Mapping the Constructicon with SYMPAThy. Italian Word Combinations between fixedness and productivity

Alessandro Lenci
Gianluca E. Lebani, Marco S. G. Senaldi
University of Pisa
alessandro.lenci@ling.unipi.it
gianluca.lebani@for.unipi.it
marco.senaldi@sns.it

Sara Castagnoli
Francesca Masini
University of Bologna
s.castagnoli@unibo.it
francesca.masini@unibo.it

Malvina Nissim
University of Groningen
m.nissim@rug.nl

Abstract

This work introduces SYMPAThy, a data representation model in which the combinatorial properties of a lexical item are described by merging surface and deeper linguistic information. The proposed approach is then evaluated by comparing, for a sample list of verbal idioms, a set of SYMPAThy-based fixedness indexes against the relevant speaker-elicited indexes available in the descriptive norms collected by Tabossi et al. (2011).

1 Word combinatorics and constructions

By "Word Combinations" (WoCs) we broadly refer to the range of constructions typically associated with a lexical item. In Construction Grammar, constructions (Cxn) are conventionalized form-meaning pairings that can vary in both complexity and schematicity (Fillmore et al., 1988; Goldberg, 2006; Hoffmann and Trousdale, 2013). The Constructicon spans from fully specified structures (*kick the bucket*) to complex, productive abstract structures such as argument patterns (e.g., the Ditransitive Cxn "Subj V Obj1 Obj2", *she baked him a cake*), passing through "intermediate" Cxns with different degrees of schematicity, complexity and productivity (e.g., *take* Obj *for granted*), in what is known as the lexicon-syntax continuum. WoCs thus comprise so-called Multiword Expressions (MWEs), i.e. a variety of recurrent expressions acting as a single unit at some level of linguistic analysis, like phrasal lexemes, idioms, collocations (Calzolari et al., 2002; Sag et al., 2002; Gries, 2008), as well as the preferred distributional properties of a word at a more abstract level, i.e. argument structures and selectional preferences (Goldberg, 1995).

Each lexeme can thus be described as having a *combinatory potential* to be defined and observed at a more constrained, surface POS-pattern level (P-level) *and* at the more abstract level of syntactic structure (S-level). These two levels are often kept separate, not only theoretically, but also computationally, as their performance varies according to the different types of combinations that we want to track (Sag et al., 2002; Evert and Krenn, 2005).

We advocate a unified and integrated view of a lexeme's combinatory potential, in order to capture both fixed combinations (MWEs of various types) and more productive aspects of the lexeme's distributional behaviour. The theoretical premises lie in the constructionist view of the mental lexicon outlined above, whereas a proposal for a computational implementation is illustrated here. Specifically, we i) present SYMPAThy, *a model of data representation* that takes into account both surface and deeper linguistic information; ii) develop and test an *index of productivity* for Italian WoCs based on SYMPAThy.

2 SYMPAThy: a joint approach to WoCs

We argue that to obtain a comprehensive picture of the combinatory potential of a word and enhance extracting efficacy for WoCs, the P-based approach (which exploits sequences of POS-patterns and association measures) and the S-based approach (which exploits syntactic dependencies and association measures) should be combined. We illustrate this point with an example based on the Target Lexeme (TL) *gettare* 'throw' (V).[1]

We want to use S-based methods to capture the fact that V occurs typically within some syntactic Frames and not others, that for each Frame we have typical Fillers (lexical items) instantiating Frame slots, and that each slot is associated with certain semantic (ontological) classes:[2]

[1] All data is from a version of the "la Repubblica" corpus (Baroni et al., 2004) POS tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency parsed with DeSR (Attardi and Dell'Orletta, 2009).

[2] Data extracted by LexIt (Lenci, 2014). The list is partial: only the first three Frames are included; Frames with the re-

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

- 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, ...}

At this point, we observe that all these words are typically associated with our TL, but we don't know in which way they are all linked to one another. For instance, we have no elements for thinking that subj#*gettare#acqua#su_fuoco* is any different from subj#*gettare#acqua#su_tavolo* or subj#*gettare#ombra#su_istituzione*. However, while *gettare acqua sul fuoco* 'defuse' is an idiom in Italian, *gettare acqua sul tavolo* only has a literal meaning ('throw water on the table'); subj#*gettare#fango#su_istituzione* is yet different, since *gettare fango su* 'defame' is a fixed expression, but the Filler *istituzione* 'institution' is just one of many possibilities, so the expression is partially fixed, resulting in something like [*gettare fango su* PERSON/INSTITUTION]. The significance of *gettare acqua sul fuoco* with respect to *gettare acqua sul tavolo* emerges much more clearly if we use a P-based method. Extracting surface material, the former expression will be ranked higher than the latter (given the pattern "V N PREPART N") as the association between all words is stronger.

So, fine-grained differences do not emerge with the S-method, while the P-based method fails to capture the higher-level generalizations we get with the S-method. In order to get the best of both worlds, we extracted corpus data into SYMPAThy (SYntactically Marked PATterns), a database where information on both levels is stored and accessible jointly:

- syntactic frames with argument slots and fillers;
- linear order of all elements for each TL;
- POS tag for each element (simple preposition vs. preposition with article, definite vs. indefinite article, modal vs. full verb, etc.);
- morphosyntactic features: gender, number, finiteness, tense, etc.

flexive form *gettarsi* 'throw oneself' and objectless forms are excluded.

3 WoC fixedness with SYMPAThy

Since constructions span along a continuum between fixedness and productivity, there have been various attempts at measuring how fixed a given WoC is, mostly based on surface features. Nissim and Zaninello (2011) assess the fixedness of a subset of complex nominals by comparing inflected and lemmatized forms, and taking into account the proportion of elements that undergo variation in a given MWE. Inflection is also used by Squillante (2014) on noun-adjective expressions, and is combined with two other measures, interruptibility and substitutability. Zeldes (2013) extends Baayen's morphological productivity approach to argument structure and estimates the productivity of a syntactic slot from the number of its hapax noun fillers. Wulff (2009) uses a set of morphosyntactic indexes of variations and a collocation-based index of compositionality as variables in a regression study to determine fixedness.

We extend the state of the art of the quantitative approach to construction fixedness by exploiting the potentialities of SYMPAThy to develop a series of corpus-based indexes able to describe the fixedness of some idiomatic expressions. Our approach is then evaluated by comparing, for a sample list of expressions, a composition of our indexes against the behavioral judgments of syntactic flexibility collected by Tabossi et al. (2011).

3.1 The combinatory behaviour of a TL

In the SYMPAThy model, the combinatory space of a Target Lexeme is assumed to be formed by a network of Cxns, varying for their degree of fixedness/productivity. For any given TL such a representation is built by means of the following four-step procedure:

1. its SYMPAThy patterns are extracted from a reference corpus;
2. the set of single and multiple slot Cxns that TL combines with are semi-automatically identified. An example for the verb *gettare* is reported and explained in Appendix 1;
3. each construction is associated with a *variational profile* formed by a number of statistics extracted from the SYMPAThy pattern to estimate: i) the variability of the fillers that instantiate the syntactic slots of constructions; ii) the

morphological variability of the constructions' components; iii) the variability with respect to determiners; iv) the variability with respect to adjectival and adverbial modifications; v) the variability in the linear order.

4. variational profiles are then used to measure the lexical, morphological and syntactic *degrees of freedom* of Cxns, providing a multidimensional quantitative characterization of their level of fixedness.

3.2 Entropy-based Cxn fixedness modeling

In what follows, we devise a way to encode the variation possibilities shown by Cxns, as well as a meaningful way to combine them. Specifically, we distinguish a series of dimensions of variation and propose to exploit Entropy (Shannon, 1948) to measure how fixed is the behavior of a Cxn in a given dimension.

Entropy is a measure of randomness, calculated as the average uncertainty of a single variable:

$$H(X) = -\sum_{x \in X} p(x) \log_2(p(x)) \qquad (1)$$

This measure of randomness can be adapted to our needs by taking the variable X as being a Cxn of interest, and the states of the system x as its values on one dimension of variation. Lower entropy values are to be understood as evidence of fixedness, while higher values suggest a more variable distribution of the states of a given variable, i.e. the target construction tends to be freer.

Observed entropy values, however, can span from 0 to the logarithm of the number of values that X can assume. As a consequence, entropy values related to different dimensions of variation are not comparable, and cannot be combined into a single fixedness index. We overcome this limitation by following Wulff (2008) and describing the randomness of each variability dimension in terms of relative entropy, computed as the ratio between the observed entropy from eq.1 and the maximum entropy H_{max} for the variable X:

$$H_{rel}(X) = \frac{H(X)}{H_{max}(X)} = \frac{H(X)}{\log_2(|X|)} \qquad (2)$$

This measure, that ranges from 0 to 1, has been employed as a flexibility measure to describe the flexibility of a given set of target Cxns along the following dimensions of variation:

LEXICAL VARIABILITY. The entropy of the lexical instantiation of the slot positions of a Frame is calculated by assuming that the states x of the random variable X are all the possible fillers that can instantiate a given slot in Cxn (e.g. in subj#*gettare*#obj:*luce*#su_X, X can be filled by *vicenda* 'matter', *mistero* 'mystery', etc.).

MORPHOLOGICAL VARIABILITY. It is calculated as the entropy of the morphological features manifested by the fillers of a Cxn (e.g., *gettare#ombra-fs* 'cast shadow-singular'; *gettare#ombra-fp* 'cast shadow-plural').

ARTICLES VARIABILITY. This index encodes how variable is the presence or absence of articles determining the available slots in a Cxn, and, if appropriate, their type (DEFinite vs. INDefinite): for instance, *gettare*#∅+*acqua*#su_*DEF*+*fuoco*.

PRESENCE OF MODIFIERS. This index encodes how variable is the presence or absence of adjectives, adverbs or prepositional phrases modifying the available slots. In this way, it is possible to account for patterns like:*gettare*#*molta*+*acqua*#su_∅+*fuoco*.

DISTANCE VARIABILITY. This index exploits information on linear order available in SYMPATHY to estimate how variable is the distance in tokens between a TL and the other constituents of a given lexically specified Cxn.

In the experiment reported in the next section, we have combined the single variability measures $H_{rel}(X)$ into an overall flexibility index $F(X)$ corresponding to four possible combinations:

- SUM: $F(X)$ is obtained by summing over all the single $H_{rel}(X)$ values;
- AVERAGE: $F(X)$ is the mean of the single $H_{rel}(X)$ values;
- AVERAGE$_{POS}$: $F(X)$ is the mean of the positive $H_{rel}(X)$ values;
- MAX: $F(X)$ is the highest $H_{rel}(X)$ value.

We leave to future research the investigation of further ways to combine the variability indexes.

4 Evaluation

In order to evaluate our approach, we set out to test if our indexes can mimic the intuitive judgments of native speakers about the fixedness of fully lexically specified constructions. To do so, we selected a subset of the idioms in the norms collected

by Tabossi et al. (2011), and tested to what degree the speaker-elicited flexibility judgments available in this repository can be modeled by a composition of our variability indexes.

4.1 The descriptive norms by Tabossi et al.

Tabossi et al. (2011) collected several normative measures for 245 Italian verbal idiomatic expressions. Using a group of 740 Italian speakers, they collected a minimum of 40 elicited judgments for each idiom on several psycholinguically relevant variables.

Among the different kinds of ratings, those concerning syntactic flexibility have been collected by inserting each idiomatic expression in 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 7-point 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.

4.2 Data extraction

Out the 245 expressions in Tabossi et al., we selected the 23 target idioms reported in Appendix 2. Each such idiom can be represented, in our approach, as a fully lexically specified transitive Cxn headed by a given verbal TL, for which the subject slot is underspecified (e.g. *gettare*#obj:*maschera*). We built the variational profiles of our target idioms by adopting an adapted version of the procedure described in Section 3:

1. for each TL, we extracted the 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 indexes described in Section 3.2 were calculated. Note that, given the nature of our experimental stimuli, the lexical variability index is not relevant;
4. we built a fixedness index for each idiom, according to the four composition methods in the previous section.

4.3 Results and discussion

In order to test the cognitive plausibility of the fixedness indexes extracted from SYMPAThy, we calculated the Pearson's Product-Moment Correlation strength between them and the syntactic flexibility ratings in Tabossi et al. (2011). Correlation values are reported in Table 1. In all cases, there is a significant ($p < .05$) positive correlation, ranging between .44 and .47, thus supporting the psycholinguistic plausibility of our corpus-based variability indexes.

Combination	***r***
SUM	.44
AVERAGE	.44
AVERAGE $_{POS}$	.46
MAX	.47

Table 1: Pearson's Correlation strength between different combination methods of the SYMPAThy-based fixedness indexes and the syntactic flexibility judgments in Tabossi et al. (2011). All reported values are associated with $p < .05$, $N = 23$.

These results, albeit preliminary, look promising especially given the different nature of the behavioral and corpus-based indexes. On the one hand, the speakers' ratings are semantically driven, since they are thought to model how much the figurative meaning of a given idiom is sensitive to its syntactic form. On the other hand, the automatically corpus-derived information exploited by our indexes does not take meaning into account. SUch indexes describe a lexically specified Cxn that can in principle have an idiomatic as well as a compositional, literal meaning (even if, presumably, the latter case is rare in the corpus).

5 Conclusion

In this study we presented a procedure for characterizing the combinatorial potential of a lexical item and the degree of fixedness of the Cxns it occurs in. Such a procedure has been preliminary tested on a small sample of idiomatic expressions and the resulting representation has been evaluated against the subject-elicited judgments collected by Tabossi et al. (2011). In the future, we are planning to extend the inventory of variability dimensions (addressing also the question of the semantic compositionality of Cxns), to study their relative weight and their interactions, and to develop more sophisticated ways to combine them.

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) funded by the Italian Ministry of Education, University and Research (MIUR).

References

[Attardi and Dell'Orletta2009] Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of NAACL 2009*, pages 261–264.

[Baroni et al.2004] 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.

[Calzolari et al.2002] 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.

[Dell'Orletta2009] Felice Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*.

[Evert and Krenn2005] 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.

[Fillmore et al.1988] 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.

[Goldberg1995] Adele Goldberg. 1995. *Constructions. A Construction Grammar Approach to Argument Structures*. The University of Chicago Press, Chicago.

[Goldberg2006] Adele Goldberg. 2006. *Constructions at work*. Oxford University Press, Oxford.

[Gries2008] 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.

[Hoffmann and Trousdale2013] Thomas Hoffmann and Graeme Trousdale, editors. 2013. *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

[Lenci2014] Alessandro Lenci. 2014. Carving verb classes from corpora. In Raffaele Simone and Francesca Masini, editors, *Word Classes. Nature, typology and representations*, Current Issues in Linguistic Theory, pages 17–36. John Benjamins.

[Nissim and Zaninello2011] Malvina Nissim and Andrea Zaninello. 2011. A quantitative study on the morphology of Italian multiword expressions. *Lingue e Linguaggio*, X:283–300.

[Sag et al.2002] 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.

[Shannon1948] Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379 – 423.

[Squillante2014] Luigi Squillante. 2014. Towards an empirical subcategorization of multiword expressions. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 77–81, Gothenburg, Sweden, April. Association for Computational Linguistics.

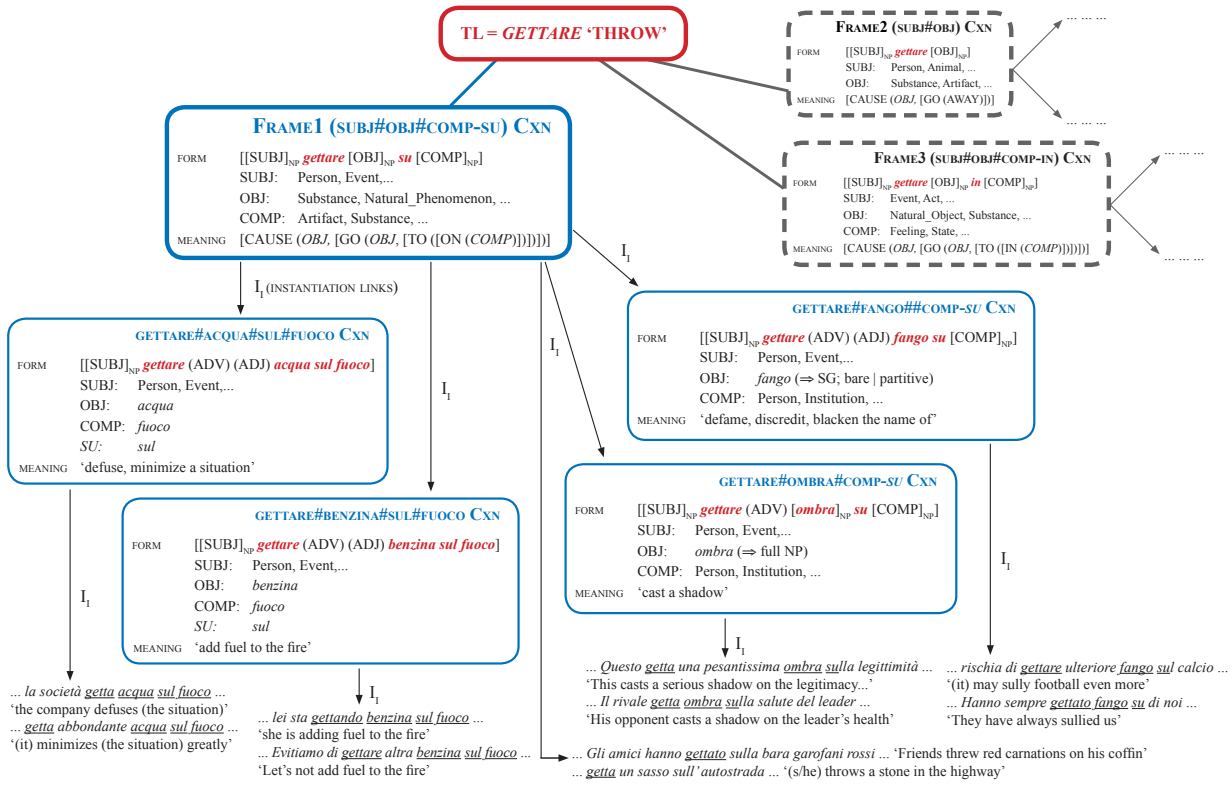
[Tabossi et al.2011] Patrizia Tabossi, Lisa Arduino, and Rachele Fanari. 2011. Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods*, 43:110–123.

[Wulff2008] Stefanie Wulff. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. Continuum.

[Wulff2009] Stefanie Wulff. 2009. Converging evidence from corpus and experimental data to capture idiomaticity. *Corpus Linguistics and Linguistic Theory*, 5(1):131–159.

[Zeldes2013] Amir Zeldes. 2013. Productive argument selection: Is lexical semantics enough? *Corpus Linguistics and Linguistic Theory*, 9(2):263–291.

Appendix 1: A SYMPAThy-based view of the network of Cxns with the verb *gettare*



The verb *gettare* 'to throw' combines with the highly schematic subj#obj#comp-su Cxn, whose slots can freely vary with respect to linear order, presence of determiners, modifiers, etc. A semi-productive instance of this construction is the subj#obj:*ombra*#comp-su Cxn, with a fixed object slot and a partially variable oblique slot, which can appear with a semantically limited range of arguments. A fully lexically specified instance of the same construction is instead the subj#obj:*acqua*#comp-su:*sul-fuoco* Cxn, which has both slots instantiated and limited degree of variability.

Appendix 2: List of idioms used as experimental stimuli

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*')