PARSEME-IT
Issues in verbal Multiword Expressions identification and classification

Johanna Monti[1], Valeria Caruso[1], Maria Pia di Buono[2]

[1]Dep. of Literary, Linguistic and Comparative Studies "L'Orientale" University of Naples, Italy
[2]TakeLab - Faculty of Electrical Engineering and Computing - University of Zagreb, Croatia
`jmonti@unior.it, vcaruso@unior.it, mariapia.dibuono@fer.hr`

Abstract

English. The second edition of the PARSEME shared task was based on new guidelines and methodologies that particularly concerned the Italian language with the introduction of new categories of verbs not considered in the previous edition. This contribution presents the novelties introduced, the results obtained and the problems that emerged during the annotation process and concerning some categories of verbs.

Italiano. *La seconda edizione del PARSEME shared task si è basata su nuove linee guida e metodologie che hanno riguardato in particolar modo la lingua italiana con l'introduzione di nuove categorie di verbi non considerate nella precedente edizione. Il contributo presenta le novità introdotte, i risultati ottenuti e le problematiche che sono emerse durante l'annotazione relativamente ad alcune categorie di verbi.*

1 Introduction

The paper reports on some final results of the second edition of an annotation trial for verbal Multiword Expressions (VMWEs) carried out on the Italian language by the PARSEME-IT research group [1], which started within the broader European PARSEME project, the IC1207 COST action ended in April 2017[2].

The initial project is expanding in this second stage of its development, thanks to a wider network of research groups, working together as one

of the ACL Special Interest Group on the Lexicon, called SIGLEX-MWE.

In its first edition, the PARSEME shared task released a corpus of 5.5 million tokens and 60,000 VMWE annotations in 18 different languages which is distributed under different versions of the Creative Commons license. To increase the computational efficiency of Natural Language Processing (NLP) applications, PARSEME focuses on a special class of Multiword Expressions which have been seldom modelled for their challenging nature, such as verbal MWEs (Savary et al., 2017).

Many of the features of this particular type of MWE are considered to be difficult to cope with, such as the discontinuity they present (*turn it off*) the syntactic variations they license (*the decision was hard to take*), the semantic variability resulting both in literal and idiomatic readings (*to take the cake*), or the syntactic ambiguity of many forms (*on* is a preposition in *to trust on somebody*, but a particle in *to take on the task*). Moreover, these units have language-specific features, and are generally modelled according to descriptive categories developed by different traditions of linguistic studies. The PARSEME research group thus addresses also the creation of a multilingual common platform for VMVEs using universal terminology, guidelines and methodologies for the identification of these units cross-linguistically. Moreover, at the end of the first annotation trial a shared task on automatic identification of VMWEs was also carried out and has proved the reliability and usefulness of the data collected so far, which have been already presented and discussed (Savary et al., 2017; Monti et al., 2017).

The paper illustrates the types of VMWEs used by the second PARSEME annotation trial more thoroughly. In Section 2 we provide a brief description of the second annotation trial of the PARSEME shared task together with the statistics. Then we present a new category of verbal MWEs,

---

[1]`https://sites.google.com/view/parseme-it/home`
[2]`https://typo.uni-konstanz.de/parseme/`

namely Inherently Clitic Verbs (Section 3) and in Section 4 two very productive categories in Italian (IRV and IDV). In Section 5, we discuss some borderline cases which posed some classification issues. Finally, we conclude and discuss future work.

2 PARSEME Shared Task Second annotation trial: a brief report

This section focuses on the novelties which have been introduced in the guidelines and methodologies used for the second annotation trial in order to cover a wider range of VMWEs which were left apart in the first stage of the project. The improvements seem to be particularly valuable for the data collection carried out on the Italian language, because they address some peculiarities of the Italian language which were not considered in the first edition of the shared task but have been taken into account in the second edition, namely:

- **Inherently clitic verbs (ICV)**, which is an extremely rich and varied VMWE category in Italian (Masini, 2015). As described in Section 3, a language specific category was created for the Italian language (LS.ICV) which takes into account only those verbs whose semantics is changed by a non-reflexive clitic pronoun, like *entrarci* when it means *to be relevant to something*, while the intranstive form of the verb *entrare* means *to enter*.

- **Inherently adpositional verbs (IAV)**, a high frequency category of VMWEs, namely those verbs whose meanings are significantly affected by an "idiomatic selected preposition", like *su* in *contare su qualcuno* (to rely on someone): without the preposition the verb means only *to determine the total number of something*. These verbs are often called *prepositional verbs*[3].

- **Multi verb constructions (MVC)**, VMWEs composed by a sequence of two adjacent verbs (in a language-dependent order), a governing verb V_gov (also called a vector verb) and a dependent verb V_dep (also called a polar verb), like in *lasciar perdere* (to give up).

---

[3]Schneider, N., Green, M., 2015, New Guidelines for Annotating Prepositional Verbs, https://github.com/nschneid/nanni/wiki/Prepositional-Verb-Annotation-Guidelines

The other classifying categories used are (a) **light verb constructions (LVCs)**, e.g. *fare una passeggiata* (to have a walk), and (b) **idioms (ID)**, e.g., *tirare le cuoia* (to kick the bucket), considered to be universal categories or categories which can be found in all languages participating in the task.

Other VMWEs are instead maintained as quasi-universal categories, since their range of application seems to cover only some language groups or languages, but not all. They are (c) **inherently reflexive verbs (IReflVs)**, and (d) **verb-particle constructions (VPCs)**. The first group (IReflVs) allows annotators to account for verbs which are never used without a reflexive clitic pronoun, e.g., (Italian) *suicidarsi* (to suicide), or for those verbs whose meaning is significantly affected by the pronoun, e.g., (Italian) *farsi* (to take drugs) while the non-pronominal form, *fare*, means *to make*. Semantic aspects are also used to identify Verb-particle constructions (VPC) because their meaning is fully non-compositional, e.g., *buttare giù* (to swallow), or only partly non-compositional, like in *tirare avanti* (to go on) since the preposition no longer owns its spatial meaning.

Table 1 presents the statistics of the various categories of VMWEs in the PARSEME-IT corpus 1.1.

3 A language specific category: Inherently clitic verbs (LS.ICV)

Inherently Clitic Verbs (LS.ICV) represent a specific category for some Romance languages, and they are particularly frequent in the Italian language. It is often challenging to distinguish LS.ICV from Inherently Reflexive Verbs (IRV), particularly because some clitics may be ambiguous, like *se/si* which is a polyfunctional clitic pronoun and grammatical marker (and can have a reflexive, reciprocal, impersonal, passivizing, aspectual, and middle function). LS.ICVs together with IRVs are pronominal verbs. LS.ICV are formed by a full verb combined with one or more non-reflexive clitic that represents the pronominalization of one or more complement (CLI).

The following verbs should be annotated as LS.ICV:

- The verb without the CLI does not exist, e.g., *infischiarsene* (do not worry about) vs *infischiare*;

	sent.	tokens	VMWEs	IAV	IRV	LS.ICV	LVC.cause/full	MVC	VID	VPC.full/semi
IT-dev	917	32613	500	44	106	9	19/100	6	197	17/2
IT-train	13555	360883	3254	414	942	20	147/544	23	1098	66/0

Table 1: PARSEME-IT corpus version 1.1

- The verb without the CLI does exist, but has a very different meaning as in *prenderle* (gl.: to take them, transl. to be beaten) vs *prendere* (to take) or *prenderci* (gl.: to take it, transl. to grasp the truth) vs *prendere* (to take);

- The verb has more than one CLI of which the second one is an invariable object complement, like in *fregarsene* (gl.: matter self of-it, transl. do not care about) or *infischiarsene* (do not worry about);

- The verb has two non-reflexive invariable CLIs, like in *farcela* (gl.: to make there it, transl. to succeed);

- The verb has a different meaning with respect to an intensive use of the same two non-reflexive invariable CLIs, like in *andarsene* (gl.: to go away self from-there, transl. to die) vs *andarsene* (to go away) or *bersela* (gl.: drink self it, transl. to believe) vs *bersela* (to drink it).

The annotation of LS.ICV was performed following a specific decision tree [4].

In the training corpus 20 different LS.ICV were annotated manually, such as *farcela*, *rimetterci*, *fregarsene* among others.

4 Very productive VMWEs: IRVs and VID

IRVs and VID represent very productive categories in Italian which pose some classifying issues due to their specific characteristics.

With reference to IRVs, the presence of the clitic pronoun *si* may generate ambiguity in the annotation process, as in Italian it refers to three different types of construction: i) reflexive, ii) impersonal, iii) inherent.

In order to distinguish these cases, we consider that in the reflexive construction, the clitic pronoun can be paraphrased by means of either an

anaphoric expression which stands for *se stesso* (oneself) or a mutual expression which refers to *gli uni e gli altri* (these and those). Another relevant aspect to consider in the classification of IRVs is the presence of an implicit thematic role due to the fact that the action includes two different entities with different thematic properties but with the same reference, e.g., in *guardarsi* (to look at oneself) the clitic signals the presence of coreference between the first argument and the second one. Another source of mis-classification of IRVs is related to the presence of anticausative constructions. In these constructions, the clitic may represent an overt marker of reduced transitivity, , e.g., *sedersi* (to sit down).

In some cases, IRVs occur in idiomatic construction and their meaning is affected by the presence of new elements, such as in *guardarsi bene da* (to be careful not to). Consequently the annotation of such occurrences is subjected to the evaluation of characteristics related to VID, as the low variability, the presence of semantic non-compositional meaning, and the literal-idiomatic ambiguity. In the VID class, the non-compositionality property is prototypical such as in *battersi all'ultimo sangue* (lit. to fight till the last blood) which means *to fight to the last*. Despite their meaning is opaque, sometimes VID may have both a literal and idiomatic meaning and the boundaries between them are difficult to trace. For example, *avere gli occhi bendati* (lit. to have the eyes covered) has both a literal meaning and an idiomatic one and in this latter case it should be translated in English as *to be blindfold*. According to Vietri (2014b), it is possible to classify ordinary-verb VID, namely VID which present a semantically full verb, on the basis of their definitional structure, identified by means of the arguments required by the operators. In the case of VID, the operator consists of the verb and the fixed element(s), while the argument may be the subject and/or a free complement. VIDs can be formed also by constructions based on the use of support verbs, namely *avere* (to have), e.g., *avere fegato* (lit. to have leaver, transl. to have guts) *essere*

(to be), e.g., *essere a cavallo* (to be golden) and *fare* (to make), e.g., *fare lo gnorri* (to play fool). The main difference between this class of VID and the one formed by ordinary verbs is that support verbs are semantically empty, and for this reason this class of VID presents a high degree of lexical and syntactic variability. This type of variability is retrievable in aspectual variants, the production of causative constructions, the possible deletion of the support verb which causes complex nominalizations (Vietri, 2014a).

5 Borderline cases: LVC and IAV compared

In this section we discuss the novelties concerning two categories used in the second edition of the PARSEME shared identification task of verbal MWEs (edition 1.1), namely LVC and IAV. As regards LVC, two new subcategories have been introduced in the second edition, LVC.full and LVC.cause, to account for a more fine-grained distinction between LVCs, where the verb is semantically totally bleached (e.g., *to have the right*), and those where the verb adds a causative meaning (and a new semantic role) to the noun (e.g., *to grant the right*). Therefore some new tests have been added to account for these subcategories, which heavily rely on the notion of semantic arguments.

In particular, constructions annotated as LVC.cause may involve: i) verbs that are typically used to express the cause of predicative nouns in general (e.g., *cause, provoke*), ii) verbs that are only used to express the cause of particular predicative nouns (e.g., *grant* in *to grant a right*).

IAV consists of a verb or VMWE and an idiomatic selected preposition or postposition that is either always required or, if absent, changes the meaning of the verb of the VMWE significantly. IAVs are verb+adposition combinations in which: i) the dependents of the adposition are not lexicalized, or ii) the adposition cannot be omitted without markedly altering the meaning of the verb. During the annotation trial, the IAV category has proved to be advantageous to cover the rich inventory of VMWEs in Italian, but some issues have also emerged, particularly with respect to the other class of LVC verbs, which also accounts for combinations of verbs plus prepositions. Prototypical examples of IAV collected so far include the following:

1.a *Tendere a* + N (to be inclined to something), base form *tendere* (to stretch), e.g., *Maria tende alla depressione* (Maria tends to be depressed);

1.b *Tendere a* + V (to be inclined to something), e.g., *Maria tende a dimagrire* (Maria tends to loose weight);

2. *Puntare su* + N (to bet), base form *puntare* (to stick), e.g., *puntare su qualcuno/qualcosa*.

These examples exhibit clear semantic changes from the non-adpositional base form of the verb; moreover, the preposition can not be omitted in questions, thus proving to be part of the verb:

- *Maria tende sempre ad esagerare.*

- *A cosa tende, scusa?*

Less prototypical IAV examples include verb instances exhibiting semantic changes pivoted by the arguments they combine with, like *andare in* (both *to go to* and *to become*), or *sapere di* (*to smell* and *to know about*). The type of semantic interaction at stake, called *co-composition* in the Generative Lexicon[5], is realized when "the complements carry information which acts on the governing verb, essentially taking the verb as argument and shifting its event type" (Pustejovsky, 1995). For example, *andare in* denotes directed motion when combined with proper or common place nouns like in *andare in città/montagna/America*, (to go to the city/mountain/America); or the medium of motion, when combined with vehicles names, like in *vado in bici/Ferrari*, (I ride my bike/drive my Ferrari). However, with nouns denoting *states*, like *andare in estasi* (to become absorbed) or *andare in panico* (to start feel panic), the verb acquires the aspectual meaning of *to go into the state X*, and can not be classified as an LVC. With names referring to events, instead, like *andare in soccorso* (lit. to go in assistance), the original spatial semantics bleaches by interacting with the name meaning: actually *to go into the event X* denotes the action expressed by the predicative name and can be classified as an LVC.

---

[5]Co-composition has been called '*accommodation* in more recent works (Pustejovsky, 2013).

6 Conclusions and Future Work

In this paper we described the novelties concerning the PARSEME shared task on automatic identification of verbal MWEs - edition 1.1 (2018), in which new verb categories have been included in comparison with the 2017 edition. Some of them are language-specific, such as ICV for some Romance languages, others are not, like IAV. The increased number of categories enables to annotate corpus data more thoroughly, and discover a broad range of combinatorial phenomena that present different degrees of opacity.

We also discussed two productive categories in Italian, namely IRV and VID, and analyzed LVC and IAV borderline cases together with observations on combinatorial phenomena that can be applied in order to annotate VMWE more effectively.

Future work includes a further linguistic analysis of borderline cases in order to contribute to the description of these phenomena.

References

Francesca Masini. 2015. Idiomatic verb-clitic constructions: lexicalization and productivity. In *Mediterranean Morphology Meetings*, volume 9, pages 88–104.

Johanna Monti, Maria Pia di Buono, and Federico Sangati. 2017. Parseme-it corpus an annotated corpus of verbal multiword expressions in italian. In *Fourth Italian Conference on Computational Linguistics-CLiC-it 2017*, pages 228–233. Accademia University Press.

James Pustejovsky. 1995. *The generative lexicon*. MIT Press.

James Pustejovsky. 2013. Type theory and lexical decomposition. In *Advances in generative lexicon theory*, pages 9–38. Springer.

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.

Simonetta Vietri. 2014a. *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*, volume 31. John Benjamins Publishing Company.

Simonetta Vietri. 2014b. The lexicon-grammar of italian idioms. In *Workshop on Lexical and Grammatical Resources for Language Processing, COLING 2014*, pages 137–146.