Multiword expressions we live by: a validated usage-based dataset from corpora of written Italian

Francesca Masini[1], M. Silvia Micheli[2], Andrea Zaninello[3], Sara Castagnoli[4], Malvina Nissim[5]
[1]Alma Mater Studiorum – University of Bologna, [2]University of Milano-Bicocca
[3]Zanichelli editore, [4]University of Macerata, [5]CLCG, University of Groningen
francesca.masini@unibo.it, maria.micheli@unimib.it
azaninello@zanichelli.it, sara.castagnoli@unimc.it, m.nissim@rug.nl

Abstract

The paper describes the creation of a manually validated dataset of Italian multiword expressions, building on candidates automatically extracted from corpora of written Italian. The main features of the resource, such as POS-pattern and lemma distribution, are also discussed, together with possible applications.

1 Introduction

The computational treatment of multiword expressions (henceforth, MWEs) is notoriously a major challenge in NLP (Ramish, 2015; Villavicencio et al., 2005). In the last decades, the (computational) linguistics community has dedicated many efforts to the development of techniques for the (semi-)automatic identification and extraction of MWEs from corpora and the consequent creation of resources, such as gold standard lists of MWEs, which are needed for evaluation tasks or machine learning training. This notwithstanding, the availability of such resources is still quite limited compared with "the ubiquitous and pervasive nature of MWEs" (Ramish, 2015), especially for 'non-mainstream' languages like Italian.

With this work, we contribute to this line of research by providing a dataset of 1,682 validated Italian multiword expressions, obtained through the manual annotation of candidates automatically extracted from corpora of written Italian within the CombiNet project (Simone and Piunno, 2017b). The dataset is to be intended as a first release that will be enriched in the future. We describe our methodology in Section 2, while in Section 3 we report on preliminary analyses carried out with respect to MWE features and distribution.

2 Methodology

For the creation of the dataset we built on data extracted within the CombiNet project, where the computational task of extracting candidate word combinations from corpora was aimed at supporting the creation of an online lexicographic resource for Italian (Simone and Piunno, 2017a). The notion of 'word combination' was large enough to encompass both MWEs (Calzolari et al., 2002; Sag et al., 2002; Gries, 2008; Baldwin and Kim, 2010) – namely strings endowed with (different degrees of) fixedness, idiomaticity or simply conventionality – and more abstract distributional properties of a word, such as argument structures, subcategorization frames or selectional preferences (Lenci et al., 2017).

As a consequence, two different extraction methods – both based on the technique of searching corpora[1] with sets of patterns, and ranking retrieved candidates using frequency and association measures – were used.[2] More precisely, the search was performed using, in turn, shallow part-of-speech (POS) sequences and syntactic relations: the former method performs better with fixed and adjacent word combinations, whereas the latter is more efficient for syntactically flexible combinations. Since for the present work we focus more on MWEs proper rather than combinatorics in general, we opted to use the data previously gathered with the POS-based method.

Candidates were obtained by feeding the EXTra software (Passaro and Lenci, 2015) with a list of 122 POS-patterns deemed representative of Italian

[1]The corpora used within CombiNet were *la Repubblica* (Baroni et al., 2004) and *PAISÀ* (Lyding et al., 2014).

[2]For a full description of the methods and their assessment, see (Lenci et al., 2017) In what follows we only provide information which is relevant for the current discussion.

MWEs, derived from both relevant literature and a corpus-driven identification task; the list includes adjectival, adverbial, nominal, prepositional and verbal patterns, up to five slots (see Lenci et al., 2017). The results were ranked by LogLikelihood.

As a first step, we selected top-ranked results by cutting at LL $\geq$ 7,500, which we observed to be a good balance between precision (high chance of being a MWE) and recall (enough variety), yielding 7,045 candidates. Then we manually annotated this list of candidates to obtain the gold standard inventory of Italian MWEs released and described in the present paper. Each candidate was validated independently by two annotators, and a third annotator judged the conflicted cases,[3] which amounted to 673 (less than 10%). We validated sequences that were deemed to display some type of conventionality (fixedness, idiomaticity, high familiarity of use). We included only MWEs in their 'full' form (e.g., *punto di partenza* 'starting point', *in breve tempo* 'in a short time'), thus excluding sequences that were clearly part of incomplete MWEs (e.g. *scanso di equivoci*, lit. avoidance of misunderstandings, as part of the larger adverbial MWE *a scanso di equivoci*, lit. at avoidance of misunderstandings, 'to avoid misunderstandings').

3 The Resource

The final list of valid MWEs amounts to 1,682 (about 24% of the candidates), and is made available to the community.[4] The resource contains the following information: (i) lemmatized MWE;[5] (ii) corresponding POS-pattern;[6] (iii) corpus/corpora where the MWE was found; (iv) LogLikelihood; (v) raw frequency.

3.1 Caveat

In order to make our resource re-usable on the very same corpora employed for the extraction, we kept all data in their original form. This means that lemmatization and POS-tagging were retained, even if erroneous.

Examples of errors and anomalies include:

(a) inconsistent lemmatization, especially for prepositions (e.g. *radere al suolo* 'raze to the ground' occurs twice, lemmatized as *radere a suolo* and *radere al suolo*, although the preposition is correctly tagged as an articulated preposition in both cases) and conjunctions (e.g. *carne e ossa* 'flesh and blood' and the almost identical *carne ed ossa*, with the euphonic *-d* on the conjunction *e* 'and', are two separate items);

(b) wrong lemmatization and tagging, especially for participial-like forms (e.g. *centro abitato* 'residential area', lit. center inhabited, lemmatized as *centro abitare*, lit. center to inhabit; or *posta elettronica* 'electronic mail' lemmatized as *porre elettronico*, lit. to put electronic, since *posta* is interpreted as the feminine past participle of *porre* 'to put' and not as the noun *posta* 'mail'), but not only (e.g. *lavori di costruzione* 'construction works' lemmatized as *lavorio* [instead of *lavoro*] *di costruzione*; or *meccanica quantistica* 'quantum mechanics' where *meccanica* is tagged as an adjective);

(c) multiple tagging for the same form (*essere vero* 'be true' occurs twice because *vero* is tagged sometimes as an adjective, sometimes as an adverb).

Tricky cases also include lexicalized forms (*guarda caso* 'strangely enough', where *guarda* is – correctly, from the technical point of view – lemmatized as *guardare* 'look' and tagged as verb, although it is no longer a verb within that lexicalized expression) and pronominal verbs (like *sentirsi in dovere* 'to feel obliged', where the verb is lemmatized as *sentire* 'to feel', and not as its reflexive form *sentirsi*, although the MWE requires the reflexive form).

3.2 POS-patterns

The validated MWEs in this first release instantiate 82 POS patterns out of the 122 used for the extraction (cf. Section 2). Non-represented patterns (over 30% of the original set) include e.g. Prep-Adj-Verb (e.g. *per quieto vivere* 'for a quiet life') as well as more complex – and arguably less frequent – patterns such as N-Prep-ArtDef-N-Adj (e.g. *lotta contro la criminalità organizzata* 'fight against organized crime').

[3] All annotations were performed by the authors.

[4] DOI: 10.6092/unibo/amsacta/6506. http://amsacta.unibo.it/id/eprint/6506

[5] MWEs are lemmatized because the extraction was performed using lemmas. A consequence of this is that we may have two identical lemmatized sequences that however differ in POS-tagging. For instance, *cambio di guardia* (lit. change of guard) occurs twice: in one case *di* 'of' is tagged as a bare preposition, in the other as an articulated preposition (*della* 'of the'), giving rise to two partially different MWEs (the latter may mean both 'changing of the guard' and 'changeover of leaders', whereas the former can refer only to the second of these meanings).

[6] The tagset is available here: http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

Table 1 shows the most attested patterns, while Table 2 the rarely attested ones (only one MWE in our dataset).

Pattern	Fq.	Example
N-Prep-N	165	*punto di vista* 'viewpoint'
V-ArtDef-N	152	*valere la pena* 'to be worth'
V-Prep-N	110	*scendere in campo* 'to take the field'
V-N	83	*avere paura* 'to be afraid'
V-ArtIndef-N	83	*correre un rischio* 'to run a risk'
N-A	80	*tavola rotonda* 'round table'
N-PrepArt-N	79	*vigile del fuoco* 'fireman'
Prep-N-Prep	77	*di fronte a* 'in front of'
PrepArt-N-Prep	75	*al fine di* 'with the aim of'
Prep-N	63	*di parte* 'biased'
V-Adv	62	*andare avanti* 'to go on'
N-N	62	*piano terra* 'ground floor'
V-Adj	55	*essere presente* 'to be there'
V-PrepArt-N	47	*entrare nel merito* 'to address'
Prep-ArtDef-N	35	*dietro le quinte* 'behind the scenes'

Table 1: Most attested POS-patterns

Overall, most attested patterns are 2- or 3-grams. The first 4-slot pattern V-Prep-ArtIndef-N only appears at rank 36, corresponding to 8 different MWEs (e.g. *rispondere a una domanda* 'to answer a question').

In terms of lexical categories, expectedly, most frequent patterns pertain to the nominal and verbal domains. The N-Prep(Art)-N type is the most common pattern for complex nominals, in agreement with theoretical literature (Masini, 2009, e.g.). Patterns headed by prepositions and giving rise to complex prepositions, conjunctions and modifiers are also numerous.

Pattern	Fq.	Example
Prep-Adj-Conj-Adj	1	*in bianco e nero* 'in black and white'
V-ArtDef-N-A	1	*dare il via libera* 'to give green light'
A-Prep-V	1	*difficile a dirsi* 'difficult to say'
V-Prep-Adj-N	1	*mettere a dura prova* 'to put a strain (on)'
Adj-Prep-N	1	*degno di nota* 'noteworthy'

Table 2: Least attested POS-patterns

3.3 Lemmas used to form MWEs

The single-word lemmas that concur to form the MWEs in our list amount to 1,235.

Not surprisingly, among the most used lemmas we find function words like prepositions (*di* 'of' fq.421; *in* 'in' fq.227; *al* 'at/to the' fq.124, *a* 'at/to' fq.55 and *ad* 'at/to' fq.10; *per* 'for' fq.50; *da* 'from' fq.34; *su* 'on' fq.24; *con* 'with' fq.20) and determiners (*il* 'the' fq.208; *un* 'a' fq.71 and *una* 'a' fq.41), which appear in many POS-patterns. Conjunctions are instead less frequent (*e* 'and' fq.21 and *ed* 'and' fq.4; *o* 'or' fq.4), like quantifiers (e.g. *ogni* 'each' fq.11).

Quite expectedly, top-ranked verbs (*essere* 'to be' fq.67; *fare* 'to do/make' fq.46; *avere* 'to have' fq.36; *mettere* 'to put' fq.35; *prendere* 'to take' fq.27; *andare* 'to go' fq.19; *dare* 'to give' fq.17) and top-ranked nouns (*tempo* 'time' fq. 32; *mano* 'hand' fq.26; *parte* 'part' fq.23; *posto* 'place' fq.17; *giorno* 'day' fq.16) are lexemes carrying a generic meaning, which favors their combinatory power. Among the mostly used words we also find numerals like *primo* 'first' (fq.30) or *secondo* 'second' (fq.18), and adverbs like *non* 'not' (fq.29).

A cursory comparison between the lemmas of the MWEs in our list and the *Vocabolario di Base* (De Mauro, 1980), which contains the 7,000 most common lemmas in Italian, shows a large convergence: well over 70% of our lemmas are included in the *Vocabolario di Base*. Thus, very frequent MWEs also feature very common lexical items.

3.4 Distribution in corpora

The distribution of MWEs in the two corpora used for the extraction is shown in Table 3.

We retrieved more MWEs from *la Repubblica*

Corpus	N. of MWEs
la Repubblica (total)	1354
PAISÀ (total)	700
la Repubblica (only)	982
PAISÀ (only)	328
Both	372

Table 3: Distribution of MWEs in the two corpora. "Only" indicates how many MWEs are specific to one corpus only and are not found in the other.

than *PAISÀ*, which is expected given that the latter is smaller in size (250M tokens vs. 380M). What is less expected is the rather low number of MWEs shared by the two corpora, amounting to 372, hence 22%. Although *la Repubblica* is a journalistic source and *PAISÀ* is a web corpus containing more varied text genres (especially from Wikimedia Foundation projects), we expected a larger convergence, considering that they both contain written (mid-)formal texts and that *PAISÀ* also contains texts from the news.

Some POS-patterns seem to be definitely more typical of one corpus over the other. As Table 4 illustrates, the N-Prep-N pattern, for instance, is much more typical of *la Repubblica*, whereas the N-Adj pattern is more attested in *PAISÀ*.

Corpus	N-Prep-N	N-Adj
la Repubblica (only)	120	36
PAISÀ (only)	27	44
Both	18	0

Table 4: Distribution of MWEs of two common POS patterns in the two corpora

Among top-ranked MWEs for both LogLikelihood and raw frequency we find *in grado di* 'able to' and *per la prima volta* 'for the first time', in both corpora. The highest ranked MWEs in *PAISÀ* is *voce correlata* 'see also', which is obviously due to the texts that form this resource. Generally, top-ranked MWEs for LogLikelihood also have high frequency, but not in all cases: *essere in essere* 'to exist', for instance, turns out to be highly significant in terms of LogLikelihood but has a very low frequency in both corpora.

4 Discussion

The sequences contained in this release are obviously quite heterogeneous.

Semantically speaking, some are very idiomatic in meaning (e.g. *braccio di ferro* 'arm wrestling', *colpo di scena* 'coup de théâtre', *mandare in onda* 'to broadcast'), some other (much) less so (e.g. *prendere le distanze* 'to distance (oneself)', *andare in pensione* 'to retire', *di servizio* 'service (adj.)'), their specialty lying more in their familiar, conventional status (e.g. *sapere benissimo* 'to know (damn) well', *essere favorevole* 'to be in favour', *nella storia* 'in history'). Still others may have more than one meaning, with different degrees of figurativity (e.g. *mettere in scena*, which can mean both 'to stage' and 'to enact').

From a formal point of view, some look rather fixed and do not admit lexical insertion (e.g. *vero e proprio* 'proper') or inflection (e.g. *tra l'altro* 'by the way', *ordine del giorno* 'agenda'), whereas others seem more flexible (e.g. *essere certo* 'to be sure', *andare bene* 'to be OK, to go well', *posto di lavoro* 'workplace'). MWE variability is one aspect that we did not address here but definitely deserves to be investigated more thoroughly (cf. e.g. (Nissim and Zaninello, 2011)). In fact, some MWEs may exhibit different behaviour and even completely different meanings according to their grammatical form, like, for example, *a suo tempo* 'in due course' (lit. in his/her time) vs. *ai suoi tempi* 'in his/her time' (lit. in his/her times). Being based on lemmatized forms, our study does not currently account for such form differences. Moreover, our study is based on contiguous sequences, therefore discontinuous or topicalized occurrences are not accounted for.

We also aim at broadening this initial list by exploring more candidates from the CombiNet data, which are obviously still rich of relevant material. This first release, although limited, is meaningful since it is the first list of commonly used MWEs available for the Italian language, except for domain-specific resources such as PANACEA (Frontini et al., 2012). Although lexicographic material is now accessible for Italian lexical combinatorics (see e.g. (Lo Cascio, 2013)), usage-based and freely available lists of MWEs are still missing and much needed, both for computational tasks and for applied (lexicographic and language teaching related) purposes.

Acknowledgments

This research relies on data extracted within the CombiNet project (PRIN 2010-2011 Word Combinations in Italian, n. 20105B3HE8), coordinated by Raffaele Simone and Alessandro Lenci, and funded by the Italian Ministry of Education, University and Research (MIUR).

References

Calzolari Nicoletta, Fillmore Charles J., Grishman Ralph, Ide Nancy, Lenci Alessandro, MacLeod Catherine and Zampolli Antonio. 2002. Towards best practice for multiword expressions in computational lexicons. In Rodríguez, M. G. and Araujo, C. P. S. (eds.), *Towards Best Practice for Multiword Expressions in Computational Lexicons*. *LREC*, 1934-40.

Baldwin Timothy and Kim Su Nam. 2010. Multiword expressions. In Indurkhya, N. and Damerau, F. J. (eds.), *Handbook of natural language processing*, 267-29. Taylor and Francis Group, Boca Raton (FL).

Baroni Marco, Bernardini Silvia, Comastri Federica, Piccioni Lorenzo, Volpi Alessandra, Aston Guy and Mazzoleni Marco. 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. (eds), *Proceedings of the Third International Conference on Language Resources and evaluation (LREC)*, 1771-4. European Language Resources Association, Lisbon.

De Mauro Tullio. 1980. *Guida all'uso delle parole*. Editori Riuniti, Roma.

Frontini Francesca, Quochi Valeria, and Rubino Francesco. 2012. Automatic Creation of quality Multi-word Lexica from noisy text data. In Kay, M, Boitet, C. (eds.), *Proceedings of COLING 2012: Sixth Workshop on Analytics for Noisy Unstructured Text Data: 24th International Conference on Computational Linguistics COLING 2012; 2012 December 8-15*. http://hdl.handle.net/10230/20422.

Gries Stefan T. 2008. Phraseology and linguistic theory: A brief survey. In Granger, S. and Meunier, F. (eds.), *Phraseology: An interdisciplinary perspective*, 3-25. John Benjamins, Amsterdam/Philadelphia.

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

Lenci Alessandro, Masini Francesca, Nissim Malvina, Castagnoli Sara, Lebani Gianluca E., Passaro Lucia C. and Senaldi Marco S. G. 2017. How to harvest Word Combinations from corpora: Methods, evaluation and perspectives. *Studi e Saggi linguistici*, 55(2): 45-68.

Lo Cascio Vincenzo. 2013. *Dizionario combinatorio italiano*. John Benjamins, Amsterdam/Philadelphia.

Lyding Verena, Stemle Egon, Borghetti Claudia, Brunello Marco, Castagnoli Sara, Dell'Orletta Felice, Dittmann Henrik, Lenci Alessandro and Pirrelli Vito 2014. The PAISA corpus of Italian web texts. *9th Web as Corpus Workshop (WaC-9)@ EACL 2014*, 36-43. EACL (European chapter of the Association for Computational Linguistics.

Masini Francesca. 2009. Phrasal lexemes, composunds and phrases. *Word Structure*, 2(2): 254-71.

Nissim Malvina and Zaninello Andrea. 2011. A quantitative study on the morphology of Italian multiword expressions. *Lingue e linguaggio*, 10(2): 283-300.

Passaro Lucia C. and Lenci Alessandro. 2015. Extracting Terms with EXTra. In Corpas Pastor, G. (ed.), *Computerised and Corpus-based Approaches to Phraseology. Monolingual and Multilingual Perspective*, 188-196. Editions Tradulex, Geneva.

Ramisch Carlos. 2015. *Multiword Expressions Acquisition - A Generic and Open Framework*. Springer, Dordrecht.

Sag Ivan A., Baldwin Timothy, Bond Francis, Copestake Ann and Flickinger Dan. 2002. Multiword expressions: A pain in the neck for NLP. *International conference on intelligent text processing and computational linguistics*, 1-15. Springer, Dordrecht.

Simone Raffaele and Piunno Valentina. 2017a. Entry word combination: lexicographical representation and lexicological aspects. *Studi e Saggi Linguistici*, 55(2): 13-44.

Simone Raffaele and Piunno Valentina, editors. 2017b. Word Combinations: phenomena, methods of extraction, tools, Special Issue of *Studi e Saggi Linguistici*, 55(2).

Villavicencio Aline, Bond Francis, Korhonen Anna, McCarthy Diana. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech and Language*, 19(4): 365-377.