Language resources for Italian: towards the development of a corpus of annotated Italian multiword expressions

Shiva Taslimipoor
University of Wolverhampton, UK
shiva.taslimi@wlv.ac.uk

Anna Desantis, Manuela Cherchi
University of Sassari, Italy
annadesantis_91@libero.it,
manuealacherchi82@gmail.com

Ruslan Mitkov
University of Wolverhampton, UK
r.mitkov@wlv.ac.uk

Johanna Monti
"L'Orientale" University of Naples, Italy
jmonti@unior.it

Abstract

English. This paper describes the first resource annotated for multiword expressions (MWEs) in Italian. Two versions of this dataset have been prepared: the first with a fast markup list of out-of-context MWEs, and the second with an in-context annotation, where the MWEs are entered with their contexts. The paper also discusses annotation issues and reports the inter-annotator agreement for both types of annotations. Finally, the results of the first exploitation of the new resource, namely the automatic extraction of Italian MWEs, are presented.

Italiano. *Questo contributo descrive la prima risorsa italiana annotatata con polirematiche. Sono state preparate due versioni del dataset: la prima con una lista di polirematiche senza contesto, e la seconda con annotazione in contesto. Il contributo discute le problematiche emerse durante l'annotazione e riporta il grado di accordo tra annotatori per entrambi i tipi di annotazione. Infine vengono presentati i risultati del primo impiego della nuova risorsa, ovvero l'estrazione automatica di polirematiche per l'italiano.*

1 Rationale

Multiword expressions (MWEs) are a pervasive phenomenon in language with their computational treatment being crucial for users and NLP applications alike (Baldwin and Kim, 2010; Granger and Meunier, 2008; Monti et al., 2013; Monti and Todirascu, 2015; Seretan and Wehrli, 2013). However, despite being desiderata for linguistic analysis and language learning, as well as for training and evaluation of NLP tasks such as term extraction (and Machine Translation in multilingual scenarios), resources annotated with MWEs are a scarce commodity (Schneider et al., 2014b). The need for such types of resources is even greater for Italian which does not benefit from the variety and volume of resources as does English.

This paper outlines the development of a new language resource for Italian, namely a corpus annotated with Italian MWEs of a particular class: verb-noun expressions such as *fare riferimento*, *dare luogo* and *prendere atto*. Such collocations are reported to be the most frequent class of MWEs and of high practical importance both for automatic translation and language learning. To the best of our knowledge, this is the first resource of this kind in Italian.

The development of this corpus is part of a multilingual project addressing the challenge of computational treatment of MWEs. It covers English, Spanish, Italian and French and its goal is to develop a knowledge-poor methodology for automatically identifying MWEs and retrieving their translations (Taslimipoor et al., 2016) for any pair of languages. The developed methodology will be used for Machine Translation and multilingual dictionary compilation, and also in computer-aided tools to support the work of language learners and translators.

Two versions of the above resource have been produced. The first version consists of lists of MWEs annotated out-of-context with a view to performing fast evaluation of the developed methodology (out-of-context mark-up). The second version consists of annotated MWEs along with their concordances (in-context annotation).

The latter type of annotation is time-consuming, but provides the contexts for the MWEs annotated.

2 Annotation of MWEs: out-of-context mark-up and in-context annotation

After more than two decades of computational studies on MWEs, the lack of a proper gold standard is still an issue. Lexical resources like dictionaries have limited coverage of these expressions (Losnegaard et al., 2016) and there is also no proper tagged corpus of MWEs in any language (Schneider et al., 2014b).

Most previous studies on the computational treatment of MWEs have focused on extracting types (rather than tokens)[1] of MWEs from corpora (Ramisch et al., 2010; Villavicencio et al., 2007; Rondon et al., 2015; Salehi and Cook, 2013). The widely-used toolboxes of MWEToolkit (Ramisch et al., 2010) or Xtract (Smadja, 1993) extract expressions if their statistical occurrences represent the likelihood of them being MWEs. The evaluation for the type-based extraction of MWEs has been mostly performed against a dictionary (de Caseli et al., 2010), lexicon (Pichotta and DeNero, 2013) or list of human-annotated expressions (Villavicencio et al., 2007). However, there are some examples like the expression *have a baby*, which in exactly the same form and structure, might be an MWE (meaning *to give birth*) in some contexts and a literal expression in others.

As for the automatic identification of the tokens of MWEs, Fazly et al. (2009) make use of both linguistic properties and the local context, in determining the class of an MWE token. They report an unsupervised approach to identifying idiomatic and literal usages of an expression in context. Their method is evaluated on a very small sample of expressions in a small portion of the British National Corpus (BNC), which were annotated by humans. Schneider et al. (2014a) developed a supervised model whose purpose is to identify MWEs in context. Their methodology results in a corpus of automatically annotated MWEs. It is not clear, however, if the methodology is able to tag one specific expression as an MWE in one context and non-MWE in another. The PARSEME shared task[2] is also devoted to annotating verbal MWEs in several languages. The shared task, while having interesting discussions on the area, has embarked upon the labour-intensive annotation of verbal MWEs.

Since there is no list of verb-noun MWEs in Italian, we first automatically compile a list of such expressions, to be annotated by human experts. This is based on previous attempts at extracting a lexicon of MWEs (as in (Villavicencio, 2005)). Annotators are not provided with any context and hence the task is more feasible in terms of time. Human annotators are asked to label the expressions as MWEs only if they have sufficient degrees of idiomaticity. In other words, a Verb + Noun MWEs does not convey literal meaning in that the verb is delexicalised.

However, we believe that idiomaticity is not a binary property; rather it is known to fall on a continuum from completely semantically transparent, or literal, to entirely opaque, or idiomatic (Fazly et al., 2009). This makes the task of out-of-context marking-up of the expression more challenging for annotators, since they have to pick a value according to all the possible contexts of a target expression. This ambiguity and the fact that there are many expressions that in some contexts are MWEs and in some contexts not, prompted us to initiate a subsequent annotation where MWEs are tagged in their contexts. The idea is to extract the concordances around all the occurrences of a Verb + Noun expression and provide annotators with these concordances in order to be able to decide the degree of idiomaticity of the specific verb-noun expression. We compare the reliability of the in-context and out-of-context annotations by way of the agreement between annotators.

2.1 Experimental expressions

Highly polysemous verbs, such as *give* and *take* in English and *fare* and *dare* in Italian widely participate in Verb+Noun MWEs, in which they contribute a broad range of figurative meanings that must be recognised (Fazly et al., 2007). We focus on four mostly frequent Italian verbs: *fare*, *dare*, *prendere* and *trovare*. We extract all the occurrences of these verbs when followed by any noun, from the itWaC corpus (Baroni and Kilgarriff, 2006), using SketchEngine (Kilgarriff et al., 2004). For the first experiment all the Verb+Noun types are extracted when the verb is lemmatised; and for the second experiment all the concor-

[1] Type refers to the canonical form of an expression, while token refers to each instance (usage) of the expression in any morphological form in text.

[2] http://typo.uni-konstanz.de/parseme/index.php/2-general/142-parseme-shared-task-on-automatic-detection-of-verbal-mwes

dances of these verbs when followed by a noun are generated.

2.2 Out-of-context mark-up of Verb+Noun(s)

The extraction of Verb+Noun candidates of the four verbs in focus and the removal of the expressions with frequencies lower than 20, results in a dataset of $3,375$ expressions. Two native speakers annotated every candidate expression with 1 for an MWE if the expression was idiomatic and with 0 for a non-MWE if the expression was literal. We have also defined the tag 2 for the expressions that in some contexts behave as MWEs and in others do not, e.g. *dare frutti*, which has a literal usage that means *to produce fruits* but in some contexts means *to produce results* and is an MWE in these contexts. While this out-of-context 'fast track' annotation procedure saves time and yields a long list of marked-up expressions, annotators often feel uncomfortable due to the lack of context. The information about the agreements between annotators in terms of $Kappa$ is shown in Table 2 and is compared with the in-context annotation of MWEs as explained in Section 2.3.

2.3 Annotating Verb+Noun(s) in context

We design an annotation task, in which we provide a sample of all usages of any type of Verb+Noun expression to be annotated. For this purpose, we employ the SketchEngine to list all the concordances of each verb when it is followed by a noun. Concordances include the verb in focus with almost ten words before and ten words after that. The SketchEngine reports only $100,000$ concordances for each query. Among them, we filter out the concordances that include Verb+Noun expressions with frequencies lower than 50 and we randomly select 10% of the concordances for each verb. As a result, there are $30,094$ concordances to be annotated. The two annotators annotate all usages of Verb+Noun expressions in these concordances, considering the context that the expression occurred in, marking up MWEs with 1 and expressions which are not MWEs, with 0. Table 1 reports on the details of annotation tasks and Table 2 shows the agreement details for them.

2.4 Discussion

As seen in Table 2, the inter-annotator agreement is significantly higher when annotating the expressions in context. One of the main causes of disagreements in out-of-context annotation is concerned with abstract nouns. The annotation of expressions composed of a verb followed by a noun with an abstract meaning is a more complicated process as the candidate expression may carry a figurative meaning. Each annotator uses their intuition to annotate them and it leads to random tags for these expression (e.g. *fare notizia*, *dare identità*, *prendere possesso*) when they are out-of-context. However, in the case of in-context annotation, concordances composed of abstract nouns have been annotated in the majority of cases with 1 by both annotators.

Table 1: Annotation details (A: Annotator)

Annotation task	A	tag 0	tag 1 (MWE)	tag 2
Out-of-context	1st	2,491	792	92
	2nd	2,112	1,127	136
In-context	1st	10,478	19,616	-
	2nd	9,058	21,036	-

Table 2: Inter-annotator agreement

Annotation task	Kappa	Observed Agreement
Out-of-context	0.40	0.73
In-context	**0.65**	**0.85**

In-context annotation is also very helpful for annotating expressions with both idiomatic and literal meanings. An interesting observation, reported in Table 3, is related to the number of expressions that are detected with the two different usages of idiomatic and non-idiomatic, in context.

Table 3: Statistics on the in-context annotation

	0 tagged	1 tagged	context depending
1st annotator	924	195	530
2nd annotator	696	424	529

As can be seen in Table 3,[3] among the $1,649$ types of expressions in concordances, 530 (32%) of them could be MWEs in some context and non-MWEs in others (context-depending), according to the first annotator. This annotator has annotated only 3% of the expressions with tag '2' without context.

[3] Note that the numbers in Table 3 cannot be interpreted to validate agreement between annotators, i.e. no conclusion about agreement can be derived from 3.

3 First use of the MWE resource: comparative evaluation of the automatic extraction of Italian MWEs

In our multilingual project (see Section 1) we regard the automatic translation of MWEs as a two-stage process. The first stage is the extraction of MWEs in each of the languages; the second stage is a matching procedure for the extracted MWEs in each language which proposes translation equivalents. In this study the extraction of MWEs is based on statistical association measures (AMs).

These measures have been proposed to determine the degree of compositionality, and fixedness of expressions. The more compositional or fixed expressions are, the more likely it is that they are MWEs (Evert, 2008; Bannard, 2007). According to Evert (2008), there is no ideal association measure for all purposes. We aim to evaluate AMs as a baseline approach against the annotated data which we prepared. We focus on a selection of five AMs which have been more widely discussed to be the best measures to identify MWEs. These are: MI3 (Oakes, 1998), log-likelihood (Dunning, 1993), T-score (Krenn and Evert, 2001), log-Dice (Rychlý, 2008) and Salience (Kilgarriff et al., 2004) all as defined in SketchEngine. We compare the performance of these AMs and also frequency of occurrence (Freq) as the sixth measure to rank the candidate MWEs. We evaluate the effect of these measures in ranking MWEs on both kinds of datasets.

3.1 Experiments on type-based extraction of MWEs

In the first experiment, the list of all extracted Verb + Noun combinations (as explained in Section 2.1) are ranked according to the above measures that are computed from itWaC as a reference corpus. To perform the evaluation against the list of annotated expressions, we process all 2,415 expressions for which the annotators agreed on tags 0 or 1. After ranking the expressions by the measures, we examine the retrieval performance of each measure by computing the 11-point Interpolated Average Precision (11-p IAP). This reflects the goodness of a measure in ranking the relevant items (here, MWEs) before the irrelevant ones. To this end, the interpolated precision at the 11 recall values of 0, 10%, ..., 100% is calculated. As detailed in Manning et al. (2008), the interpolated precision at a certain recall level, r, is defined as the highest precision found for any recall level $r' \geq r$. The average of these 11 points is reported as 11-p IAP in Table 4.

Table 4: 11-p IAP for ranking MWEs using different AMs

AMs	11-p IAP
Freq	0.49
MI3	0.51
log-likelihood	0.49
Salience	0.49
log-dice	0.48
T-Score	0.49

Table 5: Accuracy of AMs in classifying usages of Verb+Noun(s).

AMs	Accuracy
Freq	0.72
MI3	0.68
log-likelihood	0.72
Salience	0.69
log-dice	0.67
T-Score	0.69

As can be seen in Table 4, the selected association measures generally perform with similar performance in ranking this type of MWEs, with $MI3$ performing slightly better than others.

3.2 Experiments on token-based identification of MWEs

In the second experiment, we seek to establish the effect of these measures on identifying the usages of MWEs in our dataset of in-context annotations. We set a threshold for each score that we have computed for Verb+Noun expression types. By setting thresholds we compute the classification accuracy of the measures to identify MWEs among the usages of Verb+Noun expressions in a corpus. Specifically, each candidate of a Verb+Noun in the concordances is automatically tagged as an MWE if its lemmatised form has a score higher than the threshold, and as a non-MWE, otherwise. For each measure, we compute the arithmetic mean (average) of all the values of that measure for all expressions, and set the resulted average value as a threshold.

The accuracies of classifying the candidate Verb+Noun expressions are computed based on the human annotations of the concordances and are shown in Table 5. The classification accuracies of AMs are also very close to each other (see Table 5); however, this time $Log\text{-}likelihood$ and $Freq$ fare slightly better than others in classifying tokens of Verb+Noun expressions.

3.3 Usage-related features

Our new resource of concordances contains useful linguistic information related to usages of expressions and as such important features can be extracted from the resource to help identifying MWEs. One of these features can be obtained from the statistics of different possible inflections of the verb component of an expression. Based on the premise of the fixedness of MWEs, we expect that the verb component of a verb-noun MWE occurs only in a limited number of inflections. We implement this feature by dividing the frequency of occurrences of each expression by the number of inflections that the verb component occurs in. Note that to count the number of different inflections of the verb component, we rely on the subcorpus of concordances that we gathered.

We evaluate this approach only on 1,077 expressions that occur in concordances. We rank the expressions according to this newly computed score and we call this score, which depends on the inflection varieties, INF-VAR. For all verbs, the INF-VAR performs comparably to Frequency in ranking MWEs higher than non-MWEs, but for the verb *trovare*, we obtain better 11-p IAP using this score than by using Frequency (see Table 6).

Table 6: Performance of new scores in ranking MWEs in terms of 11-p IAP.

	total	*trovare*
Frequency	0.57	0.44
INF-VAR	0.58	**0.48**

4 Conclusions and future work

In this paper, we outline our work towards a gold-standard dataset which is tagged with Italian verb-noun MWEs along with their contexts. We show the reliability of this dataset by its considerable inter-annotator agreement compared to the moderate inter-annotator agreement on annotated verb-noun expressions presented without context. We also report the results of automatic extraction of MWEs using this dataset as a gold-standard. One of the advantages of this dataset is that it includes both 0-tagged and 1-tagged tokens of expressions and it can be used for classification and other statistical NLP approaches. In future work, we are interested in extracting context features from concordances in this resource to automatically recognise and classify the expressions that are MWEs in some contexts but not MWEs in others.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, second edition.*, pages 267–292. CRC Press.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations*, EACL ’06, pages 87–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, 44(1-2):59–77.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS*, 19(1):61–74.

Stefan Evert. 2008. Corpora and collocations. In *Corpus Linguistics. An International Handbook*, volume 2, pages 1212–1248.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. 2007. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41(1):61–89.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Sylviane Granger and Fanny Meunier. 2008. *Phraseology: an interdisciplinary perspective*. John Benjamins Publishing Company.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *EURALEX 2004*, pages 105–116, Lorient, France.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. *Proceedings of the ACL Workshop on Collocations*, pages 39–46.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on mwe resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Johanna Monti and Amalia Todirascu. 2015. Multiword units translation evaluation in machine translation: another pain in the neck? In *Proceedings of MUMTTT workshop, Corpas Pastor G, Monti J, Mitkov R, Seretan V (eds) (2015), Multi-word Units in Machine Translation and Translation Technology*.

Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan. 2013. Multi-word units in machine translation and translation technologies.

Michael P. Oakes. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Karl Pichotta and John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, WA, October.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, May. European Language Resources Association.

Alexandre Rondon, Helena Caseli, and Carlos Ramisch. 2015. Never-ending multiword expressions learning. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 45–53, Denver, Colorado, June. Association for Computational Linguistics.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *RASLAN 2008*, pages 6–9, Brno. Masarykova Univerzita.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 1:266–275.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL*, 2:193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Violeta Seretan and Eric Wehrli. 2013. Syntactic concordancing and multi-word expression detection. *International Journal of Data Mining, Modelling and Management*, 5(2):158–181.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Shiva Taslimipoor, Ruslan Mitkov, Gloria Corpas Pastor, and Afsaneh Fazly. 2016. Bilingual contexts from comparable corpora to mine for translations of collocations. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing'16. Springer.

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

Aline Villavicencio. 2005. The availability of verb–particle constructions in lexical resources: How much is enough? *Computer Speech & Language*, 19(4):415–432.