PARSEME-It Corpus An annotated Corpus of Verbal Multiword Expressions in Italian

Johanna Monti¹, Maria Pia di Buono², Federico Sangati³

³Indipendent Researcher, Italy

Abstract

English. This paper describes a new language resource annotated with verbal multiword expressions (VMWEs) in Italian. The paper discusses the state of the art in VMWE identification and annotation in Italian, the methodology adopted, the various VMWE categories annotated, the corpus and the annotation process. Finally, the paper ends with results, conclusion and future work.

Italiano. *Questo contributo descrive una nuova risorsa linguistica annotata con polirematiche verbali per la lingua italiana. Viene presentato lo stato dell'arte relativamente all'identificazione ed all'annotazione di polirematiche per la lingua italiana, la metodologia adottata, le diverse categorie di polirematiche verbali annotate nel corpus, il corpus stesso e il processo di annotazione. Infine vengono illustrati i risultati ottenuti, le conclusioni e le prospettive future.*

1 Introduction

This paper outlines the development of a new language resource for Italian, namely the **PARSEME-It VMWE corpus**, annotated with Italian MWEs of a particular class: verbal multiword expressions (VMWE). The PARSEME-It VMWE corpus has been developed by the PARSEME-IT research group¹ in the framework of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (Savary et al., 2017), a joint effort, carried out

within a European research network, to elaborate universal terminologies and annotation guidelines for verbal multiword expressions in 18 languages, among which also the Italian language is represented. Notably, multiword expressions represent a difficult lexical construction to identify, model and treat by Natural Language Processing (NLP) tools, such as parsers, machine translation engines among others, mainly due to their non-compositional property. In particular, among multiword expressions verbal ones are particularly challenging because they have different syntactic structures (prendere una decisione 'make a decision', decisioni prese precedentemente 'decisions made previously'), may be continuous and discontinuous (andare e venire versus andare in malora in Luigi ha fatto andare la società in malora), may have a literal and figurative meaning (abboccare all'amo 'bite the hook' or 'be deceived'). In this paper, we describe the state of the art in VMWE annotation and identification for the Italian language (section 2). We then present the methodology (section 3), the Italian VMWE categories taken into account for the annotation task (section 4), the corpus and the annotation process (section 5), and the results (section 6). Finally, we discuss conclusions and future work (section 7).

2 State of the art in VMWE identification and annotation in Italian

Several scholars have investigated different kinds of Italian VMWEs, focusing on both syntactic and semantic aspects. Among these works, we may distinguish contrastive and comparative analyses, and synchronic and diachronic studies.

In the first group, most of the scholars propose a comparison with Germanic languages (Mateu and Rigau, 2010), mainly for describing verb-particle constructions, that represent a very common phenomenon in this family.

On the other hand, synchronic and diachronic

¹https://www.researchgate.net/project/PARSEME-IT-Syntactic-Parsing-and-Multiword-Expressions-in-Italian

studies include analyses of: (i) verb-particle constructions (Masini, 2005; Iacobini and Masini, 2005; Quaglia and Trotzke, 2017), (ii) idiomatic constructions (Tabossi et al., 2011; Vietri, 2014c) with either ordinary or support verbs (Vietri, 2014b), (iii) support, or light, verbs, which represent a wider phenomenon and, for this reason, they have been largely analysed (La Fauci, 1980; D'Agostino and Elia, 1998; Cicalese, 1999; Alba-Salas, 2004; Quochi, 2007; Cicalese et al., 2016). Reflexive verbs in Italian have been investigated as occurrences of non-local anaphora (Reuland, 1990) and considering their syntactic classification (Carstea Romascanu, 1977).

To the best of our knowledge only a limited number of monolingual language resources with multiwords for the Italian language have been developed such as a dictionary for Italian idioms (Vietri, 2014a), a series of example corpora and a database of MWEs represented around morphosyntactic patterns (Zaninello and Nissim, 2010), or a corpus annotated with Italian MWEs of a particular class: verb-noun expressions such as *fare riferimento*, *dare luogo* and *prendere atto* (Taslimipoor et al., 2016). At the time of writing, therefore, the PARSEME-It VMWE corpus represents the first sample of a corpus, which includes several types of VMWEs, specifically developed for NLP applications.

3 Methodology

The development of the Italian VMWE corpus is based on the PARSEME annotation guidelines², provided for the shared task. The guidelines have been developed with the aim of delivering general definitions and prescriptions for the annotation of VMWEs in 18 languages, but, at the same time, of allowing language-specific descriptions of these linguistic phenomena (Savary et al., 2017). The annotation guidelines include three main categories:

- a universal category, which is common to all the languages involved in the task and holds light-verb constructions (LVCs) and idioms (ID);
- 2. a **quasi-universal category**, relevant for some languages or language families, that

contains inherently reflexive verbs (IReflVs) and verb-particle constructions (VPCs);

3. an **other VMWEs** category, which is a residual category for the occurrences not belonging to any of the previous groups.

In order to ease the identification and categorisation task of VMWEs, a decision tree method was devised with generic and language-specific tests. Generic tests consider general criteria that are valid for all languages, while language-specific tests consider structural, lexical, morphological and syntactic features that are specific for the individual languages. The decision tree includes three steps, (i) identification of a VMWE candidate, i.e., a combination of a verb with at least one other word, which is a potential VMWE; (ii) identification of the lexicalized elements of the expression, (iii) assignment of the VMWE to one of the VMWE categories, using general and languagespecific tests.

4 Italian VMWEs

For the Italian VMWE annotation task, according to PARSEME guidelines, multiword expressions are understood as (continuous or discontinuous) sequences of words with the following compulsory properties:

- Their component words include a head word and at least one other syntactically related word. Most often the relation they maintain is a syntactic (direct or indirect) dependency but it can also be e.g., a coordination.
- They show some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language.
- At least two components of such a word sequence have to be lexicalized.

In this task we only annotate the lexicalized components and ignore open slots. Collocations, i.e., word co-occurrences whose idiosyncrasy is of statistical nature only (e.g., *the graphic shows, drastically drop*, etc.), are excluded from the scope of this study. The VMWE which have been annotated for the Italian language are:

1. **Light verb constructions** (LVC), which typically consist of a verb and a noun or prepositional phrase, e.g., *fare una domanda* ('to

²The guidelines are available at http://parsemefr.lif.univmrs. fr/guidelines-hypertext/.

make a question'), *fare una passeggiata* ('to have a walk'). The verb has a purely syntactic operator function (performing an activity or being in a state), whereas the noun is predicative, often referring to an event (e.g., decision, visit) or a state (e.g., fear, courage);

- 2. **Idioms** (ID), which have at least two lexicalized components including a head verb and at least one of its arguments, e.g., *tirare le cuoia* ('kick the bucket'), *piovere a catinelle* ('rain cats and dogs');
- 3. **Inherently reflexive verbs** (IReflV), which are those reflexive verbal constructions which (a) never occur without the clitic e.g., *suicidarsi* ('suicide'), or when (b) the REFLV and non-reflexive versions have clearly different senses or subcategorization frames e.g., *riferirsi* ('refer');
- 4. Verb particle combinations (VPC), which are formed by a lexicalized head verb and a lexicalized particle dependent on the verb. The meaning of the VPC is noncompositional. Notably, the change in the meaning of the verb goes significantly beyond adding the meaning of the particle, e.g., *buttare giù* ('swallow'). This type of construction is very frequent in English, German, Swedish, Hungarian, but we can find them also in Italian;
- 5. Other Verbal MWEs (OTH), which gather the types not belonging to any of the categories above, e.g., *corto-circuitare* ('shortcircuit').

5 Corpus and annotation task

5.1 PARSEME Italian VMWE corpus

The PARSEME-It VMWE corpus is based on a selection of texts taken from the *PAISA* ' corpus of Italian web texts (Lyding et al., 2014). We chose this corpus because it contains documents (i) from different web sources, e.g., Wikibooks, Wikinews, Wikiversity, and several blog services from different websites, collected in 2010 by means of a Creative Commons-focused web crawling, and a targeted collection of documents from specific websites, (ii) dedicated to no specific technical domain, free from copyright issues, so as to be compatible with an open license (iii) annotated in

CoNLL format, i.e. lemmatized, POS-tagged and annotated with syntactic dependencies. For our annotation task, we selected a sub-corpus formed by 17,000 sentences (corresponding to 421,848 tokens) randomly taken from blogs, Wikipedia and Wikinews. The corpus was kept in its original state and therefore no errors or inconsistencies were corrected. The pre-annotation of the PAISA' was kept in order to ease the annotation work with reference to the identification of verbal MWEs but we asked annotators not to overestimate the system's performances, and to review the whole text, not only the pre-annotated candidates proposed by the system. A dedicated tag in FLAT was defined for this purpose. The objective was to have a final corpus of at least 3,500 annotated VMWEs per language. Since the density of VMWEs highly depend on the particular language, as well as text choice and genre, we were not able to make any reliable estimation of the corpus size needed to reach this goal from the beginning of the task.

5.2 Annotation environment

The annotation environment used for the PARSEME-It VMWE corpus is FLAT, a webbased linguistic annotation environment³ based around the FoLiA format⁴ a rich XML-based format for linguistic annotation. FLAT allows users to view annotated FoLiA documents and enrich these documents with new annotations (Figure 1), a wide variety of linguistic annotation types is supported through the FoLiA paradigm. It is a document-centric tool that fully preserves and visualises document structure. It is open source software developed at the Centre of Language and Speech Technology, Radboud University Nijmegen and is licensed under the GNU Public License v3.

5.3 Annotation task

The annotation task for the Italian language was performed in five different stages.

1. The PARSEME Annotation guidelines were agreed on⁵ and examples for the Italian language were added in order to ease the annotation task by the Italian annotators. To this end, a two-phase pilot annotation in Italian

³http://flat.readthedocs.io/en/latest/

⁴http://proycon.github.io/folia

⁵http://parsemefr.lif.univ-mrs.fr/parseme-stguidelines/1.0/?page=home

Figure 1: Example of annotated data in FLAT

was carried out. This step was useful in identifying the Italian VMWE categories to be annotated, but also to promote cross-language convergences with the other languages foreseen in the shared task. Each pilot annotation phase provided feedback from annotators and was followed by enhancements of the guidelines, corpus format and processing tools.

2. A pre-processing step of the *PAISA* ' corpus was needed: a 'no space' column was added to the files in order to add the 'nsp' tag if a token should have been appended to the pre-vious one without a space.

Α	В	С
Rank	Token	No space (nsp)
1	Ľ	nsp
2	espresso	
3	è	
4	spesso	
5	sinonimo	
6	di	
7	una	
8	sosta	
9	veloce	nsp
10	,	

Figure 2: Example of the use of an nsp tag

3. The annotation task of the training set (approx. 16,000 sentences) was manually performed in running texts using the FLAT environment by five Italian native speakers with linguistic background. Each annotator was given a certain number of files, containing 1,000 sentences in CoNLL format. All the doubts about the annotation were collected in a shared file and discussed during the annotation phase. Difficulties in annotating VMWE mainly concerned (i) the boundaries of the VMWE such as in *Sei ovviamente nel pieno diritto di esprimere [...]* where it is difficult to decide if the VMWE should be *sei ... nel ... diritto* or *sei ... nel pieno diritto*,

(ii) the category attribution concerning for instance the *fare* + N VMWE type, since in some cases the category is LVC such as in *fare rumore* and in some others is ID such as in *fare schifo*, (iii) the identification of nested VMWEs like in *Mi guardo bene* where the annotator has to decide if in the ID *guardarsi bene* there is also a IRefIV *guardarsi* or not.

- 4. A few files were double-annotated to evaluate the inter-annotator agreement (IAA). Measuring IAA is not a trivial task because of the challenges posed by VMWEs and described in the Introduction. The available IAA results organized per-VMWE F-score (F_{unit}), estimated Cohens K (K_{unit}) and finally standard K(K_{cat}) (Savary et al., 2017) scores are presented in Table 1.
- 5. Further 1,000 sentences were used as test-set during the shared task. The VMWE annotations were automatically annotated by the systems that took part in the shared task and performed according to the same guidelines.

	#S	#T	#A1	#A ₂	Funit	Kunit	Kcat
IT	2000	52639	336	316	0.417	0.331	0.78

Table 1: AA scores for Italian annotation: #S, and #T show the number of sentences and tokens in the corpora used for measuring the IAA, respectively. $\#A_1$ and $\#A_2$ refer to the number of VMWE instances annotated by each of the annotators (Savary et al., 2017).

6 Results

The PARSEME-It VMWE corpus is composed of 2,454 entries (Table 2), and it is freely available⁶, released under Creative Commons licenses.

The data have been annotated using the official parseme-tsv format⁷ (Figure 3), adapted from the CoNLL format.

⁶http://hdl.handle.net/11372/LRT-2282

⁷http://typo.uni-konstanz. de/parseme/index.php/2general/ 184-parseme-shared-task-format-of-the-finalannotation.

Category	Occurrences
ID	1163
IReflV	730
LVC	482
VPC	73
OTH	6
Total	2454

Table 2: Overview of VMWEs in the PARSEME-It VMWE corpus, including train and test sets.

1	In	_	_	
2	prossim	ità		
3	della		-	_
4	tornata	-	-	
5	alattar		-	
5	etettoi	ale	_	-
0	per	-	-	
/	la	_	_	
8	rielezi	one	_	_
9	delle	_	_	
10	cariche	_	_	
11	di	_	_	
12	assesso	ri	_	
13	alla	· -	-	_
14	Dogiono	-	-	
14	Keytone	_	-	
15	veneto	_	_	
16	qualcun	0	_	_
17	vuole	_	_	
18	far-	_	1:ID	
19	ali			
20	le	_	1	
21	scarne	 nsn	1	
22	2	пэр	-	
22		_	_	

Figure 3: Example of annotated data in parsemetsv format

In the official parseme-tsv format, as described in Savary et al. (2017), the information about each token are represented by 4 tab-separated columns featuring (i) the position of the token in the sentence or a range of positions (e.g., 1-2) in case of multiword tokens such as contractions, (ii) the token surface form, (iii) an optional flag indicating that the current token is adjacent to the next one, and (iv) an optional VMWE code composed of the VMWEs consecutive number in the sentence and for the initial token in a VMWE its category (e.g., 2:ID if a token starts an idiom which is the second VMWE in the current sentence). In case of nested, coordinated or overlapping VMWEs multiple codes are separated with a semicolon. Furthermore, in order to provide data usable as features in the shared task systems, also companion files in a format close to CoNLL-U⁸ have been released. These companion files contain extra linguistic information, i.e., lemmas, POS-tags, morphological features, and syntactic dependencies.

7 Conclusion and Future Work

In this paper, we described a linguist resource of Italian VMWE, developed within the PARSEME Shared Task on Automatic Identification of VMWE. We consider this work an initial contribution for elaborating an Italian universal terminology of VMWE. Future work includes the extension of the current corpus and a fine-grained linguistic analysis of the annotation in order to contribute to the description of these phenomena.

Acknowledgments

The work described in this paper has been supported by the IC1207 PARSEME COST action. The annotation work was carried out also thanks to the help of Maarten van Gompel who adapted the FLAT annotation platform to the needs of the community.

Our thanks go also to the Italian annotators, Valeria Caruso, Manuela Cherchi, Anna De Santis, Annalisa Raffone, for their contributions.

Autorship contribution is as follows: Johanna Monti is author of Sections 1, 3, 4 and 5.3; Maria Pia di Buono of Sections 2 and 6 and 7; and Federico Sangati of Sections 5.1 and 5.2.

References

- Josep Alba-Salas. 2004. Fare light verb constructions and italian causatives: Understanding the differences. *ITALIAN JOURNAL OF LINGUISTICS*, 16(2):283.
- M Carstea Romascanu. 1977. I tipi di verbi riflessivi in italiano. *Revue Roumaine de Linguistique Bucuresti*, 22(2):125–130.
- Anna Cicalese, Emilio D'Agostino, Alberto Maria Langella, and Ilaria Villari. 2016. Els verbs locatius com a variants de verbs de suport. *Quaderns d'Italià*, 21:153–166.
- Anna Cicalese. 1999. Le estensioni di verbo supporto. uno studio introduttivo. *Studi italiani di linguistica teorica ed applicata*, 28(3):447–485.
- Emilio D'Agostino and Annibale Elia. 1998. Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. AA. VV, Ai limiti del linguaggio. Bari: Laterza, pages 287–310.

⁸http://universaldependencies.org/format.htm

- Claudio Iacobini and Francesca Masini. 2005. Verbparticle constructions and prefixed verbs in italian: typology, diachrony and semantics. In *Mediterranean Morphology Meetings*, volume 5, pages 157–184.
- Nunzio La Fauci. 1980. Aspects du mouvement de wh, verbes supports, double analyse, complétives au subjonctif en italien: pour une description compacte. *Lingvisticae Investigationes*, 4(2):293–341.
- Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice DellOrletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisa corpus of italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43.
- Francesca Masini. 2005. Multi-word expressions between syntax and the lexicon: the case of italian verb-particle constructions. SKY Journal of Linguistics, 18(2005):145–173.
- Jaume Mateu and Gemma Rigau. 2010. Verb-particle constructions in romance: A lexical-syntactic account. *Probus*, 22(2):241–269.
- Stefano Quaglia and Andreas Trotzke. 2017. Italian verb particles and clausal positions. In *IATL 31: The 31st annual meeting Israel Association for Theoretical Linguistics*, pages 67–82.
- Valeria Quochi. 2007. A usage-based approach to light verb constructions in italian: Development and use.
- Eric Reuland. 1990. Reflexives and beyond: Non-local anaphora in italian revisited. *Grammar in progress: glow essays for Henk van Riemsdijk*, 36:351.
- Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemizadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop* on Multiword Expressions (MWE 2017), pages 31– 47.
- Patrizia Tabossi, Lisa Arduino, and Rachele Fanari. 2011. Descriptive norms for 245 italian idiomatic expressions. *Behavior Research Methods*, 43(1):110–123.
- Shiva Taslimipoor, Anna Desantis, Manuela Cherchi, Ruslan Mitkov, and Johanna Monti. 2016. Language resources for italian: towards the development of a corpus of annotated italian multiword expressions. In Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016). ceur-ws.
- Simona Vietri. 2014a. The italian module for nooj. In *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it*, pages 389–393.

- Simonetta Vietri. 2014b. *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*, volume 31. John Benjamins Publishing Company.
- Simonetta Vietri. 2014c. The lexicon-grammar of italian idioms. In Workshop on Lexical and Grammatical Resources for Language Processing, COLING 2014, pages 137–146.
- Andrea Zaninello and Malvina Nissim. 2010. Creation of lexical resources for a characterisation of multiword expressions in italian. In *LREC*.