides, there is no formal distinction between arguments and adjuncts: a subcategorization frame is represented as an unordered pattern of syntactic dependencies whose combination is strongly associated to the target predicate. LexIt also abstracts away from surface morphosyntactic patterns and actual word order.

In our case, all the occurrences of TLs in different syntactic frames were extracted together with the lexical fillers of the relevant syntactic slots. The main difference between EXTra and LexIt is that the former targets linear sequences, while the latter can exploit the syntactic annotation of a parsed corpus to identify discontinuous and more schematic Cxns too. In both cases only candidate WoCs with frequency > 5 were considered, and ranked (as lemmas) according to one of the most common statistical measures used to estimate the association strength of MWEs and WoCs in general, namely Log Likelihood (Evert, 2009).

In order to illustrate the different outputs obtained using the P-based and the S-based methods, Table 2 shows some candidates extracted by the two tools for the TL *toccare* "to touch", which correspond to the two idiomatic expressions *toccare con mano* "to experience firsthand" (lit. touch with hand) and *toccare il fondo* "to hit rock bottom" (lit. touch the bottom). As regards the former, P-based results include different combinations of the two lemmas *toccare* "to touch" and *mano* "hand", extracted on the basis of three specific POS-sequences (namely, V+PREP+ART+N, V+PREP+N, V+ART+N)[5]; the idiom *toccare con mano* is more frequent than the others, which are basically literal combinations (e.g., *toccare con la mano* and *toccare di mano* "to touch with the hand", *toccare la mano* "to touch the hand"). On the other hand, only one candidate is extracted which corresponds to the second idiom, *toccare il fondo* (V+ART+N). S-based results are based on the two abstract syntactic relations *verb-comp_con* (complement introduced by the preposition *con* "with") and *verb-object*, and are thus limited to one candidate per combination. We will briefly comment on the limitations of both methods in § 4.1 below.

[5] Parts-Of-Speech are indicated in Tab. 2 by letters enclosed in round brackets: (v) for verb, (e) for preposition, (rd) for definite article, (s) for noun.

	Log Likelihood	Freq	Candidates							
P-based	4553.28	37	toccare	(v)	con	(e)	il	(rd)	mano	(s)
(verbal patterns)	4553.28	14	toccare	(v)	di	(e)	mano	(s)		
	4553.28	739	toccare	(v)	con	(e)	mano	(s)		
	4553.28	15	toccare	(v)	il	(rd)	mano	(s)		
	3793.50	503	toccare	(v)	il	(rd)	fondo	(s)		
S-based	8616.49	892	toccare	(v)	*comp_con*	mano (s)				
	2067.49	484	toccare	(v)	*obj*	fondo (s)				

Table 2. *Selected P-based and S-based candidates for the verbal TL* toccare *"to touch".*

3. *Evaluating P-based and S-based methods*

The performance of the two extraction methods was evaluated contrastively and with reference to external benchmarks by means of three evaluation experiments, described in the following sections.

3.1. *Evaluation against gold standard dictionary*

First, an automated comparison was set up between candidate WoCs extracted for the 25 selected TLs and combinations for the same lemmas included in the most comprehensive combinatory dictionary available for the Italian language, i.e. *Dizionario Combinatorio Italiano* (DiCI, Lo Cascio, 2013). The dataset derived from this printed, manually compiled dictionary was taken as a gold standard to calculate recall, that is the percentage of DiCI combinations successfully retrieved by the two methods.

The results of the experiment (fully described in Castagnoli *et al.*, 2016) indicate that recall varies greatly across lemmas. As regards the P-method, recall ranges between 73-74% for TLs like *rosso* "red" and *economico* "economic" and 37% for lemmas like *tenere* "keep/hold" and *fine* "end/goal". As for the S-method, we get 74% recall for *economico* "economic" and 14% for *mano* "hand". Such considerable differences in recall might be due to the nature of the lemmas involved: the fact that recall for both methods is lower for highly polysemous words like *fine* "end/goal" and *mano* "hand" might indicate that only part of their respective word senses is represented in a

skewed corpus like *la Repubblica* (see § 2). Overall, average recall is higher for the P-method than for the S-method, namely 53% vs. 48%. However, R-precision, which measures precision at the rank position corresponding to the number of combinations found in DiCI, is almost always higher for LexIt than for EXTra. This suggests that S-based data is less noisy, i.e. there are fewer irrelevant results among the highest-ranking candidates extracted with the S-method than is the case with the P-method.

The comparison also reveals that, although the two methods perform similarly for about 76% of gold standard combinations – i.e. they both do or do not retrieve such combinations – they can also be considered as strongly complementary: the P-method was found to have a better performance for nominal and adjectival TLs, whereas the S-method had a higher recall for virtually all verbal TLs. This result was quite expected, given that nominal MWEs are generally more fixed than verbal ones in Italian (Voghera, 2004; cf. also § 2).

Further investigations might be needed to ascertain the extent to which the results are influenced by the specific features and settings of the extraction tools, as well as by the types of combinations and the way they are represented in the gold standard.

3.2. *Human evaluation*

A second stage of evaluation focused on collecting human judgments on 2,000 candidate WoCs extracted from the corpus – 1,000 from each system, taking the top 100 candidates for 10 of the 25 TLs used in the automatic evaluation – in a twofold perspective. On the one hand, a group of linguists was asked to provide expert judgments on the status of such candidates as valid or non-valid WoCs; besides performing a comparative assessment of the two methods, the main aim was to assess the proportion of valid WoCs that are extracted from the corpus but unattested in a manually compiled resource like DiCI, thus providing information to improve dictionary coverage. On the other hand, the candidate dataset was submitted – with detailed instructions – to 'linguistically naïve' evaluators recruited through the Crowdflower platform[6], asking them to rate on a 1-5 scale the *typicality* of candidate combinations (in other words, whether these were proper WoCs deserving inclusion in a combinatory dictionary) as well as their *idiomatic-*

[6] Cfr. *http://www.crowdflower.com*.

ity (i.e., their degree of (non-)compositionality). This second crowdsourcing experiment (fully described in Nissim *et al.*, 2015) was primarily designed to shed light on whether experts' and non-experts' judgements differ in the assessment of WoCs, and to detect potential differences in the degree of idiomaticity of the WoCs the two methods extract. The results of both experiments are summarized in Table 3.

	Expert evaluation		Non-expert evaluation		Both
	Valid candidates	Valid candidates *not* in DiCI	Valid candidates	Valid and idiomatic	Valid candidates
EXTra	40.8% (408/1,000)	64.7% (264/408)	64.1% (641/1,000)	42.9% (275/641)	33.4% (334/1,000)
LexIt	44.7% (447/1,000)	58.4% (261/447)	58.2% (582/1,000)	37.8% (220/582)	33.1% (331/1,000)
EXTra+LexIt	42.8% (855/2,000)	52.6% (525/855 - *of which 75 in common* -> 450/855)	61.2% (1,223/2,000)	40.5% (495/1,223)	33.3% (665/2,000)

Table 3. *Results of human evaluation.*

As regards expert evaluation, positive judgments were expressed for 42.8% of candidate WoCs extracted by the two systems (855/2,000); the S-method was judged to be more precise than the P-method (44.7% vs. 40.8%). Quite interestingly, more than half of WoCs judged as valid by experts were found not to be recorded in DiCI (450/855), with an almost equal contribution from the two extraction methods. Validated WoCs not recorded in DiCI include, for example, *prendere atto* "to acknowledge", *prendere la mira* "to take aim", *prendere il via* "to start", *tappeto rosso* "red carpet", *finale di stagione* "end of (the) season", *famiglia tradizionale* "traditional family", *basso impatto* "low impact" *pagare – retta/parcella/abbonamento/dividendo* "pay – fee/invoice/subscription/dividend", *prendere – rivincita* "to take – revenge", *tenere a battesimo* "to inaugurate". The results thus seem to confirm the potential of corpus-based WoC extraction for lexicographic tasks.

The comparison between expert and non-expert evaluation shows disagreement in the perception of what a valid/typical WoC is. Non-expert evaluators were more inclusive than experts, judging as valid 61.2% vs. 42.8% of candidates; however, only around half of the WoCs judged as valid by non-experts were also considered valid by experts, and there is some evidence that unexperienced judgments may sometimes be inaccurate. For instance, lay evaluators validated candidates like *dichiarare una guerra* "declare a war"

and *tenere l'ostaggio* "take the hostage", which resemble proper WoCs but in fact differ in some details (proper WoCs would be, respectively, *dichiarare guerra* "declare war" and *tenere in ostaggio* "take someone hostage"), as well as dubious WoCs like *famiglia italiana* "Italian family" and *prendere - carta* "take - paper". On the other hand, non-experts also identified a few WoCs that experts failed to validate, such as *prendere corpo* "to take shape", *guerra punica* "punic war", and *prendere a prestito* "to borrow". Moreover, while the S-method had higher precision according to experts and to dictionary comparison, lay evaluators attributed a better performance to the P-method: the reason for this may lie in the fact that LexIt candidates correspond to more abstract and schematic WoCs, which could be harder to map onto specific instances by lay evaluators. Similarly, the higher ratio of candidates annotated as idiomatic for EXTra than for LexIt may indicate that it was easier for evaluators to identify valid idiomatic WoCs when they were given full strings (e.g. *toccare il fondo*) rather than word couples (e.g., *toccare – fondo*).

3.3. *Discussion*

A number of interesting insights emerge from the three evaluation experiments described above. To start with, they provide evidence that relying on the (semi)automatic extraction of word combinations from corpora proves to be a very fruitful methodology: not only do the tested extraction systems have a good recall against the existing combinatory dictionary chosen as benchmark, but – according to expert evaluators – they also make it possible to acquire a large number of previously unregistered WoCs that would be worth adding to the dictionary. Human inspection of raw extraction results remains a necessary step, in order to identify valid WoCs.

Differences between expert and non-expert evaluations suggest that the notion of Word Combination may need to be better defined and refined: experts' judgments may be biased by their own conception of WoCs, which in turn may depend on their field of expertise (e.g., a syntactician will possibly confer to the concept of WoC a different interpretation than a semanticist would), while non-linguists may find the concept itself, not to mention the notion of idiomaticity, too difficult to grasp. However, it cannot be ignored that about one third of candidates (665/2,000, cf. Table 3) were judged valid by all evaluators: this 'core' set, which results from a sort of 'double-checked crowdsourcing' compensating for limitations on either sides, can possibly be taken as the pool of best candidates to populate the lexicographic resource.

Finally, all evaluation experiments suggest that P-based and S-based extraction methods are largely complementary, rather than competing with one another. On the one hand, in the first automatic evaluation experiment, recall appears to be related to the POS of the TL (P-method performs better with nouns/adjectives, S-method performs better with verbs). On the other hand, with respect to human evaluation, out of the total number of valid WoCs extracted by the two systems and not recorded in DiCI (525) only 75 combinations overlap, which suggest that the two systems extract different types of valid WoCs. These findings thus point to the need for hybrid extraction systems that leverage both P-based and S-based information, as recently suggested also by Heid (2015) and Squillante (2015).

4. *SYMPAThy – SYntactically Marked PATterns*

4.1. *Combining P-based and S-based information to extract WoCs from corpora*

To demonstrate how beneficial it is to combine the P-based and the S-based perspectives, we provide an example with the TL *gettare* "to throw". Resorting to the S-method, we can observe that our TL typically occurs within some syntactic frames, that for each frame we have typical fillers (lexical items) instantiating frame slots, and that each slot is associated with certain semantic (ontological) classes (selected LexIt data):

- 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, ...}

However, such a procedure does not allow us to distinguish a frame like subj*#gettare#acqua#su_fuoco* lit "to throw water on the fire" which –

given an appropriate context – may stand for the idiom *gettare acqua sul fuoco* "to defuse", from a literal combination like subj#*gettare#acqua#su_tavolo* "to throw water on the table" and an in-between case such as subj#*gettare#fango#su_istituzione* lit. "to throw mud on an institution", which contains the fixed idiomatic part *gettare fango su_X* "to defame" with a free slot being instantiated by a potentially open class of lexemes belonging to the PERSON/INSTITUTION/EVENT semantic type.

Such a difference can be grasped by a P-method, which searching for the previously identified POS pattern "V+N+PREPART+N" would detect a stronger association between the components of an idiom like *gettare acqua sul fuoco* with respect to the components of a literal combination like *gettare acqua sul tavolo*. The main problem for a P-based approach, however, is that many WoCs, especially verbal ones, allow for a considerable degree of syntactic flexibility (Villavicencio *et al.*, 2007). In an idiom like *gettare acqua sul fuoco*, for instance, the determiner can vary (*gettare (dell)'acqua sul fuoco* lit. "to throw (some) water on the fire"), the object can be modified (*gettare molta acqua sul fuoco* lit. "to throw a lot of water on the fire", meaning "to defuse a lot"), and passivization is allowed (*viene gettata acqua sul fuoco* lit. "water is thrown on the fire"). This would require taking into account and specifying all possible variations a priori. On top of that, P-based approaches are not able to address more abstract combinatory information (e.g., argument structures) and are thus typically limited to MWEs, but not to semi-fixed combinations such as *gettare fango su X*.

In sum, while fine-grained differences between different types of WoCs do not emerge with the S-method, the P-method fails to capture the higher-level generalizations we get with the S-method. For this reason, we devised SYMPATHY (Lenci *et al.*, 2014; 2015), a data representation format that integrates both methods.

4.2. *The representational format of SYMPATHY data*

For every occurrence of a given TL in a dependency parsed and POS-tagged corpus, the SYMPATHY extraction algorithm derives a data format that simultaneously encodes the following linguistic information for every terminal node depending on the TL:

- its lemma;
- its POS tag;

– its morphosyntactic features;
– its linear distance from the TL;
– the dependency path linking it to TL.

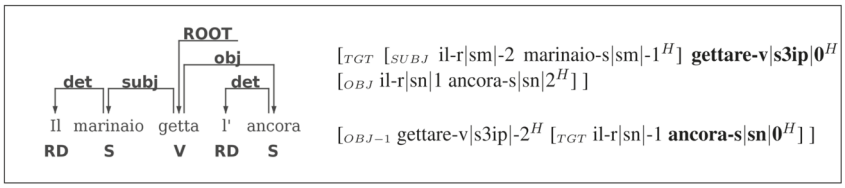

Figure 1. *LEFT: dependency tree for the sentence* il marinaio getta l'ancora *"the sailor throws the anchor"; RIGHT: SYMPAThy patterns for the TLs* gettare *"throw" (above) and* ancora *"anchor" (below).*

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 refers to the singular masculine form (sm) of the lemma *il* "the", that is an article (r) linearly placed two tokens on the left of TL[7]. The difference between the *gettare* and the *ancora* patterns gives an idea of the target-dependent nature of the SYMPAThy format: both syntactic annotation and linear order are represented with respect to the TL (e.g., see the inverse OBJ-1 dependency in the *ancora* pattern). Moreover, only the constituents that are directly or indirectly governed by TL and the constituent that governs it are extracted. Finally, the structural information encoded by our patterns abstracts from the one-to-one dependency relations identified by the parser and builds macro-constituents that somehow remind of the tree structure typical of phrase structure grammars. Such constituents represent meaningful linguistic chunks, 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 (but is usually neglected in S-based approaches), and

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

whose linear placement should be posited a priori in a P-based perspective. This information makes it possible to tell apart an idiom like *gettare acqua sul fuoco* and an otherwise identical compositional expression like *gettare acqua su quel grande fuoco* ("throw water on that big fire").

In the next section we report on an experiment that Lenci *et al.* (2015) conducted on a small set of 23 Italian idiomatic expressions to verify how SYMPAThy could be exploited to study the degree of fixedness/productivity of Italian WoCs.

4.3. *Italian idioms between fixedness and productivity: a test case for SYMPAThy*

Previous corpus studies have assessed the fixedness of MWEs by means of indices of inflection, interruptibility and substitutability (Nissim and Zaninello, 2011; Squillante, 2014) or predicting speakers' judgments of idiomaticity with corpus measures of morphosyntactic variability and compositionality (Wulff, 2009). We exploited the potentialities of SYMPAThy to develop a series of corpus indices that described the fixedness of 23 Italian idiomatic expressions. Our approach was then evaluated by comparing a composition of our indices against the behavioral judgments of syntactic flexibility collected by Tabossi *et al.* (2011), in order to test if our indices correlate with the intuitive judgments of native speakers about the fixedness of fully lexically specified constructions.

4.3.1. *Measuring WoC fixedness with Shannon's Entropy*

Shannon's (1948) Entropy was used to compute how flexible each idiom in our dataset was in a series of dimensions of formal variability. Entropy is a measure of randomness that computes the average degree of uncertainty in a random variable *X*:

$$H = -\sum_{x \in X} p(x) log_2 \left(p(x)\right)$$

In the above formula, the variable *X* stands for a given idiom of interest, while each state of the system *x* represents any possible value the idiom can assume in a certain variational dimension. Lower entropy values are to be understood as evidence of fixedness, while higher values suggest a higher variability of the Cxn in the variational axis at hand. Observed entropy values, however, can span from 0 to the logarithm of the number of values that

X can assume. As a consequence, our entropy values related to different dimensions of variation were not comparable, and could not be combined into a single fixedness index. We overcame this limitation by following Wulff (2008) and resorting to relative entropy, computed as the ratio between the observed entropy from the above equation and the maximum entropy H_{max} for the variable X:

$$H_{rel}(X) = \frac{H(X)}{H_{\max(X)}} = \frac{H(X)}{\log_2(|X|)}$$

We hereby obtained an entropic index that spanned from 0 to 1 for each of the following dimensions of variation:

MORPHOLOGICAL VARIABILITY. The variability of the morphological features manifested by the fillers of a Cxn. Let's take the idiom *tirare la cinghia* "to tighten one's belt" as an example. Out of its 157 occurrences in the corpus, it occurs 156 times with the argument *obj:cinghia* being feminine singular and just 1 time with the noun being feminine plural. The probability of the first state will therefore be about 0.99, while the probability of the second state will be about 0.01. Applying the entropy formula above, we'll have:

$$H_{morph}(tirare\ la\ cinghia) = \frac{-(0.99 * \log_2 0.99 + 0.01 * \log_2 0.01)}{\log_2 2} = 0.08$$

Actually, in computing our indices we did not take into account states that occurred just once (like *cinghia* appearing as feminine plural), because we did not consider them to be informative. In the case above, for instance, the actual morphological variability value was eventually set to 0, indicating that *tirare la cinghia* hardly ever varies the morphology of its noun argument.

ARTICLES VARIABILITY. The variability in the presence or absence of articles and, if appropriate, their type (definite vs. indefinite). *Tirare la cinghia*, for instance, occurs 1 time with no adjectives, 1 time with an indefinite article (e.g., *tirare una cinghia* "to tighten one belt") and 155 times with a definite article, so the resulting article variability entropy is 0.02.

PRESENCE OF MODIFIERS. The variability in the presence or absence of intervening adjectives and PPs. *Tirare la cinghia* occurs 13 times with modifying adjectives or PPs (e.g. *tirare la cinghia dei tassi d'interesse*

"to tighten the belt of interest rates") and 144 times without modifiers, with an overall modifier entropy of 0.41.

DISTANCE VARIABILITY. The variability in the token distance of the Cxns constituents from the verbal TL. In our example case, *cinghia* occurs 7 times at a 4-token distance from *tirare*, 10 times at a 3-token distance from *tirare* and the remaining 136 times at a 2-token distance from the verb. The distance variability entropy in this case is equal to 0.40.

We experimented four ways to combine the entropic indices above in an all-embracing flexibility index *F(X)* for each idiom, namely SUM, AVERAGE, average of the positive values ($\text{AVERAGE}_{\text{POS}}$) and finally considering just the highest value (MAX).

4.3.2. *The descriptive norms by Tabossi* et al. *(2011)*

Tabossi *et al.* (2011) collected human ratings on 245 Italian verbal idioms from 740 subjects on a series of psycholinguistically relevant variables, including syntactic flexibility. Each variable was evaluated by a minimum of 40 speakers. To collect syntactic flexibility judgments, each idiomatic expression was put into a sentence in which one of the following five syntactic modifications occurred: adverb insertion, adjective insertion, left dislocation, passive and movement. Participants were asked to evaluate on a 1-7 scale how much the meaning of the idiomatic expression in the syntactically modified sentence was similar to its unmarked meaning as expressed in a paraphrase prepared by the authors.

4.3.3. *Idiom extraction and analysis*

Out the 245 expressions in Tabossi *et al.* (2011), we selected the 23 target idioms reported in Appendix 2 and proceeded this way:

1. for each verbal TL of each idiom, we extracted its SYMPAThy patterns from the *la Repubblica* corpus;
2. the patterns involving one of our target idioms were identified and selected;
3. for each idiom, the variability indices described in § 4.3.1 were calculated;
4. we built a fixedness index for each idiom, according to the four composition methods presented in § 4.3.1.

4.3.4. *Results and discussion*

In Table 4 we report the individual (MORPHOLOGICAL, ARTICLES, MODIFIERS, DISTANCE) and aggregated (SUM, AVERAGE, AVERAGE$_{POS}$ and MAX) entropic scores for the 5 most and 5 least flexible idioms with respect to the AVERAGE fixedness score.

Idiom	Morphological Entropy	Articles Entropy	Modifiers Entropy	Distance Entropy	SUM	AVG	AVG$_{POS}$	MAX
Mettere il dito sulla piaga	0.45	0.61	0.73	0.54	2.33	0.58	0.58	0.73
Prendere una cotta	0.30	0.37	0.90	0.63	2.20	0.55	0.55	0.90
Prendere un granchio	0.30	0.44	0.77	0.61	2.11	0.53	0.53	0.77
Perdere il treno	0.29	0.46	0.65	0.40	1.79	0.45	0.45	0.65
Mettere i puntini sulle i	0.22	0.22	0.67	0.37	1.48	0.37	0.37	0.67
Gettare la spugna	0.00	0.02	0.20	0.18	0.40	0.10	0.13	0.20
Tirare le cuoia	0.00	0.00	0.00	0.25	0.25	0.06	0.25	0.25
Tirare i remi in barca	0.00	0.00	0.00	0.13	0.13	0.03	0.13	0.13
Mettere il carro davanti ai buoi	0.00	0.13	0.00	0.00	0.13	0.03	0.13	0.13
Mettere nero su bianco	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4. *Individual and aggregated entropic flexibility indices for the 5 most and 5 least flexible idioms according to the* AVERAGE *overall flexibility score.*

The idiom that obtained the lowest overall flexibility value with the SUM, the AVERAGE and the MAX composition methods was *mettere nero su bianco* ("to put into writing", lit. "to put black on white") with a score of 0.0 that meant a complete lack of flexibility along the variational axes explored, while the most fixed idiom according to the AVERAGE$_{POS}$ composition method was *mettere il carro davanti ai buoi* ("to put the cart before the horse", lit. "to put the cart before the oxen") with a score of 0.13. By contrast, the SUM, the AVERAGE and the AVERAGE$_{POS}$ methods showed *mettere il dito sulla piaga* ("to touch a sore point", lit. "to put the finger on the sore") to be the most flexible idiom, with three scores of 2.33, 0.58 and 0.58, while the most variable one with respect to the MAX index was *prendere una cotta*

("to get a crush", lit. "to get a cooking"), with an overall score of 0.90. To validate the psycholinguistic plausibility of these formal flexibility indices we obtained with SYMPATHY, we calculated the Pearson's Product-Moment Correlation strength between them and the syntactic flexibility ratings in Tabossi *et al.* (2011). In all cases we found a significant ($p < 0.05$) positive correlation, ranging between 0.44 and 0.47 (Fig. 2). Albeit preliminary, these results look promising given the different nature of the behavioral and corpus-based indices and suggest the psycholinguistic plausibility of our SYMPATHY-based entropic values. We must keep in mind that the speakers' ratings are semantically driven, since they point towards the preservation of idiomatic meaning in the syntactically modified forms, while our entropic indices are not. In addition, they refer to strings that can in principle have an idiomatic as well as a compositional, literal meaning (even if, presumably, the latter case was rare in the corpus).

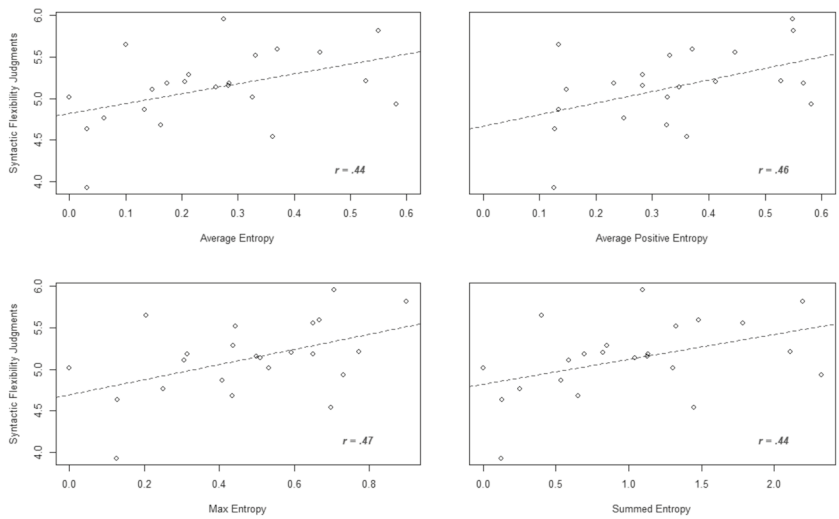

Figure 2. *Pearson's Correlation strength between different combination methods of the SYMPATHY-based fixedness indices and the syntactic flexibility judgments in Tabossi* et al. *(2011). All reported values are associated with $p < .05$, $N = 23$.*

5. *Conclusions*

The goal of this paper was to describe and compare the performance of two methods for the (semi)automatic extraction of Word Combinations

from corpora – based on part-of-speech sequences (P-method) and syntactic patterns (S-method), respectively – with a view to evaluate their usefulness for lexicographic applications, and in particular for the development of the CombiNet dictionary. Our results indicate that both approaches provide valuable data for populating a lexicographic resource. Moreover, they confirm previous claims that P-methods and S-methods are largely complementary, as they tend to retrieve – in addition to a common WoC set – different types of Word Combinations.

In order to leverage the pros of both methods, we devised a new data representation format – called SYMPAThy – that merges features of the two approaches and can thus be exploited to extract a larger variety of Word Combinations within a single environment. In addition, SYMPAThy can represent a useful resource to study the degree of fixedness/productivity of Italian WoCs, and to collocate them along an idiomaticity continuum. Further research is underway to assess the potentialities of SYMPAThy for lexicography as well as for research on lexical combinatorics.

Bibliography

Attardi, G. and Dell'Orletta, F. (2009), *Reverse revision and linear tree combination for dependency parsing*, in Ostendorf, M., Collins, M., Narayanan, S., Oard, D.W. and Vanderwende, L. (2009, eds.), *Proceedings of Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Companion Volume: Short Papers*, Association for Computational Linguistics, Boulder, pp. 261-264.

Baldwin, T. and Kim, S. N. (2010), *Multiword expressions*, in Indurkhya, N. and Damerau, F. J. (2010, eds.), *Handbook of Natural Language Processing*, CRC Press, Taylor and Francis Group, Boca Raton (FL), pp. 267-292.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. and Mazzoleni, M. (2004), *Introducing the* La Repubblica *Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian*, in Lino, M.T., Xavier, M.F., Ferreira, F., Costa, R. and Silva, R. (2004, eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2004)*, European Language Resources Association, Lisbon, pp. 1771-1774.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009), *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*, in «Language Resources and Evaluation», 43, 3, pp. 209-226.

Calzolari, N., Fillmore, C.J., Grishman, R., Ide, N., Lenci, A., MacLeod, C. and Zampolli, A. (2002), *Towards best practice for multiword expressions in computational lexicons*, in Rodríguez, M.G. and Araujo, C.P.S. (2002, eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, European Language Resources Association, Las Palmas, pp. 1934-1940.

Castagnoli, S., Lebani, G.E., Lenci, A., Masini, F., Nissim, M. and Passaro, L.C. (2016), *POS-patterns or Syntax? Comparing Methods for Extracting Word Combinations*, in Corpas Pastor, G. (2016, ed.), *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Editions Tradulex, Geneva, pp. 116-128.

Daille, B. (2003), *Conceptual structuring through term variations*, in Bond, F., Korhonen, A., McCarthy, D. and Villavicencio, A. (2003, eds.) *Proceedings of the ACL workshop on multiword expressions: analysis, acquisition and treatment (MWE 2003)*, Association for Computational Linguistics, Sapporo, pp. 9-16.

Dell'Orletta, F. (2009), *Ensemble system for Part-of-Speech tagging*, in Magnini, B. and Cappelli, A. (2009, eds.), *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, pp. 1-8.

Evert, S. (2009), *Corpora and collocations*, in Lüdeling, A. and Kytö, M. (2009, eds.), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, pp. 1212-1248.

Fillmore, C.J., Kay, P. and O'Connor, M.C. (1988), *Regularity and idiomaticity in grammatical constructions: the case of let alone*, in «Language», 64, 3, pp. 501-538.

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

Gries, S. Th. (2008), *Phraseology and linguistic theory: a brief survey*, in Granger, S. and Meunier, F. (2008, eds.), *Phraseology: an interdisciplinary perspective*, John Benjamins, Amsterdam/Philadelphia, pp. 3-25.

Heid, U. (2015), *Extracting linguistic knowledge about collocations from corpora*, plenary talk delivered at the EUROPHRAS 2015 conference, Malaga, Spain, June 29 – July 1, 2015.

Hoffmann, Th. and Trousdale, G. (2013, eds.), *The Oxford Handbook of Construction Grammar*, Oxford University Press, Oxford.

Lenci, A. (2014), *Carving verb classes from corpora*, in Simone, R. and Masini, F. (2014, eds.), *Word Classes. Nature, typology and representations*, John Benjamins, Amsterdam/Philadelphia, pp. 17-36.

Lenci, A., Lebani, G.E., Castagnoli, S., Masini, F. and Nissim, M. (2014), *SYMPATHy: Towards a comprehensive approach to the extraction of Italian Word Combinations*, in Basili, R., Lenci, A. and Magnini, B. (2014, eds.), *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it 2014)*, Pisa University Press, Pisa, pp. 234-238.

Lenci, A., Lebani, G.E., Senaldi, M.S.G., Castagnoli, S., Masini, F. and Nissim, M. (2015), *Mapping the Constructicon with SYMPATHy. Italian Word Combinations between fixedness and productivity*, in Pirrelli, V., Marzi, C. and Ferro, M. (2015, eds.), *Proceedings of the NetWordS final conference on Word Knowledge and Word Usage - Representations and Processes in the Mental Lexicon*, CEUR-WS.org., Pisa, pp. 144-149.

Lo Cascio, V. (2013), *Dizionario combinatorio italiano*, John Benjamins, Amsterdam/Philadelphia.

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

Nissim, M. and Zaninello, A. (2011), *A quantitative study on the morphology of Italian multiword expressions*, in «Lingue e Linguaggio», 10, pp. 283-300.

Nissim, M., Castagnoli, S. and Masini, F. (2014), *Extracting MWEs from Italian corpora: A case study for refining the POS-pattern methodology*, in Kordoni, V., Savary, A., Egg, M., Wehrli, E. and Evert, S. (2014, eds.), *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, Association for Computational Linguistics, Gothenburg, pp. 57-61.

Nissim, M., Castagnoli, S., Masini, F., Lebani, G.E., Passaro, L.C. and Lenci, A. (2015), *Automatic extraction of Word Combinations from corpora: evaluating methods and benchmarks*, in Bosco, C., Tonelli, S. and Zanzotto, F.M. (2015, eds.), *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*, Accademia University Press, Torino, pp. 204-209.

Passaro, L.C. and Lenci, A. (2015), *Extracting Terms with EXTra*, in Corpas Pastor, G. (2015, ed.), *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Editions Tradulex, Geneva, pp. 188-196.

Piunno, V., Masini, F. and 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. (2015), *Multiword Expressions Acquisition – A Generic and Open Framework*, Springer, Dordrecht.

Rossini Favretti, R., Tamburini, F. and De Santis, C. (2002), *A corpus of written Italian: a defined and a dynamic model*, in Wilson, A., Rayson, P.

and McEnery, T. (2002, eds.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa, Munich, pp. 27-38.

Sag, I.A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2002), *Multiword expressions: A pain in the neck for NLP*, in Gelbukh, A.F. (2002, ed.), *Computational Linguistics and Intelligent Text Processing, Third International Conference (CICLing 2002)*, Springer, Mexico City, pp. 1-15.

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

Shannon, C.E. (1948), *A mathematical theory of communication*, in «The Bell System Technical Journal», 27, 3, pp. 379-423.

Simone, R. and Piunno, V. (2017), *Combinazioni di parole che costituiscono entrata. Rappresentazione lessicografica e aspetti lessicologici*, in«Studi e Saggi Linguistici», 55, 2, pp. 13-44.

Squillante, L. (2014), *Towards an empirical subcategorization of multiword expressions*, in Kordoni, V., Savary, A., Egg, M., Wehrli, E. and Evert, S. (2014, eds.), *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, Association for Computational Linguistics, Gothenburg, pp. 77-81.

Squillante, L. (2015), *Polirematiche e collocazioni dell'italiano. Uno studio linguistico e computazionale*, Ph.D. dissertation Università di Roma "La Sapienza".

Tabossi, T., Arduino, L. and Fanari, R. (2011), *Descriptive norms for 245 Italian idiomatic expressions*, in «Behavior Research Methods», 43, pp. 110-123.

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

Vincze, V., Nagy, T.I. and Berend, G. (2011), *Detecting noun compounds and light verb constructions: a contrastive study*, in Kordoni, V., Ramisch, C. and Villavicencio, A. (2011, eds.), *Proceedings of the ACL workshop on multiword expressions: from parsing and generation to the real world (MWE 2011)*, Association for Computational Linguistics, Portland, pp. 116-121.

Voghera, M. (2004), *Polirematiche*, in Grossmann, M. and Rainer, F. (2004, eds.), *La formazione delle parole in italiano*, Niemeyer, Tübingen, pp. 56-69.

Wulff, S. (2008), *Rethinking Idiomaticity: A Usage-based Approach*, Continuum, London.

Wulff, S. (2009), *Converging evidence from corpus and experimental data to capture idiomaticity*, in «Corpus Linguistics and Linguistic Theory», 5, 1, pp. 131-159.

Appendix 1

Adjectival and Nominal patterns (38)	Verbal patterns (28)	Prepositional patterns (36)	Miscellaneous patterns (20)
['a', 'a'] ['a', 'cc', 'a'] ['a', 'e', 's'] ['a', 'e', 'v'] ['a', 'ea', 's'] ['a', 's'] ['a', 'v'] ['b', 'a'] ['no', 's'] ['s', 'a'] ['s', 'cc', 's'] ['s', 'e', 'a'] ['s', 'e', 'rd', 's'] ['s', 'e', 'rd', ['a', 's']] ['s', 'e', 'rd', ['no', 's']] ['s', 'e', 'rd', ['s', 'a']] ['s', 'e', 'ri', 's'] ['s', 'e', 'ri', ['a', 's']] ['s', 'e', 'ri', ['no', 's']] ['s', 'e', 'ri', ['s', 'a']] ['s', 'e', 's'] ['s', 'e', 'v'] ['s', 'e', ['a', 's']] ['s', 'e', ['s', 'a']] ['s', 'ea', 'a'] ['s', 'ea', 's'] ['s', 'ea', ['a', 's']] ['s', 'ea', ['s', 'a']] ['s', 'ea', ['no', 's']] ['s', 's'] ['s', 'v'] ['s', ['s', 'a']] ['s', ['s', 'cc', 's']] [['s', 'a'], 'ea', 's'] [['s', 'a'], 'a'] [['s', 'a'], 'e', 's'] [['s', 's'], 'a'] [['s', 's'], 'cc', 's']	['v', 'a'] ['v', 'b'] ['v', 'cs', 'rd', 's'] ['v', 'cs', 'ri', 's'] ['v', 'e', 'a'] ['v', 'e', 'rd', 's'] ['v', 'e', 'ri', 's'] ['v', 'e', 's'] ['v', 'e', 'v'] ['v', 'e', ['a', 's']] ['v', 'e', ['n', 's']] ['v', 'e', ['no', 's']] ['v', 'e', ['s', 'a']] ['v', 'ea', 's'] ['v', 'ea', ['a', 's']] ['v', 'ea', ['s', 'a']] ['v', 'rd', 'a'] ['v', 'rd', 's'] ['v', 'rd', ['a', 's']] ['v', 'rd', ['s', 'a']] ['v', 'ri', 'a'] ['v', 'ri', 's'] ['v', 'ri', ['a', 's']] ['v', 'ri', ['s', 'a']] ['v', 's'] ['v', ['n', 's']] ['v', ['s', 'e', 's']] [['v', 's'], 'e', 's']	['e', 'a'] ['e', 'rd', 'a'] ['e', 'rd', 's'] ['e', 'rd', ['a', 's']] ['e', 'rd', ['no', 's']] ['e', 'rd', ['s', 'a']] ['e', 'ri', 'a'] ['e', 'ri', 's'] ['e', 'ri', ['a', 's']] ['e', 'ri', ['no', 's']] ['e', 'ri', ['s', 'a']] ['e', 's'] ['e', ['a', 'cc', 'a']] ['e', ['a', 's']] ['e', ['a', 'v']] ['e', ['b', 's']] ['e', 'di', 's'] ['e', ['no', 's']] ['e', ['s', 'a']] ['e', ['s', 'cc', 's']] ['e', ['s', ['e', 'a']]] ['e', ['s', ['e', 's']]] ['e', ['s', ['ea', 's']]] ['ea', 'a'] ['ea', 's'] ['ea', ['a', 's']] ['ea', ['a', 'v']] ['ea', ['b', 'a']] ['ea', ['b', 's']] ['ea', ['no', 's']] ['ea', ['s', ['e', 'a']]] ['ea', ['s', ['e', 's']]] ['ea', ['s', ['ea', 's']]] ['ea', ['s', 'a']] [['e', 's'], 'e'] [['ea', 's'], 'e']	['di', 's'] ['n', 's'] ['no', 's'] ['e', ['n', 's']] ['e', ['no', 's']] ['e', 'rd', ['no', 's']] ['ea', ['no', 's']] ['ap', 's'] ['s', 'ap'] ['e', ['ap', 's']] ['e', ['s', 'ap']] ['ea', ['ap', 's']] ['ea', ['s', 'ap']] ['cs', 's'] ['cs', 'a'] ['v', 'cc', 'v'] ['bn', ['v', 'rd', 's']] ['bn', ['v', 'ri', 's']] ['bn', ['v', 's']] ['bn', ['v', 'b']]

a = adjective; ap; possessive adjective; b = adverb; cc = coordinating conjunction; cs = subordinating conjunction; e = preposition; ea = articulated preposition; n = cardinal number; no = ordinal number; ri = indefinite article; rd = definite article; s = noun; v = verb.

Appendix 2

Gettare la maschera ("to reveal oneself")
Gettare la spugna ("to give up")
Gettare acqua sul fuoco ("to defuse a situation")
Gettare olio sul fuoco ("to inflame a situation")
Mettere la mano sul fuoco ("to stake one's life on sth")
Mettere il carro davanti ai buoi ("to put the cart before the horse")
Mettere le carte in tavola ("to lay one's cards on the table")
Mettersi il cuore in pace ("to resign oneself to sth")
Mettere nero su bianco ("to put sth down in black and white")
Mettere il dito sulla piaga ("to hit someone where it hurts")
Mettere i puntini sulle i ("to be nitpicking")
Mettere zizzania ("to sow discord")
Perdere la testa ("to lose one's head")
Perdere il treno ("to miss an opportunity")
Perdere il filo ("to lose the thread")
Perdere la bussola ("to lose one's bearings")
Prendere il toro per le corna ("to take the bull by the horns")
Prendere una cotta ("to get a crush on somebody")
Prendere un granchio ("to make a blunder")
Tirare i remi in barca ("to rest on one's oars")
Tirare la cinghia ("to tighten one's belt")
Tirare le cuoia ("to die")
Tirare la corda ("to take sth too far")

Alessandro Lenci
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
alessandro.lenci@unipi.it

Francesca Masini
Dipartimento di Lingue, Letterature e Culture Moderne
Università di Bologna
Via Cartoleria 5
40124 Bologna (Italy)
francesca.masini@unibo.it

Malvina Nissim
Faculty of Arts
University of Groningen
Oude Kijk in 't Jatstraat 26
9712 EK Groningen (The Netherlands)
m.nissim@rug.nl

Sara Castagnoli
Dipartimento di Scienze della Formazione, dei Beni Cultirali e del Turismo
Università di Macerata
P.le Luigi Bertelli 1
62100 Macerata (Italy)
sara.castagnoli@unimc.it

Gianluca E. Lebani
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
gianluca.lebani@for.unipi.it

Lucia C. Passaro
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
Università di Pisa
Via Santa Maria 36
56126 Pisa (Italy)
lucia.passaro@for.unipi.it

Marco S.G. Senaldi
Laboratorio di Linguistica "G. Nencioni"
Scuola Normale Superiore
Piazza dei Cavalieri 7
56127 Pisa (Italy)
marco.senaldi@sns.it