IJCoL
Italian Journal of Computational Linguistics

5-2 | 2019
Further Topics Emerging at the Fifth Italian Conference on Computational Linguistics

PARSEME-It: an Italian corpus annotated with verbal multiword expressions

Johanna Monti and Maria Pia di Buono

Electronic version
URL: http://journals.openedition.org/ijcol/483
DOI: 10.4000/ijcol.483
ISSN: 2499-4553

Publisher
Accademia University Press

Printed version
Number of pages: 61-93

Electronic reference
Johanna Monti and Maria Pia di Buono, "PARSEME-It: an Italian corpus annotated with verbal multiword expressions", *IJCoL* [Online], 5-2 | 2019, Online since 01 December 2019, connection on 28 January 2021. URL: http://journals.openedition.org/ijcol/483 ; DOI: https://doi.org/10.4000/ijcol.483

PARSEME-It: an Italian corpus annotated with verbal multiword expressions

Johanna Monti*
Università degli Studi di Napoli
L'Orientale

Maria Pia di Buono**
Università degli Studi di Napoli
L'Orientale

The paper describes the PARSEME-It corpus, developed within the PARSEME-It project which aims at the development of methods, tools and resources for multiword expressions (MWE) processing for the Italian language. The project is a spin-off of a larger multilingual project for more than 20 languages from several language families, namely the PARSEME COST Action. The first phase of the project was devoted to verbal multiword expressions (VMWEs). They are a particularly interesting lexical phenomenon because of frequent discontinuity and long-distance dependency. Besides they are very challenging for deep parsing and other Natural Language Processing (NLP) tasks. Notably, MWEs are pervasive in natural languages but are particularly difficult to be handled by NLP tools because of their characteristics and idiomaticity. They pose many challenges to their correct identification and processing: they are a linguistic phenomenon on the edge between lexicon and grammar, their meaning is not simply the addition of the meanings of the single constituents of the MWEs and they are ambiguous since in several cases their reading can be literal or idiomatic. Although several studies have been devoted to this topic, to the best of our knowledge, our study is the first attempt to provide a general framework for the identification of VMWEs in running texts and a comprehensive corpus for the Italian language.

1. Introduction

Multiword expressions (MWEs) represent a difficult lexical construction to identify, model and treat in Natural Language Processing (NLP) tasks, e.g., parsing (Constant, Sigogne, and Watrin 2012), machine translation (Venkatapathy and Joshi 2006; Monti et al. 2013; Mitkov et al. 2018) and keyphrase extraction (Newman et al. 2012), mainly due to their non-compositional property. The lack of compositionality, which concerns the lexical, morphological, syntactic, semantic, pragmatic and statistical level of analysis, namely, (Baldwin 2006), characterizes the behaviour of such linguistic phenomena.

Different types of lexical constructions can be classified as MWEs, with different levels of representation in each language based on their frequency and language-specificity (Salehi, Cook, and Baldwin 2016), e.g., compound nouns are very common in languages such as English (Copestake 2003; Ó Séaghdha 2008) and German (Im Walde, Müller, and Roller 2013), light verb constructions (LVCs) in English (Butt 2010), Persian (Karimi-Doostan 1997), and Italian (Alba-Salas 2002).

* UNIOR NLP Research Group, Dept. of Literary, Linguistic and Comparative Studies - Via Duomo, 219 80138 Napoli, Italy. E-mail: jmonti@unior.it
** UNIOR NLP Research Group, Dept. of Literary, Linguistic and Comparative Studies - Via Duomo, 219 80138 Napoli, Italy. E-mail: mpdibuono@unior.it

Scholars usually do not converge on a unique definition and classification of MWEs nor include in their classifications the same types of MWEs. For our study we refer to the definition of MWEs as "lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity[1]" (Baldwin and Kim 2010).

Among these types, verbal multiword expressions (VMWEs) are particularly challenging and, as we will discuss in the next sections, a fine-grained classification is needed. They may present different syntactic structures, e.g., *prendere una decisione* (to make a decision), *decisioni prese precedentemente* (decisions made previously), may be continuous and discontinuous, e.g., *andare e venire* (to come and go) versus *andare in malora* (go to ruin) in *Luigi ha fatto **andare** la società **in malora*** (Luigi made the company go ruin), may have a literal and figurative meaning, e.g., *abboccare all'amo* (to take the bait). Moreover, these units have language-specific features and are generally modelled according to descriptive categories developed by different traditions of linguistic studies.

In this paper, we describe the PARSEME-It VMWE corpus[2], which represents the main outcome of the PARSEME-It project[3], a spin-off project of the European IC1207 COST[4] action PARSEME[5], carried out by the UNIOR NLP Research Group[6]. The main aim of the project is i) to bridge the gap between linguistic precision and computational efficiency in NLP applications by investigating the syntactic and semantic representation of MWEs in language resources and ii) the integration of MWE analysis in syntactic parsing and translation technology. Deliverables include mainly enhanced monolingual language resources (lexicons, grammars and annotated corpora) in Italian or multilingual linguistic resources with the Italian language. The UNIOR NLP Research group, together with the language leaders working on other languages, has contributed to developing the general and language-specific guidelines for the PARSEME annotation process.

We discuss related researches in linguistic studies on VMWEs and more in general in MWE processing, including a description of the PARSEME COST Action and its aims (Section 2). Then, the PARSEME-It corpus (Section 3) is introduced. In Section 4, we present the VMWE categories included in the annotation scheme and in Section 5 the annotation guidelines, the identification tests and decision trees used. The description of the annotation process (Section 6) and annotation issues (Section 7), the analysis of productive categories and borderline cases (Section 8) follow. Finally, we discuss conclusions and future work (Section 9).

2. Related Work

As a diverse and complex phenomenon present in all natural languages (Jackendoff 1997; Sag et al. 2002), MWEs have attracted the interest of many disciplines.

1 As defined by Lyse and Andersen (2012), "statistical idiomacity is the phenomenon of particular combinations of words occurring with markedly higher frequency in comparison to alternative phrasings of the same concept".
2 https://github.com/UNIORNLP/PARSEME-It-Corpus
3 https://sites.google.com/view/parseme-it/home
4 https://www.cost.eu
5 https://typo.uni-konstanz.de/parseme
6 https://sites.google.com/view/unior-nlp-research-group/home?authuser=0

Recently Constant et al. (2017) proposed a classification including MWE categories which are non-exhaustive and may overlap:

- **Idiom**: a group of lexemes whose meaning is established by convention and cannot be deduced from the individual lexemes composing the expression (e.g., *tirare le cuoia* → to kick the bucket).
- **Light-verb construction**: it is formed by a head verb with light semantics that becomes fully specified when combined with a (directly or indirectly) dependent predicative noun (e.g., *fare una passeggiata* → to take a walk).
- **Verb-particle construction**: it comprises a verb and a particle, usually a preposition or adverb, which modifies the meaning of the verb and which needs not be immediately adjacent to it (e.g., *buttare giù* → to swallow). Verb-particle constructions are also referred to as *phrasal verbs*.
- **Compound**: a lexeme formed by the juxtaposition of adjacent lexemes, occasionally with morphological adjustments (e.g., *carta di credito* → credit card). Compounds can be subdivided according to their syntactic function. Thus, nominal compounds are headed by a noun (e.g., *lettera aperta* → open letter) whereas noun compounds and verb compounds are concatenations of nouns (e.g., *treno merci* → freight train) or verbs (e.g., *lasciar andare* → let go). Some authors (Stymne, Cancedda, and Ahrenberg 2013; Shapiro 2016; Gagné and Spalding 2009) refer to closed compounds when they are composed of a single token (e.g., *banconota* → banknote), and open compounds when they consist of lexemes separated by spaces or hyphens (e.g., *fuggi-fuggi* → rush).
- **Complex function word**: it is a function word formed by more than one lexeme, encompassing multiword conjunctions (e.g., *non appena* → as soon as), prepositions (e.g., *fino a* → up until), and adverbials (e.g., *in linea di massima* → by and large).
- **Multiword named entity**: a multiword linguistic expression that rigidly designates an entity in the world, typically including people, organizations, and locations (e.g., *Organizzazione delle Nazioni Unite* → United Nations).
- **Multiword term**: a multiword designation of a general concept in a specific subject field (e.g., *missione scientifica a breve termine* → short-term scientific mission).

More specifically, MWEs are characterized by a set of properties, pointed out by Markantonatou et al. (2018), which increase the difficulty of their automatic processing:

1. **Semantic non-compositionality**. In numerous cases, the meaning of VMWEs cannot be deduced on the basis of their syntactic structure and of the meanings of their components. For instance, the meaning of *me lo ha detto l'uccellino* (a bird told me that) as *qualcuno me lo ha detto in segreto* (someone told me that in secret) cannot be deduced by the meanings of *dire* (tell) and *uccellino* (little bird).

2. **Lexical and grammatical inflexibility**[7]**.** Lexical and syntactic constraints of VMWEs may be unpredictable, e.g., *ha messo il carro davanti ai buoi* (lit. 'he put the cart in front of the oxen' → he put the cart in front of the horse) e non **ha messo i carri davanti ai buoi* (lit. 'he put the carts in front of the oxen') or **ha messo il calesse davanti ai buoi* (lit. 'he put the calesh in front of the oxen').
3. **Regular variability.** Even though VMWEs present lexical and grammatical inflexibility, they may present some regular variability as well, e.g., *prendere una decisione* (to make a decision): *La decisione che prendemmo* (the decision we made).
4. **Discontinuity.** Elements in a VMWE may not be adjacent, e.g., *fornire un contributo* (to make a contribution): *Ha fornito un rilevante contributo al progetto* (He made a significant contribution to the project).
5. **Categorical ambiguity.** VMWEs sharing the same syntactic structure and lexical choices, as in *fare un discorso* (to give a speech) and *fare un dolce* (to make a cake), may belong to different categories, i.e., *fare un discorso* is a light verb construction, while *fare un dolce* is not an MWE in that the element co-occurring with the verb is a concrete noun (Ninio 2011).
6. **Syntactic ambiguity.** VMWE occurrences may be syntactically ambiguous, e.g., *giù* is an adverb in *buttare giù la palla* and a particle in *buttare giù un boccone*, where it takes the meaning of *to swallow*.
7. **Literal-idiomatic ambiguity.** Some VMWEs may present both a literal and idiomatic meaning, e.g., *Ha preso il toro per le corna* (lit. 'he took the bull by its horns' → grasp the nettle).
8. **Non-literal translatability.** VMWEs usually may not be translated by means of a word-for-word process. *Il mattino ha l'oro in bocca* (lit. 'the morning has gold in its mouth' → the early bird catches the worm).
9. **Cross-lingual divergence.** VMWE behaviours change across different languages, as they are the result of different linguistic traditions. For instance, in Germanic languages *off* has a status of stand-alone word and forms verb-particle constructions, while in Slavic languages is a prefix and becomes an inherent part of verbal lexemes (Markantonatou et al. 2018) as in (PL) *wyłączyć* 'part. connect' → turn off).
10. **Wordplay proneness.** VWMEs allow playful usage and creativity in some specific contexts. For instance, *vuole che rimetta tutto nel sacco dopo che l'ho svuotato* (lit. 'He wants me to put everything again in the bag after I have emptied it') from *svuotare il sacco* with the idiomatic meaning of *to blow the whistle*.

Two threads of research are relevant to our work: (i) linguistic studies on Italian VMWEs, mainly with the contribution of scholars working on the Italian language; and (ii) MWE Processing. The former aims at presenting current research outputs in

7 Sheinfux et al. (2019) provide an interesting discussion on the concept of inflexibility of VMWEs, starting from the work by Gibbs et al. (1989) and Nunberg et al. (1994).

contrastive/comparative analyses and synchronic and diachronic studies. The latter takes into account computational researches on MWE processing, as the developed corpus is intended to improve the automatic processing of these linguistic phenomena.

Linguistic Studies on VMWEs. Several scholars have investigated different categories 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 studies include analyses of: (i) verb-particle constructions (Simone 1997; Masini 2005; Iacobini and Masini 2005; Quaglia and Trotzke 2017), (ii) idiomatic constructions (Tabossi, Arduino, and Fanari 2011; Vietri 2014c) with either ordinary or support verbs (Vietri 2014a), (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; Jezek 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 2014b), 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). With reference to Italian word combinations, it is worth mentioning the CombiNET project[8], which represents an important contribution to MWE extraction from Italian corpora (Nissim, Castagnoli, and Masini 2014), and SYMPAThy, a new approach to the extraction of this type of occurrences (Lenci et al. 2014). 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.

MWE Processing. MWEs have been the focus of the PARSEME COST Action, which enabled the organization of an international and highly multilingual research community (2015). This community launched in 2017 the first edition of the PARSEME shared task on automatic identification of VMWEs (Savary et al. 2017), which was replicated in 2018 (Ramisch et al. 2018) with the aim of developing universal terminologies, guidelines and methodologies for several languages, including the Italian language. To increase the computational efficiency of Natural Language Processing (NLP) applications, PARSEME focused on a special class of Multiword Expressions, namely VMWEs. The main outcomes include unified definitions and annotation guidelines for several types of VMWEs, as well as a large multilingual openly available VMWE annotated corpus.

In the first edition, eighteen languages were addressed, including 4 non-Indo-European languages. The task was co-located with the 13th Workshop on Multiword Expressions (MWE 2017) (Markantonatou et al. 2017), which took place during the

8 https://sites.google.com/site/enwcin/home

European Chapter of the Association for Computational Linguistics (EACL 2017). A corpus of 5.5 million tokens and 60,000 VMWE annotations in the 18 languages was released and distributed under different versions of the Creative Commons license.

In the second edition the annotation methodology was enhanced and the set of languages was changed reaching twenty languages. The task was co-located with the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018) (Savary et al. 2018) at COLING 2018 (Santa Fe, USA). A corpus of 6 million tokens and 79,000 VMWE annotations in the 20 languages was released and also, in this case, it is distributed under different versions of the Creative Commons license.

A focused overview of how MWEs are handled in NLP applications, with particular attention to the nature of interactions between MWE processing and downstream applications in NLP, such as MWE parsing and Machine Translation (MT) can be found in Constant et al. (2017).

With reference to MT, MWE-aware technologies have been proved successful in several cases (Pal, Naskar, and Bandyopadhyay 2013; Cap et al. 2015). In order to improve the quality of translation, various strategies, depending on the MT paradigm, have been proposed to overcome problems related to MWE processing (Ren et al. 2009; Kordoni and Simova 2014; Ramisch, Besacier, and Kobzar 2013; Barreiro et al. 2014). Also in neural approaches to MT, some recent contributions show that the proper handling of MWE improves the translation of MWEs by adding synthetic MWE data to the training corpora (Rikters and Bojar 2017) or by annotation and data augmentation, using external linguistic resources (Zaninello and Birch 2020).

Finally, the workshop series titled Multiword Units in Machine Translation and Translation Technology (MUMTTT) (Monti et al. 2013; Pastor et al. 2015; Monti et al. 2018; Pastor et al. 2019) and the recent volume on the same topic (Mitkov et al. 2018) provide an overview of state-of-the-art research in this field and highlight the importance of proper computational treatment of these lexical units in MT and translation technology (TT).

Besides NLP tasks, cross-lingual studies of multiwords and automatic extraction of translation equivalents represent an important field of research. With the aim of building MWE repositories, Wehrli and Villavicencio (2015) propose an extraction methodology based on aligned corpora for English, Portuguese and French. They combine a symbolic parser with a high-recall statistically-based extraction method and identify correspondences in the language pairs using alignment and distributional methods (de Caseli et al. 2010; Laranjeira et al. 2014).

Acknowledging the diversity of idiomatic structures,Villavicenzio et al. (2004) propose a framework for the cross-lingual collection of idioms and mapping of their equivalent parts which allows the identification of similarity at semantic, syntactic and lexical levels.

Statistical methods have been applied to parallel corpora (Wehrli and Villavicencio 2015) to evaluate their cross-lingual applicability for idiomatic pattern identification, while Taslimipoor et al. (2016) improve the performance of monolingual association measures by augmenting them with information about translation equivalents and using them to produce a ranking of expressions according to their idiomaticity.

3. PARSEME-It VMWE Corpus

This section outlines the **PARSEME-It VMWE corpus** (version 1.1), annotated with VMWEs for the Italian language. As described in the previous sections, the corpus is

the main outcome of the PARSEME-It project together with the general and language-specific guidelines for the PARSEME annotation process.

The corpus is based on a selection of texts taken from the *PAISÀ* corpus of Italian web texts[9] (Lyding et al. 2014). We chose this corpus because its documents are:

1. representative of 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;
2. dedicated to no specific technical domain, free from copyright issues, so as to be compatible with an open license;
3. 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 15,728 sentences (corresponding to 430,789 tokens) randomly taken from blogs, Wikipedia and Wikinews. Due to the heterogeneous sources, e.g., social media, blogs, forum posts, consumer reviews, texts present variable characteristics: inconsistent punctuation and capitalization, use of slang and technical jargons, specific syntactic constructions related to genres. Nevertheless, the corpus was kept in its original state and therefore no errors or inconsistencies were corrected. The automatically pre-annotated information in the original corpus, namely morpho-syntactic and dependency annotations[10], were kept to ease the annotation work regarding the identification of VMWEs, but we asked annotators not to overestimate the system's performances, and to review the whole text, not only the pre-annotated candidates, namely all the verbs (V). A dedicated tag in FLAT, the web-based annotation environment used in the project (Section 6), was defined for this purpose.

The objective was to have a final corpus of at least 3,500 annotated VMWEs. Since the density of VMWEs in the corpus is highly dependent 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.

4. VMWE Categories

For the Italian VMWE annotation task, according to the 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, for instance, it can also be a coordination.

9 https://www.corpusitaliano.it/en/

10 The tag sets for such annotation have been developed by the Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) and the University of Pisa in the framework of the TANL (Text Analytics and Natural Language processing) project.

- 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.

Only the lexicalized components[11] were annotated and open slots[12] ignored, as in *prendere qualcuno di sorpresa*, only *prendere ... di sorpresa* are annotated, while *qualcuno* is not because it can be replaced by a noun or a pronoun. Collocations, i.e., word co-occurrences whose idiosyncrasy is of statistical nature only (e.g., *the graphic shows, drastically drop*, etc.), were excluded from annotation as well. Therefore, 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 make a question). 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). This category has two subclasses: i) LVCs in which the verb is semantically totally bleached (**LVC.full**), e.g. *fare una passeggiata* (to have a walk) and ii) LVCs in which the verb adds a causative meaning to the noun (**LVC.cause**), e.g. *dare il mal di testa* (to give a headache);
2. **Idioms** (VID), which have at least two lexicalized components including a head verb and at least one of its arguments, e.g., *tirare le cuoia* (to kick the bucket), *piovere a catinelle* (to rain cats and dogs);
3. **Inherently reflexive verbs** (IRV), account for those reflexive verbal constructions (a) which are never used without a reflexive clitic pronoun e.g., *suicidarsi* (to suicide), or (b) when the IRV and non-reflexive versions have clearly different senses or subcategorization frames e.g., *farsi* (to take drugs) while the non-pronominal form, *fare*, means *to make*.
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 non-compositional. Notably, the change in the meaning of the verb goes significantly beyond adding the meaning of the particle, e.g., *fare fuori* (lit. 'to do out' $\rightarrow$ to kill). This type of construction is very frequent in English, German, Swedish, Hungarian, but we can find it also in Italian. The VPC category is split in two subcategories as well: fully non-compositional VPCs (**VPC.full**), in which the particle totally changes the meaning of the verb as in *fare fuori* and semi non-compositional VPCs (**VPC.semi**), in which the particle adds a partly predictable but non-spatial

11 According to Savary and Cordeiro (2018), the lexicalized components of an MWE are those which are always realized by the same lexeme. For instance in *to pay a visit* the head verb is always a form of *pay* and the object is always *visit*: these two elements are therefore lexicalized components of the VMWE.

12 An open slot (Savary and Cordeiro 2018) is a component of a compulsory argument which can be realized by a free lexeme taken from a relatively large semantic class. If we consider again the example of the VMWE *to pay a visit*, an open slot is represented by the determiner *a*, which can be freely replaced, as in *paid many visits*.

meaning to the verb like in *tirare avanti* (to go on) since the preposition *avanti* no longer owns its spatial meaning (forward).

5. **Multi Verb Constructions** (MVC), which are composed by a sequence of two adjacent verbs (in a language-dependent order), a governing verb (also called a vector verb) and a dependent verb (also called a pole/polar verb), e.g. *lasciar perdere* (lit. 'let lose' → to forget).

6. **Inherently Adpositional Verbs** (IAV), which consist of a verb or VMWE and an idiomatic selected preposition or postposition that is either always required like in *appartenere **a*** (to belong to) or, if absent, changes the meaning of the verb of VMWE significantly, like in *contare **su*** where the preposition *su* is required to express the meaning of 'to rely on' compared to the verb without the preposition which means *to count*. It is a special optional and experimental category, corresponding to what is sometimes called in English *prepositional verbs*[13].

Besides these categories, shared by all languages involved in the PARSEME COST Action, language specific categories have been introduced in edition 1.1 of the PARSEME Shared Task. For the annotation of the Italian language, the Inherently Clitic Verbs (LS.ICV) category was proposed and carefully defined by means of linguistic tests that allow to distinguish this category from IRVs.

A language specific category: Inherently clitic verbs. LS.ICVs are an extremely rich and varied VMWE category for some Romance languages, and they are particularly frequent in the Italian language (Masini 2015).

LS.ICVs together with IRVs are pronominal verbs (De Mauro 2000): they are formed by a full verb combined with one or more non-reflexive clitics that represent the pronominalization of one or more complements (CLI) (Viviani 2006; Berruto 1987). LS.ICV is annotated when (a) the verb never occurs without one non-reflexive clitic, e.g. *entrarci* colloquial form, or (b) when the LS.ICV and the non-clitic versions have clearly different senses or subcategorization frames, like *entrar**ci*** when it means *to be relevant to something*, while the intransitive form of the verb *entrare* means *to enter*.
It is often challenging to distinguish LS.ICV from IRV, particularly because some clitics may be ambiguous, like *se/si* (Cinque 1988; Cordin 2001; Pescarini 2015) which is a polyfunctional clitic pronoun and grammatical marker (and can have a reflexive, reciprocal, impersonal, passivizing, aspectual, and middle function).

The following verbs are annotated as LS.ICV:

- The verb without the CLI does not exist, e.g., *infischiarsene* (do not worry about) vs **infischiare, *infischiarsi*;
- The verb without the CLI does exist, but has a very different meaning as in *prenderle* (lit. 'to take them' → to be beaten) vs *prendere* (to take) or *prenderci* (lit. 'to take it' → to grasp the truth) vs *prendere* (to take);

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

- The verb has more than one CLI of which the second one is an invariable object complement, like in *fregarsene* (lit. 'to matter self of-it' → do not care about) or *infischiarsene* (do not worry about);
- The verb has two non-reflexive invariable CLIs, like in *farcela* (lit 'to make there it'→ 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* (lit. 'to go away self from-there' → to die) vs *andarsene* (to go away) or *bersela* (lit. 'to drink self it', → to believe) vs *bersela* (to drink it).

A language-specific decision tree to annotate LS.ICV occurrences was developed, as described in Section 5.

5. Annotation Guidelines and Decision Trees

The PARSEME annotation guidelines have been developed with the aim of delivering general definitions and prescriptions for the annotation of VMWEs in all languages involved in the shared task, but, at the same time, of allowing language-specific descriptions of these linguistic phenomena (Savary et al. 2017). We describe here the guidelines and methodologies used for the second annotation trial of the Shared Task, which introduced some novelties to cover a wider range of VMWEs, left apart in the first edition. The improvements of the second edition were particularly valuable for the data collection carried out on the Italian language, because they addressed some peculiarities of the Italian language which were not considered previously, such as the LS.ICV category.

For the second edition of PARSEME annotation task, the following categories were identified:

1. two **universal** categories, common to all languages involved in the task and hold both LVC categories, namely **LVC.full**, and **LVC.cause**, and idioms (**VID**);
2. three **quasi-universal** categories, relevant for some languages or language families but non-existent or very exceptional in others. This category encompasses **IRV**, the two subclasses of VPCs, namely **VPC.full** and **VPC.semi** and finally **MVC**;
3. the **optional** VMWEs category **IAV**;
4. **language-specific** categories, defined for a particular language in separate documentation, as in the case of the Italian language, the **LS.ICV**.

5.1 Identification tests

In order to ease the identification and categorisation task of VMWEs, a decision 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. Each iteration of the annotation process includes three steps:

1. Identification of a VMWE candidate, i.e., a combination of a verb with at least one other word, which is a potential VMWE;
2. Identification of the lexicalized elements of the expression;
3. Assignment of the VMWE to one of the VMWE categories, using general and language specific tests.

The first two steps largely rely on the annotator's linguistic intuition and knowledge. As reported by Markantonatou et al. (2018), the identification of a VMWE, regardless of the category, may be accomplished by five generic tests on compositional aspects.

- Test 1 [CRAN] : Presence of a cranberry word, e.g., *mangiare a ufo* (to eat without paying) → *a ufo* is not a stand-alone word;
- Test 2 [LEX] : Lexical inflexibility, e.g., *non dire gatto se non ce l'hai nel sacco* (lit. 'don't say cat if you don't have in the sack' → don't count on something before it happens) vs **non dire cane se non ce l'hai nel sacco* (lit. 'don't say dog if you don't have in the sack');
- Test 3 [MORPH] : Morphological inflexibility, e.g., *andare a letto con le galline* (lit. 'to go to bed with the hens' → to go to bed early) vs **andare a letto con la gallina* (to go to bed with the hen);
- Test 4 [MORPHOSYNT] : Morpho-syntactic inflexibility, e.g., *farò del mio meglio* (I will do my best) vs **Farò del tuo meglio* (*I will do your best);
- Test 5 [SYNT] : Syntactic inflexibility, e.g., *vivi e lascia vivere* (live and let live) → **lascia vivere e vivi* (let live and live).

Besides these five tests, a specific hypothesis has been formulated to identify LVC candidates, which do not pass Tests 1 and 3-5 and for which Test 2 is hard to apply due to their high productivity, even though they present some restrictions.
LVC hypothesis: In a verb+(prep)+noun candidate the verb is a pure syntactic operator and the noun expresses an activity or a state, e.g. *fare un discorso* (to make a speech). If a candidate group passes any of the previous tests, it can be annotated as VMWE. To confirm the LVC hypothesis a specific test, namely Test 6 described in Section 5.2, has to be applied.

5.2 Category Decision Trees

In order to select a category for the identified VMWEs, a decision tree formed of both structural and category tests is provided (Figure 1). The decision tree is formed by a set of tests which help the annotator to identify and annotate VMWE candidates.

Tests S.1-S.4 (prev. 6-8) are structural, which means that the categorization is based on the syntactic structure of VMWE canonical form and defined by means of four tests:

- Test S.1 (prev. 6) [1HEAD]: Presence of a unique verb functioning as the syntactic head of the whole expression, e.g., *fare fuori* (lit. 'to make out' → to kill) → *fare* is the head and *fuori* is a particle depending on it;
- Test S.2 (prev. 7) [1DEP]: Among the phrases dependent on the head verb exactly one contains lexicalised components, e.g., *prendere in considerazione*

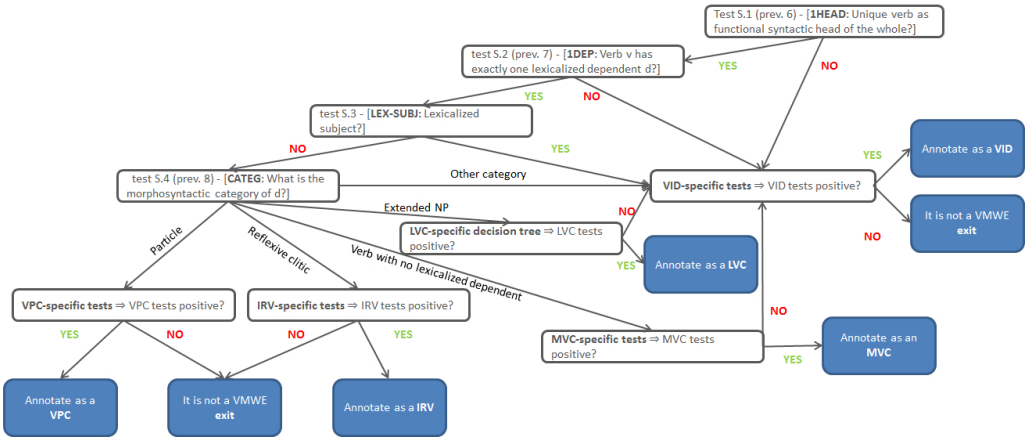

Figure 1
Decision tree for VMWE categorization

(to take into consideration) → the single dependent is a prepositional phrase, *in considerazione*;

- Test S.3 [LEX-SUBJ]: a single lexicalized (functional) syntactic dependent of the head verb is its subject, e.g., *me lo ha detto l'uccellino* (a bird told me) → *l'uccellino* (a bird) is the subject of *ha detto*;
- Test S.4 (prev. 8) [CATEG] Morphosyntactic category of the verb's dependent. This is a closed list of different values, namely (i) reflexive clitic (refl), e.g., *suicidarsi* (to suicide), (ii) particle (part), e.g., *far fuori* (lit. 'to make out' → to kill), (iii) no lexicalized dependent, e.g., *lasciar andare* (lit. 'to let go' → to unhand), (iv) adposition (preposition or postposition, as opposed to a particle), e.g., *confidare su* (to trust in), (v) extended nominal phrase, e.g., *rompere il silenzio* (to break the silence) → *il silenzio* is a noun phrase composed of an article and a singular noun, (vi), e.g., adjective *vedere nero* (to see black), (vii) adverb, e.g., *fare passi avanti* (lit. 'to make steps forward' → to progress), (viii) pronoun, e.g., *farcela* (lit. 'to make it' → to manage), (ix) verb with a lexicalised dependent including fully lexicalized clauses, e.g., *non avere peli sulla lingua* (lit. 'not have hair on the tongue' → to be outspoken), (x) other.

The other tests, i.e., VID-specific tests, LVC-specific decision trees, IRV-specific tests, VPC-specific tests, and MVC-specific tests are categorial and allow to categorize each of the classes identified initially. A complete analysis of those decision trees is provided by Markantonatou et al. (2018). Among these tests, we present the one created to classify the Italian language-specific category of ICVs.

The annotation of LS.ICV was performed following a specific decision tree[14] (Figure 2).

Three types of LS.ICV have been identified:

14 http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=060_Language-specific_tests/015_Inherently_clitic_verbs__LB_LS.ICV_RB_

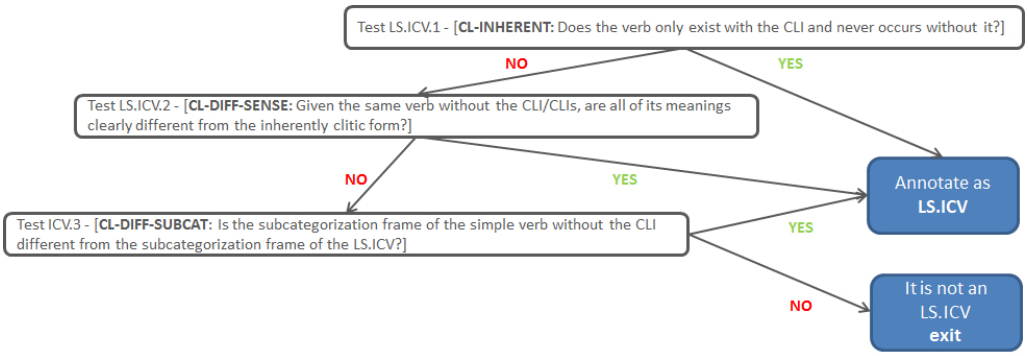

Figure 2
LS.ICV-specific decision tree

- Test LS.ICV.1 [CL-INHERENT]: the verb does exist only in the form with the clitic, e.g., *infischiarsene* (not worry about) vs *infischiare.
- Test LS.ICV.2 [CL-DIFF-SENSE]: the verb without the clitic exists but has a different meaning, e.g. *prenderle* (lit. 'to take them' → be beaten) vs *prendere* (take).
- Test LS.ICV.3 [CL-DIFF-SUBCAT]: the subcategorization frame[15] of the verb without the clitic is different from the subcategorization of the same verb with the clitic, e.g., *X se la prende con Y* (X is angry with Y) vs *X prende Y* (X takes Y).

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

6. Annotation Process and Inter-Annotator Agreement

For the annotation of the PARSEME-It VMWE corpus we used FLAT[16], a web-based linguistic annotation environment based around the FoLiA format[17] a rich XML-based format for linguistic annotation. FLAT is a document-centric tool that fully preserves and visualises document structure and allows users to view annotated FoLiA documents and enrich these documents with new annotations (Figure 3)[18]: it offers a wide variety of linguistic annotation types supported through the FoLiA paradigm.

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

15 A subcategorization frame of a verb describes how syntactic arguments are realized as the verb's dependents, for a given sense of the verb. A subcategorization frame indicates morphological and syntactic features of a verb's dependents, namely the required prepositions, postpositions and case markers of the subject, direct and oblique objects.
https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/?page=glossary#subcat-frame

16 FLAT is an open-source software developed at the Centre of Language and Speech Technology, Radboud University Nijmegen and is licensed under the GNU Public License v3 -
http://flat.readthedocs.io/en/latest/

17 http://proycon.github.io/folia

18 Translation of the example in fig. 3: *Perhaps, inadvertently, Monckton and Fielding did not make such a foolish request*.

1. The PARSEME Annotation guidelines were agreed on[19] 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 was

Figure 3
Example of annotated data in FLAT

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 PAISÀ corpus was needed. Although the tokenization follows the original tokenization of the PAISÀ corpus, some pre-processing has been applied to the original files of the corpus in order to split compound prepositions (*dei, nei, delle*, etc.), e.g., *dei* is split in the preposition *di* + the determiner *i* to allow the annotation of the preposition only, for instance, as lexicalised component of IAVs. To this end, we added new tokens corresponding to the components of the compound prepositions (see example below[20] in Table 1-2) and we also realigned all the dependency index: the heuristic being used is that the preposition is the head of the prepositional article (all tokens pointing to the prepositional article will point to the preposition in the split version and the determiner also points to the preposition). For instance the original CONLL-U sentence in Table 1[21]. In addition, we also introduced the

Table 1
Original CoNLL-U sentence

Rank	Surf	Lemma	PosG	PosF	Morph	DepIndex	DepLabel
1	Perchè	Perchè	C	CS	_	4	mod
2	la	il	R	RD	num=s|gen=f	3	det
3	ragione	ragione	S	S	num=s|gen=f	4	subj
4	sta	stare	V	V	num=s|per=3|mod=i|ten=p	0	ROOT
5	nel	in	E	EA	num=s|gen=m	4	comp
6	mezzo	mezzo	S	S	num=s|gen=m	5	prep
7	no	no	B	BN	_	4	neg
8	?	?	F	FS	_	4	punc

`SpaceAfter=No` tag on the word preceeding a clitic belonging to the

19 http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/?page=home
20 source: https://gitlab.com/parseme/sharedtask-data//tree/master/1.1/IT
21 For more information about CoNLL-U format, see https://universaldependencies.org/format.html

Table 2
Transformed sentence

Rank	Surf	Lemma	PosG	PosF	Morph	DepIndex	DepLabel
1	Perché	Perché	C	CS	_	4	mod
2	la	il	R	RD	num=s|gen=f	3	det
3	ragione	ragione	S	S	num=s|gen=f	4	subj
4	sta	stare	V	V	num=s|per=3|mod=i|ten=p	0	ROOT
5-6	nel	_	_	_	_	_	_
5	in	in	E	E	_	4	comp
5	il	il	R	RD	_	5	det
7	mezzo	mezzo	S	S	num=s|gen=m	5	prep
8	no	no	B	BN	_	4	neg
9	?	?	F	FS	_	4	punc

same token, e.g., *lavar-si*. These are annotated as two separate words in the original corpus.

3. The annotation task of the training set (approx. 14,000 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 VID 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 VID *guardarsi bene* there is also a IRV *guardarsi* or not.
4. A few files were double-annotated to evaluate the inter-annotator agreement (IAA).
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.

The current version of the PARSEME-IT corpus (1.1) represents a substantial improvement (Monti et al. 2018) in comparison to its first version (Monti, di Buono, and Sangati 2017) both in terms of categories of VMWEs taken into account for the annotation and total amount of annotated VMWEs.

Table 3 presents the statistics of the various categories of VMWEs in the PARSEME-It corpus 1.0[22], where only five categories were taken into account, namely ID (corresponding to the current VID category), IReflV (corresponding to the current IRV category), LVC, VPC and a OTH category for the VMWEs which could not be included in the previous categories. This version of the PARSEME-It corpus encompasses 1,954 VMWE annotations.

Table 3
PARSEME-It corpus version 1.0

Sent.	Tokens	VMWE	ID	IReflV	LVC	VPC	OTH
15728	387325	1954	913	580	395	62	4

Table 4, instead, shows information about the corpus version 1.1 released for the second edition of the PARSEME shared task, where a total amount of 3,754 VMWEs are annotated.

Table 4
PARSEME-It corpus version 1.1

Lang-split	Sent.	Tokens	Avg. length	VMWE	VID	IRV	LVC	VPC	IAV	MVC	LS.ICV
IT-train	13555	360883	26.6	3254	1098	942	691	66	414	23	20
IT-dev	917	32613	35.5	500	197	106	119	19	44	6	9
IT-test	1256	37293	29.6	503	201	96	129	23	41	5	8
IT-Total	15728	430789	27.3	4257	1496	1144	939	108	499	34	37

PARSEME-It VMWE corpus 1.1. includes i) the manually annotated training set, ii) manually annotated development set and finally iii) the automatically annotated test set. For each of those morphosyntactic data (parts of speech, lemmas, morphological features and/or syntactic dependencies) are also provided.

The data have been annotated using the official parseme-tsv format (Figure 4), adapted from the CoNLL format.

In the official parseme-tsv format, as described in Savary et al. (2017), the information about each token is represented by 4 tab-separated columns featuring:

- 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;
- the token surface form;
- an optional flag indicating that the current token is adjacent to the next one;
- an optional VMWE code composed of the VMWE's consecutive number in the sentence and – for the initial token in a VMWE – its category (e.g., 2:ID

22 The corpus is provided in the parseme tsv format, inspired by the CONLL-U format https://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation

```
1       In      _       _
2       prossimità      _       _
3       della   _       _
4       tornata _       _
5       elettorale      _       _
6       per     _       _
7       la      _       _
8       rielezione      _       _
9       delle   _       _
10      cariche _       _
11      di      _       _
12      assessori       _       _
13      alla    _       _
14      Regione _       _
15      Veneto  _       _
16      qualcuno        _       _
17      vuole   _       _
18      far-    _       1:ID
19      gli     _       _
20      le      _       1
21      scarpe  nsp     1
22      ?       _       _
```

Figure 4
Example of annotated data in parseme-tsv format

if a token starts an idiom which is the second VMWE in the current sentence).

In the 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-U [23] have been released. These companion files contain extra linguistic information, i.e., lemmas, POS-tags, morphological features, and syntactic dependencies.
Measuring inter-annotator agreement (IAA) is not a trivial task because of the challenges posed by VMWEs and described in the Introduction. Yet, for most languages, including Italian, the majority of the corpus has been annotated by a single annotator because of time and resource constraints. Thus, a small representative part of the corpus has been annotated by two annotators in order to calculate the IAA. The proposed IAA measures intend to assess different aspects, such as the resulting annotation, as well as the effort required in the annotation task and the guidelines and methodology applied. The available IAA results for the first edition of the PARSEME Shared Task, organized per-VMWE F-score (F_{unit}), estimated Cohen's K (K_{unit}) and finally standard K(K_{cat}) (Savary et al. 2017) scores are presented in Table 5.
To measure the unitising value [24] the MWE-based F-score (F_{unit}), as defined in Savary et al. (2017), has been calculated on one annotator with respect to the other considering the double-annotated data.
As noted by Markantonatou et al. (2018), measuring IAA, especially for Cohen's kappa (κ_{unit}), is not straightforward due to the lack of negative examples, namely spans formed of combination of a verb with other tokens (of any length) in a sentence for which both annotators agreed that they are not VMWEs. To reduce the bias in this measure with

23 http://universaldependencies.org/format.htm
24 Unitising is referred to the identification of the boundaries of a VMWE in the text;

Table 5
IAA scores for PARSEME-It VMWE corpus 1.0: #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.

	#S	**#T**	**$\#A_1$**	**$\#A_2$**	**F_{unit}**	**κ_{unit}**	**κ_{cat}**
IT	2000	52639	336	316	0.417	0.331	0.78

reference to the F-score, the total number of possible VMWE candidates in the corpus has been assumed to be equivalent to the number of verbs, which is actually higher than the number of sentences and nevertheless estimated as the number of sentences plus the VMWE annotated at least by one annotator (Savary et al. 2017).
The standard κ (κ_{cat}) is applied to calculate the agreement on categorization, considering just the double-annotated VMWE spans. Italian, as other languages in the PARSEME annotation task, e.g., Spanish[25], shows low IAA scores, especially in unitising.
Table 6 shows the IAA scores for the second edition of PARSEME-It VMWE corpus.

Table 6
IAA scores for PARSEME-It VMWE corpus 1.1

	#S	**$\#A_1$**	**$\#A_2$**	**F_{span}**	**κ_{span}**	**κ_{cat}**
IT	1000	341	379	0.586	0.550	0.882

The IAA has been evaluated on a sample of 1,000 sentences, with A_1 and A_2 VMWEs annotated by each annotator. F_{span} is the F-measure between annotators, κ_{span} is the agreement on the annotation span and κ_{cat} is the agreement on the VMWE category (Ramisch et al. 2018). Although the IAA values increased in the second annotation campaign due to the presence of more fine-grained categories and better training of the annotators, these values are not so high, which can be explained by several reasons: (i) annotating some types of texts, i.e., Web texts in our corpus, are more difficult than annotating other types of texts, e.g., newspaper; (ii) double-annotated samples are quite small; (iii) guidelines and annotator training have to be improved. At any rate, these results call for a deep analysis of the issues arisen during the annotation, as presented in the following section.

7. Annotation Issues

In this section, we discuss the main annotation issues which emerged during the annotation finalized at assessing the IAA in the second edition of the Shared task. During this phase a set of 1,000 sentences was double-annotated by two different skilled annotators. The two annotators annotated almost the same number of VMWEs, namely ANNOTATOR1 341 VMWEs and ANNOTATOR2 379 VMWEs, but they completely agreed on the number of constituents and category only in 191 cases. VMWE annotation is a very hard task and disagreements occurred in different forms:

25 For IAA values for other languages, see Markantonatou et al. (2018).

1. Partial matches (labeled): this type refers to disagreements concerning the number of constituents of a VMWE labeled in the same way by both annotators;
2. Exact matches (unlabeled): this type refers to disagreements concerning the category of VMWE only;
3. Partial matches (unlabeled): this type refers to disagreements concerning the number of constituents and the category of a VMWE;
4. Single-annotated occurrences: this type refers to VMWEs annotated only by one annotator.

The disagreements will be discussed in the next subsections.

7.1 Partial Matches Labeled

A first source of disagreement is represented by the inclusion or exclusions of one or more constituents of VMWEs. Differences in annotation arise in relation to the judgment about the lexicalization of a component word of a VMWE, which might prove particularly difficult in presence of determiners, adjectives, pronouns/clitics, negations. In 25 cases different decisions were taken by the annotators on whether these words were part or not of VMWEs, resulting only in partial overlapping in the annotations, like in the examples provided below.

Inclusion/exclusion of determiners. The example provided in (1) refers to the VMWE *dare aiuto* (to help), which has been labeled as LVC.full by both annotators, but while ANNOTATOR1 identified the VMWE as *dare ... aiuto* ANNOTATOR2 included the determiner *un* as lexicalised constituent of the VMWE and therefore labeled *dare un aiuto*. In fact, it is possible to test whether a determiner is lexicalized by searching alternatives in dictionaries, corpora, or on the web. Borderline cases exist, in which alternatives are rare but possible, specially for LVCs and decomposable VIDs. The general rule, however, is that when alternatives are possible and the determiner varies, then it should not be included in the annotation.

1. `Source: PARSEME-It VMWE 2`
   Se sarà vero è una questione che dovranno risolverla tra loro e se qualcuno è a conoscenza dei fatti accaduti può **dare un aiuto** ad uno o all'altro contendente.
   (If it is true, they will have to solve the issue among themselves and if someone is aware of the events that have occurred, they can help one or the other contender.)

Inclusion/exclusion of adjectives. Another example of disagreement between annotators is given by the presence of an adjective which might be considered as part of a VMWE although it is not completely fixed, as in example (2) where both annotators identified the VID *porre in ... luce* but there was a different judgement with reference to the adjective *cattiva* as being part or not of the lexicalised constituents of the VMWE. This is due to the possibility to have alternative adjectives like *buona* as in *porre in buona luce* or *chiara* as in *porre in chiara luce*. The problem to be solved in this respect is to decide if the different adjectives convey a different meaning for the VMWE to be annotated.

2. Source: PARSEME-It VMWE 392
La stampa ha presentato la cosa in modo non corretto, ed alcuni commentatori l'avevano utilizzata, **ponendo in cattiva luce** l'immagine della Giolo, che si era limitata a fotografare per l'eventuale utilizzo in caso di ricorso.
(The press presented this incorrectly, and some commentators had used it, putting in a bad light the image of Giolo, who limited herself to take pictures for a possible use in case of appeal.)

Inclusion/exclusion of negations. Negations are usually also considered non lexicalized. However, this is not always the case and they might also represent a source of different judgments between annotators. For instance, the VID *non fare una cippa*, a substandard expression with the meaning of 'don't do anything' in example (3) presents a lexicalised negation which nevertheless causes some doubts in ANNOTATOR1 who does not annotate it as part of the VMWE.

3. Source: PARSEME-It VMWE 408
A me sembra, da esterna che segue da anni la manifestazione perchè a Rovigo quest'anno a mio parere **non hanno fatto una cippa** che stiate cercando di spremere un limone già secco.
(It seems to me, who has been following the event for years from outside since, in my opinion, they haven't done a bit in Rovigo this year, that you are trying to squeeze an already dry lemon.)

Inclusion/exclusion of clitics. Clitics also challenge very often judgments as to whether they are part or not of VMWEs like in *fare le spese*. In example (4) only ANNOTATOR2 annotated the non-reflexive clitic *-ne* as a constituent of a VID, considering it as a fixed element of the VMWE.

4. Source: PARSEME-It VMWE 492
è tutto uno scaricabarile... e a **farne le spese** sono i ragazzi.
(it's all passing the buck ... and the boys are the ones who pay for it.)

Inclusion/exclusion of pronouns. Pronouns, indeed are also usually non-lexicalised since they can vary, but example (5) caused another disagreement as to whether the pronoun is a constituent or not of a VMWE. Here the judgment of the annotator that included the personal dative pronoun *ti* in the annotation of the VMWE *stare bene* probably is based on the idea that the meaning of the VMWE *stare bene a qualcuno* (to look good on someone) is different from the meaning of *stare bene* (to feel well). In this case, the presence of the pronoun conveys a completely different meaning although it is not invariable as other personal pronouns are equally acceptable, e.g. *(mi/ti/gli/...) sta bene*.

5. Source: PARSEME-It VMWE 966
Certo che **ti sta** proprio **bene**... è questa la sorpresa?
(It looks good on you ... is this the surprise?)

Mistakes in annotations. In the category of partial matches labeled there are also 4 mistakes, such as annotation of single words instead of multiwords, or un-annotated

elements of a VMWE. For instance, in example (6) ANNOTATOR1 did not annotate the verb of the VID *mettete*, while ANNOTATOR2 annotated it.

6. `Source: PARSEME-It VMWE 646`
Perché non ne abbiamo già abbastanza di fastidi tra Spinello e Barbujani e **vi ci mettete** anche voi?
(Don't we have enough of annoyances between Spinello and Barbujani and do you contribute too?)

7.2 Exact Matches Unlabeled

In this case, annotators identify the same constituents but disagree on the category of VMWEs. The disagreements (18 cases) mainly concern LVCs (full and cause) and VPCs (full and semi): these categories are very fine-grained and pose some problems in the assessment of the grade of non- compositionality. Another frequent source of disagreement concerns different decisions as to whether a VMWE belongs to the VID or LVC category(both full and cause). Disagreements concerning exact matches were eliminated in version 1.2 of the corpus [26].

VPC. As already mentioned, in fully non-compositional VPC (VPC.full) the change in the meaning of the verb goes significantly beyond adding the meaning of the particle: like for *buttare giù* with the meaning of *to swallow*. In semi-non-compositional VPCs (VPC.semi), the particle adds a partly predictable but non-spatial meaning to verb: like in *lasciare dietro* with the meaning of *to leave behind*. The LVC *mettere insieme* causes some uncertainties as to whether it is a VPC.full (ANNOTATOR1) or a VPC.semi (ANNOTATOR2).

7. `Source: PARSEME-It VMWE 7`
... ringrazio il sindaco Barbujani e la giunta che ha permesso di **mettere insieme** un programma di tutto rispetto.
(I thank Mayor Barbujani and the council that made it possible to put together a very respectable program.)

LVC. The verb is "light" in that it contributes to the meaning of the whole only by bearing morphological features: person, number, tense, mood, as well as morphological aspects. This implies that the syntactic subject of the verb is the semantic argument of the noun [27]. In this case, we annotate the construction as LVC.full like in *fare una presentazione* (to make a presentation). If the verb is "causative" in that it indicates that the subject of the verb is the cause or source of the event or state expressed by the noun, the VMWE should be annotated as LVC.cause like in *dare le vertigini* (to make dizzy). In example (8) annotators do not agree on the LVC type of the verb *dare fiducia* and ANNOTATOR1 labels it as LVC.cause while ANNOTATOR2 as LVC.full.

8. `Source: PARSEME-It VMWE 810`

26 https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2/IT
27 https://parsemefr.lislab.fr/parsemestguidelines/1.1/?page=050_Cross-lingual_tests/020_Light-verb_constructions__LB_LVC_RB_

La nostra volontà la vogliamo portare in Consiglio Comunale approvando il PAT, che è a portata di mano, che è in dirittura d'arrivo e servirà **dare fiducia** ai cittadini.
(We want to bring our will to the City Council by approving the PAT, which is close at hand, which is in the home stretch and will serve to trust citizens.)

Disagreement about VID and LVC. A frequent disagreement between the annotators concerned the VID and LVC categories, like example (9) where the VMWE *fare la parte* was annotated as a VID by ANNOTATOR1 and as a LVC by ANNOTATOR2. This uncertainty may be due to different judgments given to the tests applied in the decision process. In particular the annotators might have taken different decisions with respect to some tests concerning VIDs, like Test VID.2 - [LEX] - Lexical inflexibility: this test requires to assess whether the regular replacement of one of the components by related words taken from a relatively large semantic class leads to ungrammaticality or to an unexpected change in meaning, for instance in this case whether the replacement of the verb *fare* with *sostenere* or of the determiner *la* with the indefinite article *una* leads to different meanings. In case of a negative answer, annotators should have taken Test VID.3 - [MORPH] - Morphological inflexibility which requires to assess whether regular morphological change that would normally be allowed by general grammar rules leads to ungrammaticality or to an unexpected change in meaning, for instance, whether *fare le parti* has a different meaning with respect to *fare la parte*. Therefore, while ANNOTATOR1 answered positively to one of the abovementioned tests, ANNOTATOR2 answered negatively to them and answered positively to one of the tests for LVCs[28].

9. `Source: PARSEME-It VMWE 425`
Lasciate lavorare la maggioranza e lasciate l'opposizione **fare la parte** che gli compete.
(Let the majority work and let the opposition do its part.)

7.3 Partial Matches unlabeled

The only case of partial match unlabeled concerns a different interpretation of the VMWE both in terms of the number of constituents and category attribution. The example (10) presents the VMWE *buttarsi (nella calca)* which was labeled by ANNOTATOR1 as *buttar-si in la calca* (VID) and by ANNOTATOR2 *buttar-si* (IRV).

10. `Source: PARSEME-It VMWE 864`
Chi è rimasto nei pressi della propria città approfittandone per sistemare casa, alzandosi la mattina tardi, passeggiando per il corso attendendo il venerdì sera per **buttarsi nella calca** del divertimento...
(Those who stayed close to their city taking advantage of it to settle home, getting up late in the morning, walking along the street waiting for Friday evening to mix in the crowd of fun ...)

28 https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/?page=lvc#decision-tree-lvc

7.4 Single-Annotated Occurrences

The main problems of disagreement lie in the high number of VMWEs annotated by only one annotator: 250 cases split in 106 for ANNOTATOR1 and 144 for ANNOTATOR2 (Table 7).

Table 7
Single-annotated occurrences for each category

IAV	VID	MVC	LVC.full	LVC.cause	VPC.full	VPC.semi	IRV	LS.ICV
28	66	5	72	10	11	1	47	10

From these figures, it emerges very clearly that the most problematic VMWE category is represented by LVCs. One possible reason is that the verbs of LVCs are very common ones such as *fare* (*fare ricorsi, fare errori*), *dare* (*dare allucinazioni, dare informazioni*), *prendere* (*prendere visione*), *avere* (*avere difficoltà, avere esperienza*) and since these verbs share the same meaning with other lexical constructions which are not LVCs, annotators may not identify them as such. For instance, the verb *avere* does not change its meaning from *avere una sedia* (non-VMWE) to *avere difficoltà* (VMWE). Besides, it is clear from the annotations that sometimes meaning-preserving variants of a (candidate) VMWE such as verbal expressions with analytical tenses and modals, like in *hanno preso una decisione*, nominal groups (headed by nominal complements from the prototypical VMWEs) with relative clauses (e.g., *i cuori che abbiamo spezzato*), non-finite verbal clauses (e.g., *decisioni prese precedentemente*), diathesis alternation (*decisioni importanti sono state prese*) may cause problems in the identification of VMWEs. Also, VID seems to be quite problematic (66 cases): our intuition about this type of disagreement is that some VIDs are not considered sufficiently established in the common vocabulary such as *mettere su pignatta* but also because it is often challenging to distinguish VIDs when only one dependent of the head verb is lexicalized or when they occur in variants which, as already stated, might cause overlooking VMWEs and inattentions in the annotation.

8. Linguistic Observations

In this section, we discuss some linguistic observations on IRVs and VIDs, which are very productive categories, and a comparison between LVC and IAV, as their categorization may rise some borderline cases. Even though it is an interesting phenomenon, we do not offer a deep analysis on the status of VPCs in Italian since the number of occurrences in the PARSEME-It VMWE corpus is not so high and, therefore, not representative. In fact, as a Romance language, Italian was expected not to exhibit VPCs, but several dozens of VPC annotations do occur in the corpus, e.g., *volata via* (lit. 'flew part' $\rightarrow$ slipped away), *tira fuori* (lit. 'he pulls part' $\rightarrow$ he shows), or *va avanti* (lit. 'go part' $\rightarrow$ go on). This shows the possibly ambiguous status of the element co-occurring with the verb, that is, in previous examples, *via* (by/away), *avanti* (on/forward), *fuori* (out/outside), which can be either adverbs or particles, triggering the VID or the VPC category, respectively. These constructions require to be examined more closely, thus a higher number of occurrences in the corpus is required.

8.1 Very productive VMWEs: IRVs and VID

As described in Monti et al. (2018), IRVs and VID represent very productive categories in Italian which pose some classification issues due to their specific characteristics.

With reference to **IRVs**, the first source of ambiguity in the annotation process is the presence of the clitic pronoun *si* in that in Italian it may be used in three types of different constructions: i) reflexive, ii) impersonal, iii) inherent.

In order to exclude from the annotation reflexive verbs as IRVs, we consider that in reflexive constructions, the clitic pronoun *si* marks the reflexive or reciprocal construction, that is, the clitic plays the role of *self* in English and 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).

To prevent the annotation of impersonal constructions, not belonging to the IRV class, we observe that in these cases the clitic *si* co-occurs with either an intransitive verb or a transitive verb in third person singular. In these occurrences, both classes, originally presenting one or two arguments, reduce their usual number of valency slots to zero, namely they present an empty subject slot, as they convey an absolute and universal meaning expressed by a generic and underspecified subject, e.g., *si muore* (lit. *dies itself $\rightarrow$ dying), *si pensa* (lit. *thinks itself $\rightarrow$ thinking).

Furthermore, as already stated previously, inherent uses of the pronoun *si* are annotated as IRVs, if the verb without the clitic does not exist, e.g., *vergognarsi* (to feel ashamed), or if the verb without the clitic does exist and conveys a very different meaning, e.g., *raffreddarsi* (to get a cold), *raffreddare* (to cool down).

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.

Among sources of mis-classification of IRVs, we notice that the presence of unaccusative constructions (Perlmutter 1978) may generate ambiguity. In fact, in these occurrences, formed through a pseudo-reflexive construction, the clitic, usually representing an overt marker of reduced transitivity, e.g., *seder**si*** (to sit down), is not marked by the accusative case. Unaccusative verbs may be distinguished by applying both semantic and syntactic criteria. Semantically, unaccusative verbs are characterized in that their meaning stands for a change of state, in other words these verbs express telicity, as *sedersi*. From a syntactic point of view, these verbs select a specific temporal auxiliary verb, that is they combine with *be*, while unergative constructions use the verb *have*.

In some cases, IRVs occur in idiomatic constructions 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 VIDs may have both a literal and idiomatic meaning and the boundaries between them are difficult to trace. For instance, *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 (2014c), it is possible to classify ordinary-verb VIDs, namely VIDs 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. The fixed dependent can be of different types:

- Subject, e.g., *un uccellino mi ha detto* (a bird told me)
- Direct object, e.g., *tirare le cuoia* (kick the bucket)
- Circumstantial or adverbial complement, e.g., *prendere qualcosa con le pinze* (to take something with a pinch of salt)

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' → 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, production of causative constructions, possible deletion of the support verb which causes complex nominalizations (Vietri 2014a).

8.2 Borderline cases: LVC and IAV compared

During the annotation process, other borderline cases were identified in two categories, namely LVC and IAV[29], used in the second edition of the shared task.

As previously stated, the former, already annotated in the first edition of the task, has been modified to account for a more fine-grained distinction, i.e., it has been split into LVC.full and LVC.cause.

On one hand, **LVC.full** accounts for occurrences in which the verb contributes to the MWE meaning in that it bears only morphological features, namely person, number, tense, mood, as well as morphological aspect. This implies that the syntactic subject of the verb is the semantic argument of the noun. Such a definition of LVC is different from the one usually proposed by many authors (Hopper and Traugott 2003; Hacker 1958; Hook 1991, 1993) for two main aspects. At first, we do not include aspectual support verbs, unless the aspect is morphological. We do not consider aspectual verbs contributing to a change of the MWE meaning, (e.g., *iniziare* → to start) since, despite the fact they are very productive, they do not form interesting VMWEs (Savary et al. 2017). Therefore, we annotate occurrences in which a predicative noun, e.g., *passeggiata* (walk), co-occurs with a light verb, e.g., *fare*, such in *fare una passeggiata* (have a walk), nevertheless discarding occurrences with aspectual verbs, e.g., *iniziare una passeggiata* (to start a walk). Then, in addition to the standard definition, we take into account also verbs presenting a light semantics per se, which are not considered bleached support verbs. In this perspective, the occurrence *commettere un suicidio* (to commit a suicide) passes all tests and may be accounted as an LVC.full in that it preserves its meaning defined by the presence of any negatively charged achievement noun, e.g. suicide, crime, fraud, felony.

29 This section is partially based on the analysis presented by Caruso in Monti et al. (2018).

On the other hand, **LVC.cause** constructions, expected to be less idiomatic than other VMWEs, can be understood as complex predicates with a causal support verb[30]. In these occurrences, the verb is considered causative when the subject of V is the cause or the main source of the event or state expressed by the noun, e.g., *dare il mal di testa* (to give a headache). LVC.cause constructions may involve:

- verbs that are typically used to express the cause of predicative nouns in general (e.g., *cause, provoke*)
- verbs that are only used to express the cause of particular predicative nouns (e.g., *grant* in *to grant a right*).

Some new tests have been added to account for these subcategories, which heavily rely on the notion of semantic arguments. These tests aim at distinguishing cases in which: (i) the noun is predicative; (ii) the verb's subject is a noun's semantic argument; (iii) the verb presents a light semantics; (iv) the verb reduction is applicable; (v) the verb's subject is the noun's cause.

As already described, **IAVs** are a special optional and experimental category, and correspond to what is also sometimes called in English prepositional verbs, as they consist of a verb or VMWE and an idiomatic selected preposition or postposition. Since in some cases the idiomatic adpositional valency, namely when the co-occurrence of a verb with an adposition opens a slot for an argument, may be mistaken with verb-particle constructions, a language-specific test, mainly concerning English and German, has been provided. Generally speaking, these two phenomena can be distinguished by analyzing the adposition behaviour. If it can occur after the object, e.g., *to wake somebody up*, then the adposition is a particle and the group can not be categorized as IAV. If the adposition cannot occur after the object, as in **to come something across*, then the MWE belongs to the IAV category.

During the annotation trial, the IAV category has proved to be advantageous to cover a rich inventory of VMWEs in Italian, but some issues have also emerged, particularly with respect to the LVC verbs, which also account for combinations of verbs plus prepositions. Prototypical examples of IAV collected so far include the following:

- *Tendere a* + N (to be inclined to something), base form *tendere* (to stretch), e.g., *Maria tende alla depressione* (Maria tends to be depressed);
- *Tendere a* + V (to be inclined to something), e.g., *Maria tende a dimagrire* (Maria tends to loose weight);
- *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 cannot be omitted in questions, thus proving to be part of the verb, as in the following example.

11. Maria **tende** sempre **ad** esagerare. (*Maria always tends to overstate*)
A cosa **tende**, scusa? *(What does she tend to?)*

30 https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/?page=050_Cross-lingual_tests/020_Light-verb_constructions__LB_LVC_RB_

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[31], is realized when "the complements carry information which acts on the governing verb, essentially taking the verb as an argument and shifting its event type" (Pustejovsky 1995). For instance, *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 panic), the verb acquires the aspectual meaning of *to go into the state X*, and cannot 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. Therefore, a more fine-grained analysis is needed in order to annotate these categories appropriately, and capture significant semantic differences. As a counter-example, giving evidence to the broad coverage of the IAV class, one can refer to *portare a* (carry/bring to), because its causative semantics, derived from an original spatial meaning, remains unchanged in different lexical and syntactic contexts. Both with nouns denoting a state (e.g., *portare alla follia*, lit. 'to bring someone to madness'), with those referring to events (*portare a ebollizione*, lit. 'to bring something to boiling point'), and with full-sentence arguments (*portare a conoscere*, lit. 'to bring someone know something') *portare a* preserves its causative meaning.

9. Conclusion and future work

In this paper, we described a linguist resource of Italian VMWEs, developed within the PARSEME Shared Task on Automatic Identification of VMWEs. To the best of our knowledge, PARSEME-It represents the first annotated corpus for Italian VMWEs. Firstly, we introduced current works focused on MWE processing from different perspectives, i.e., linguistic studies and NLP applications. Then, we described aims and methodologies used within the PARSEME Cost Action to define the research objects and to identify such linguistic phenomena. Subsequently, we described the development of the PARSEME-It VMWE corpus and the VMWE categories we took into account within the framework of the PARSEME Shared Task on Automatic Identification of VMWEs (Savary et al. 2017; Ramisch et al. 2018).
Then, we discussed the annotation guidelines together with the identification tests and the category decision trees applied to identify and classify VMWEs. The PARSEME-It VMWE corpus is based on a selection of texts, formed by approx. 16,000 sentences (corresponding to 430,789 tokens) taken from the *PAISÀ* corpus of Italian web texts. The annotation process together with the IAA is presented. A deep analysis of the issues arisen during the double-annotation task shows the disagreement cases in IAA scores. Several sources of disagreement have been identified, namely partial matches labeled, exact matches unlabeled, partial matches unlabeled, and finally VMWE annotations by only one annotator. Yet, among the annotated occurrences, we proposed an analysis of productive categories, i.e., IRVs and VIDs, and a comparison of LVC and IAV categories.

31 Co-composition has been called *accommodation* in more recent works (Pustejovsky 2013).

Due to the high complexity of this type of phraseological units, we consider this work an initial contribution for elaborating an Italian universal terminology of VMWEs, which could ease the challenge of MWE automatic processing, in particular verbal ones.
Furthermore, the analysis of these linguistic phenomena could represent the foundation for semantic representation, suitable to encompass cross-lingual comparisons and applications.
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, increasing the quality of multilingual dictionaries and allowing their full integration into emerging language technologies (LTs). These technologies are based on a semantic formalized representation, which encodes several levels of linguistic information, suitable to guarantee the interoperability among resources from different sources and languages.
The properties of verbal multiword expressions in Italian may contribute to improving their semantic representation according to W3C standards used in current LTs, namely the OntoLex Lemon model[32]. This model aims at providing a rich linguistic grounding for ontologies, including the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e., the meaning of these lexical entries with respect to an ontology or vocabulary (McCrae et al. 2017). The use of this type of formalization to describe linguistic data and resources represents a straight way to contribute to the development of a Linguistic Linked Open Data (LLOD) cloud[33], creating, sharing, and (re-)using language resources in accordance with Linked Data principles (Bizer, Heath, and Berners-Lee 2008).

Acknowledgments

This work has been partially supported by Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 - Fondo Sociale Europeo, Azione I.2 "Attrazione e Mobilità Internazionale dei Ricercatori" Avviso D.D. n 407 del 27/02/2018 and by the IC1207 PARSEME COST action (2013-2017).
We are particularly grateful to Federico Sangati who always supported the annotation team and actively took part in the planning and the implementation of the project. Finally, our thanks go also to all the Italian annotators, which took part in editions 1.0 and 1.1 of the PARSEME shared task on verbal MWE identification, namely Valeria Caruso, Manuela Cherchi, Anna De Santis, Antonio Pascucci, Annalisa Raffone, and Anna Riccio.
Authorship contribution is as follows: Johanna Monti is author of sections 1, 3, 4, 5 and 7. Maria Pia di Buono is author of sections 2, 6, 8 and 9. Abstract is in common.

References

Alba-Salas, Josep. 2002. *Light Verb Constructions in Romance: A syntactic analysis*. Ph.D. thesis, Cornell University, NYC, New York.
Alba-Salas, Josep. 2004. Fare light verb constructions and Italian causatives: Understanding the differences. *Italian Journal of Linguistics*, 16(2):283.
Baldwin, Timothy. 2006. Compositionality and multiword expressions: six of one, half a dozen of the other? In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, July.
Baldwin, Timothy and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRCPress, Boca Raton, USA, pages 267–292.

32 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification
33 https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data

Barreiro, Anabela, Johanna Monti, Brigitte Orliac, Susanne Preuß, Kutz Arrieta, Wang Ling, Fernando Batista, Isabel Trancoso, et al. 2014. Linguistic evaluation of support verb constructions by openlogos and google translate. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 35–40, Reykjavik, Iceland, May.

Berruto, Gaetano. 1987. *Sociolinguistica dell'italiano contemporaneo*, volume 33. Carocci, Roma.

Bizer, Christian, Tom Heath, and Tim Berners-Lee. 2008. Linked Data: Principles and State of the Art. In *17th International World Wide Web Conference*, volume 1, page 40, Beijing, China, April.

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker, and Mark Harvey, editors, *Complex predicates in cross-linguistic perspective*. Cambridge University Press Cambridge, MA, Cambridge, pages 48–78.

Cap, Fabienne, Manju Nirmal, Marion Weller, and Sabine Schulte Im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28, Denver, Colorado, June.

Carstea-Romascanu, Mihaela. 1977. I tipi di verbi riflessivi in italiano. *Revue Roumaine de Linguistique Bucuresti*, 22(2):125–130.

Cicalese, Anna. 1999. Le estensioni di verbo supporto. uno studio introduttivo. *Studi italiani di linguistica teorica ed applicata*, 28(3):447–485.

Cicalese, Anna, Emilio D'Agostino, Alberto Maria Langella, and Ilaria Villari. 2016. Els verbs locatius com a variants de verbs de suport. *Quaderns d'Italià*, 21(21):153–166.

Cinque, Guglielmo. 1988. On si constructions and the theory of arb. *Linguistic inquiry*, 19(4):521–581.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Constant, Matthieu, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative Strategies to Integrate Multiword Expression Recognition and Parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 204–212, Jeju, Republic of Korea. Association for Computational Linguistics.

Copestake, Ann. 2003. Compounds revisited. In *Proceedings of the 2nd International Workshop on Generative Approaches to the Lexicon*, pages 129–154, Geneva, Switzerland, June.

Cordin, Patrizia. 2001. I pronomi riflessivi. In Lorenzo Renzi, Giampaolo Salvi, and Anna Cardinaletti, editors, *Grande grammatica italiana di consultazione*, volume 1. Il Mulino, Bologna, pages 607–17.

D'Agostino, Emilio and Annibale Elia. 1998. Il significato delle frasi: un continuum dalle frasi semplici alle forme polirematiche. In Federico Albano Leoni, Daniele Gambarara, Stefano Gensini, Franco Lo Piparo, and Raffaele Simone, editors, *Ai limiti del linguaggio*. Laterza, Bari, pages 287–310.

de Caseli, Helena Medeiros, 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.

De Mauro, Tullio. 2000. *Grande dizionario italiano dell'uso (GRADIT)*. Utet, Torino.

Gagné, Christina L and Thomas L Spalding. 2009. Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, 60(1):20–35.

Gibbs, Raymond W, Nandini P Nayak, John L Bolton, and Melissa E Keppel. 1989. Speakers' assumptions about the lexical flexibility of idioms. *Memory & cognition*, 17(1):58–68.

Hacker, Paul. 1958. *Zur Funktion Einiger Hilfsverben im Modernen Hindi*. Verlag der Akademie der Wissenschaften und der Literatur in Mainz, München.

Hook, Peter Edwin. 1991. The Emergence of Perfective Aspect in Indo-Aryan languages. *Approaches to grammaticalization*, 2:59–89.

Hook, Peter Edwin. 1993. Aspectogenesis and the Compound Verb in Indo-Aryan. In Manindra K. Verma, editor, *Complex predicates in South Asian languages*. Manohar Publishers & Distributors, New Delhi, pages 97–113.

Hopper, Paul J and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge University Press, Cambridge.

Iacobini, Claudio and Francesca Masini. 2005. Verb-particle constructions and prefixed verbs in Italian: typology, diachrony and semantics. In *Mediterranean Morphology Meetings*, volume 5, pages 157–184, Fréjus, France, September.

Im Walde, Sabine Schulte, Stefan Müller, and Stefan Roller. 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, June.

Jackendoff, Ray. 1997. *The architecture of the language faculty*. MIT Press, Cambridge.

Jezek, Elisabetta. 2004. Types et degrés de verbes supports en italien. *Lingvisticæ Investigationes*, 27(2):185–201.

Karimi-Doostan, Gholamhossein. 1997. *Light Verb Constructions in Persian*. Ph.D. thesis, University of Essex, Colchester, Essex, UK.

Kordoni, Valia and Iliana Simova. 2014. Multiword expressions in machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1208–1211, Reykjavik, Iceland, May.

La Fauci, Nunzio. 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.

Laranjeira, Bruno, Viviane Pereira Moreira, Aline Villavicencio, Carlos Ramisch, and Maria José Bocorny Finatto. 2014. Comparing the quality of focused crawlers and of the translation resources obtained from them. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3572–3578, Reykjavik, Iceland, May.

Lenci, Alessandro, Gianluca Lebani, Sara Castagnoli, Francesca Masini, and Malvina Nissim. 2014. Sympathy: Towards a comprehensive approach to the extraction of Italian word combinations. In *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it*, pages 234–238, Pisa, December. Pisa University Press.

Lyding, Verena, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISÀ corpus of Italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 36–43, Gothenburg, Sweden, April.

Lyse, Gunn Inger and Gisle Andersen. 2012. Collocations and statistical analysis of n-grams. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, volume 49. John Benjamins Publishing, Amsterdam/New York, pages 79–109.

Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors. 2017. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain, April. Association for Computational Linguistics.

Markantonatou, Stella, Carlos Ramisch, Agata Savary, and Veronika Vincze. 2018. *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Language Science Press, Berlin, Germany.

Masini, Francesca. 2005. Multi-word expressions between syntax and the lexicon: The case of Italian verb-particle constructions. *SKY Journal of Linguistics*, 18(2005):145–173.

Masini, Francesca. 2015. Idiomatic verb-clitic constructions: Lexicalization and productivity. In *Proceedings of Mediterranean Morphology Meetings*, volume 9, pages 88–104, Haifa, Israel, September.

Mateu, Jaume and Gemma Rigau. 2010. Verb-particle constructions in Romance: A lexical-syntactic account. *Probus*, 22(2):241–269.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and Applications. In *Proceedings of eLex 2017 conference*, pages 19–21, Leiden, Netherlands, September.

Mitkov, Ruslan, Johanna Monti, Gloria Corpas Pastor, and Violeta Seretan. 2018. *Multiword Units in Machine Translation and Translation Technology*, volume 341. John Benjamins Publishing Company, Amsterdam/New York.

Monti, Johanna, Anabela Barreiro, Brigitte Orliac, and Fernando Batista. 2013. When Multiwords Go Bad in Machine Translation. In *Machine Translation Summit XIV*, pages 26–33, Nice, France, September. The European Association for Machine Translation.

Monti, Johanna, Valeria Caruso, and Maria Pia di Buono. 2018. PARSEME-It-Issues in verbal Multiword Expressions identification and classification. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253, Torino, Italy, December. Accademia University Press.

Monti, Johanna, Silvio Cordeiro, Carlos Ramisch, Federico Sangati, Agata Savary, and Veronika Vincze. 2018. Advances in Multiword Expression Identification for the Italian language: The

PARSEME shared task edition 1.1. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253, Torino, Italy, December. Accademia University Press.

Monti, Johanna, 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, Rome, Italy, December. Accademia University Press.

Monti, Johanna, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan, editors. 2013. *Workshop Proceedings. Multi-Word Units in Machine Translation and Translation Technologies. MUMTTT 2013*, Switzerland, September. Tradulex.

Monti, Johanna, Mitkov Ruslan, Seretan Violeta, and Gloria Corpas Pastor. 2018. *Proceedings of the 3rd Workshop on Multi-Word Units in Machine Translation and Translation Technology (MUMTTT 2017)*. Tradulex, Switzerland, November.

Newman, David, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian Text Segmentation for Index term Identification and Keyphrase Extraction. In *Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics: Technical Papers*, pages 2077–2092, Mumbai, India, December.

Ninio, Anat. 2011. *Syntactic development, its input and output*. Oxford University Press, Oxford, UK.

Nissim, Malvina, Sara Castagnoli, and Francesca Masini. 2014. Extracting MWEs from Italian corpora: A case study for refining the pos-pattern methodology. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics - EACL 2014*, pages 57–61, Gothenburg, Sweden, April.

Nunberg, Geoffrey, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Ó Séaghdha, Diarmuid. 2008. Learning compound noun semantics. Technical report, University of Cambridge, Computer Laboratory, Cambridge.

Pal, Santanu, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013. A hybrid word alignment model for phrase-based statistical machine translation. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 94–101, Sofia, Bulgaria, August.

Pastor, Gloria Corpas, Ruslan Mitkov, Maria Kunilovskaya, and María Araceli Losey León, editors. 2019. *Computational and Corpus-based Phraseology Proceedings of the Third International Conference EUROPHRAS 2019 (short papers, posters and MUMTTT workshop contributions)*, Switzerland, September. Tradulex.

Pastor, Gloria Corpas, Johanna Monti, Violeta Seretan, and Ruslan Mitkov, editors. 2015. *Workshop Proceedings. Multi-Word Units in Machine Translation and Translation Technologies. MUMTTT 2015*, Switzerland, July. Tradulex.

Perlmutter, David M. 1978. Impersonal Passives and the Unaccusative Hypothesis. In *Proceedings of the 4th Annual Meeting of the Berkeley Linguistics Society*, volume 4, pages 157–190, Berkeley, California, February.

Pescarini, Diego. 2015. Costruzioni con si: una classificazione razionale? *Grammatica e Didattica*, pages 15–32.

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

Pustejovsky, James. 2013. Type theory and lexical decomposition. In James Pustejovsky, Pierrette Bouillon, Hitoshi Isahara, Kyoko Kanzaki, and Chungmin Lee, editors, *Advances in generative lexicon theory*. Springer, Berlin, Germany, pages 9–38.

Quaglia, Stefano and Andreas Trotzke. 2017. Italian verb particles and clausal positions. In *Proceedings of The 31st annual meeting Israel Association for Theoretical Linguistics - IATL 31*, pages 67–82, Ramata Gan, Israel, October.

Quochi, Valeria. 2007. *A usage-based approach to light verb constructions in Italian: Development and use*. Ph.D. thesis, University of Pisa, Pisa.

Ramisch, Carlos, Laurent Besacier, and Alexander Kobzar. 2013. How hard is it to automatically translate phrasal verbs from English to French. In *Proceedings of the MT Summit 2013 Workshop on Multi-word Units in Machine Translation and Translation Technology*, pages 53–61, Nice, France, September.

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal

Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA, August.
Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54, Singapore, August. Association for Computational Linguistics.
Reuland, Eric. 1990. Reflexives and Beyond: Non-local Anaphora in Italian Revisited. *Grammar in progress: glow essays for Henk van Riemsdijk*, 36:351.
Rikters, Matīss and Ondřej Bojar. 2017. Paying attention to multi-word expressions in neural machine translation. *arXiv preprint arXiv:1710.06313*.
Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15, Mexico City, Mexico, February.
Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2016. Determining the Multiword Expression Inventory of a Surprise Language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 471–481, Osaka, Japan, December.
Savary, Agata and Silvio Cordeiro. 2018. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 16)*, pages 64–72, Prague, Czech Republic, January.
Savary, Agata, 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, Valencia, Spain, April.
Savary, Agata, Carlos Ramisch, Jena D. Hwang, Nathan Schneider, Melanie Andresen, Sameer Pradhan, and Miriam R. L. Petruck, editors. 2018. *Proceedings of the Joint Workshop on Linguistic Annotation, MultiWord Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
Savary, Agata, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, et al. 2015. Parseme–parsing and multiword expressions within a European multilingual network. In *Proceedings of the 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November.
Shapiro, Naomi Tachikawa. 2016. Splitting compounds with ngrams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 630–640, Osaka, Japan, December.
Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2019. Verbal multiword expressions: Idiomaticity and flexibility. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions*. Language Science Press, Berlin, Germany, pages 35–68.
Simone, Raffaele. 1997. Esistono verbi sintagmatici in italiano? In Tullio De Mauro, editor, *Lessico e grammatica. Teorie linguistiche e applicazioni lessicografiche - Atti del Convegno interannuale della Società di linguistica italiana*, pages 155–170, Madrid.
Stymne, Sara, Nicola Cancedda, and Lars Ahrenberg. 2013. Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4):1067–1108.
Tabossi, Patrizia, Lisa Arduino, and Rachele Fanari. 2011. Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods*, 43(1):110–123.
Taslimipoor, Shiva, Anna de Santis, Johanna Monti, et al. 2016. Language resources for Italian: towards the development of a corpus of annotated Italian multiword expressions. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 285–290, Napoli, Italy, December. Accademia University Press.
Venkatapathy, Sriram and Aravind Joshi. 2006. Using Information about Multi-Word Expressions for the Word-Alignment Task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27, Sydney, Australia, July.

Vietri, Simonetta. 2014a. *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*, volume 31. John Benjamins Publishing Company, Amsterdam/New York.
Vietri, Simonetta. 2014b. The Italian module for Nooj. In *Proceedings of the First Italian Conference on Computational Linguistics, CLiC-it*, pages 389–393, Pisa, Italy, December.
Vietri, Simonetta. 2014c. The Lexicon-Grammar of Italian idioms. In *Workshop on Lexical and Grammatical Resources for Language Processing, COLING 2014*, pages 137–146, Dublin, Ireland, August.
Villavicencio, Aline, Timothy Baldwin, and Benjamin Waldron. 2004. A Multilingual Database of Idioms. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May.
Viviani, Andrea. 2006. *I verbi procomplementari tra grammatica e lessicografia*. Le Lettere, Firenze, Italy.
Wehrli, Eric and Aline Villavicencio. 2015. Extraction of Multilingual MWEs from Aligned Corpora. In *PARSEME 5th General Meeting*, Iași, Romania, September.
Zaninello, Andrea and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France, May. European Language Resources Association.
Zaninello, Andrea and Malvina Nissim. 2010. Creation of lexical resources for a characterisation of multiword expressions in Italian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, La Valletta, Malta, May.