PARSEME multilingual corpus of verbal multiword expressions

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Cebiroğlu Eryiğit, Voula Giouli, Maarten van Gompel, et al.

Chapter 4

PARSEME multilingual corpus of verbal multiword expressions

Agata Savary[1], Marie Candito[2], Verginica Barbu Mititelu[3], Eduard Bejček[4], Fabienne Cap[5], Slavomír Čéplö[6], Silvio Ricardo Cordeiro[7], Gülşen Eryiğit[8], Voula Giouli[9], Maarten van Gompel[10], Yaakov HaCohen-Kerner[11], Jolanta Kovalevskaitė[12], Simon Krek[13], Chaya Liebeskind[11], Johanna Monti[14], Carla Parra Escartín[15], Lonneke van der Plas[6], Behrang QasemiZadeh[16], Carlos Ramisch[7], Federico Sangati[17], Ivelina Stoyanova[18] & Veronika Vincze[19]

[1]Université de Tours (France), [2]Université Paris Diderot (France), [3]Romanian Academy Research Institute for Artificial Intelligence (Romania), [4]Charles University (Czech Republic), [5]Uppsala University (Sweden), [6]University of Malta (Malta), [7]Aix Marseille University (France), [8]Istanbul Technical University (Turkey), [9]Athena Research Center in Athens (Greece), [10]Radboud University in Nijmegen (Netherlands), [11]Jerusalem College of Technology (Israel), [12]Vytautas Magnus University in Kaunas (Lithuania), [13]Jožef Stefan Institute in Ljubljana (Slovenia), [14]"L'Orientale" University of Naples (Italy), [15]ADAPT Centre, Dublin City University (Ireland), [16]University of Düsseldorf (Germany), [17]independent researcher (Italy), [18]Bulgarian Academy of Sciences in Sofia (Bulgaria), [19]University of Szeged (Hungary)

Multiword expressions (MWEs) are known as a "pain in the neck" due to their idiosyncratic behaviour. While some categories of MWEs have been largely studied, verbal MWEs (VMWEs) such as *to take a walk*, *to break one's heart* or *to turn off* have been relatively rarely modelled. We describe an initiative meant to bring about substantial progress in understanding, modelling and processing VMWEs. In this joint effort carried out within a European research network we elaborated

Agata Savary et al. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop, 87-147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471591

a universal terminology and annotation methodology for VMWEs. Its main outcomes, available under open licenses, are unified annotation guidelines, and a corpus of over 5.4 million words and 62 thousand annotated VMWEs in 18 languages.

1 Introduction

One of the basic ideas underlying linguistic modelling is compositionality (Baggio et al. 2012), seen as a property of language items (Janssen 2001; Partee et al. 1990) or of linguistic analyses (Kracht 2007). Counterexamples which challenge the compositionality principles (Pagin & Westerståhl 2001) include multiword expressions (MWEs) (Sag et al. 2002; Kim 2008), and notably verbal MWEs (VMWEs), such as (1–4).[1]

(1) *Ida **skriva glavo v pesek**.* (SL)
Ida hide.3.SG head in sand
Ida hides her head in the sand. 'Ida pretends not to see a problem.'

(2) *Er **legt** die Prüfung **ab**.* (DE)
he lay.3.SG the exam PART
He lays the exam PART. 'He takes the exam.'

(3) *Η Ζωή **παίρνει** μία **απόφαση**.* (EL)
i zoi perni mia apofasi
the Zoe take.3.SG a decision
Zoe takes a decision. 'Zoe makes a decision.'

(4) *Alina **se face** doctor.* (RO)
Alina REFL.3.SG make.3.SG doctor
Alina REFL makes doctor. 'Alina becomes a doctor.'

VMWEs pose special challenges in natural language processing (NLP):

1. SEMANTIC NON-COMPOSITIONALITY: The meaning of many VMWEs cannot be deduced in a way deemed grammatically regular on the basis of their syntactic structure and of the meanings of their components. For instance, the meaning of sentence (1) cannot be retrieved from the meanings of its component words (SL) *glava* 'head' and *pesek* 'sand', except when very specific interpretations of these words and of their combination are admitted.

[1]See the preface for the description of the conventions used to present multilingual examples.

2. Lexical and grammatical inflexibility: VMWEs are frequently subject to unpredictable lexical or syntactic constraints. For instance, when the individual lexemes in (EN) *to **throw** somebody **to the lions*** are replaced by their synonyms or the noun is modified by an adjective, the expression loses its idiomatic meaning:[2] (EN) *#to fling sb to the lions*, *#to throw sb to the hungry lions*. Similarly, the predicative noun in the light-verb construction (EN) *she **took** a **glance** at the headline* cannot take a modifier denoting an agent, especially if different from the verb's subject (**she **took** Paul's **glance** at the headline*).

3. Regular variability: Despite this inflexibility the VMWEs can still exhibit some regular variability, e.g.: (i) inflection or passivisation, as in (EN) *he was **thrown to the lions***, (ii) a restricted lexical replacement and an adjectival modification of the predicative noun, as in (EN) *he **took/had** a quick **glance** at the headline*, (iii) omission of components without change in meaning, as in (EL) ***meno me ti glika (sto stoma)*** 'I stayed with the sweetness (in.the mouth)' ⟹ 'I was very close to enjoy something desired but I failed to'.

4. Discontinuity: The components of a VMWE may not be adjacent, e.g. (EN) *a **mistake** was frequently **made**, never **turn** it **off***.

5. Categorical ambiguity: VMWEs of different categories may share the same syntactic structure and lexical choices. For instance, (EN) *to **make** a **mistake*** and (EN) *to **make a meal** of something* 'to treat something as more serious than it really is' are combinations of the same verb with a direct object but the former is a light-verb construction (since the verb is semantically void and the noun keeps its original predicative meaning), while the latter is an idiom (since the noun loses its original sense).

6. Syntactic ambiguity: Occurrences of VMWEs in text may be syntactically ambiguous, e.g. (EN) *on* is a particle in *to **take on** the task* 'to agree to be in charge of the task', while it is a preposition in (EN) *to **sit on the fence*** 'not to take sides in a dispute'.

7. Literal-idiomatic ambiguity: A VMWE may have both an idiomatic and a literal reading. For instance the VMWE (EN) *to **take the cake*** 'to be the

[2]Henceforth, an asterisk (*) preceding a sentence will mean that the sentence is ungrammatical, while a dash (#) will signal a substantial change in meaning with respect to the original expression.

most remarkable of its kind' is understood literally in (EN) *to take the cake out of the fridge*.

8. Non-literal translatability: Word-for-word translation of VMWEs is usually incorrect, e.g. (EN) *to* ***take the cake*** 'to be the most remarkable of its kind' does not translate to (FR) *prendre le gâteau* 'to take the cake'.

9. Cross-language divergence: VMWEs behave differently in different languages and are modelled according to different linguistic traditions. For instance, functional tokens, such as (EN) *off*, have a status of stand-alone words and can form verb-particle constructions in Germanic languages, e.g. (EN) *to* ***turn off***. In Slavic languages, conversely, they function as prefixes, as in (PL) *wyłączyć* 'PART.connect' ⟹ 'turn off', and are seen as inherent parts of verbal lexemes. Therefore, they cannot trigger MWE-related considerations (cf. §8). Also, the scope of light (or support) verb constructions may greatly vary from one linguistic tradition to another, e.g. depending on whether the copula *to be* is considered a light verb or not (cf. §9.1).

10. Wordplay proneness: In particular contexts, VMWEs can be a subject of ad hoc creativity or a playful usage, as in (EN) *they want us to put the cat back inside the bag* 'they want us to pretend that the revealed secret remains unrevealed'.

Due to these unpredictable properties, the description, identification, analysis and translation of VMWEs require dedicated procedures. For example, due to 2 and 3, the description of VMWEs can be constrained neither to the level of the lexicon nor to the one of the syntax only. Challenge 4 hinders VMWE identification with traditional sequence labelling approaches and calls for syntactic analysis. Challenges 5, 6 and 7, however, mean that their identification and categorisation cannot be based on solely syntactic patterns. Challenges 1, 2, 7 and 8 constitute central issues in machine translation. Challenge 9 affects cross-lingual VMWE modelling. Finally, challenge 10 goes far beyond the state of the art in semantic modelling and processing of VMWEs.

A consistent linguistic and NLP terminology is required in order to better understand the nature of VMWEs, compare their properties across languages, hypothesise linguistic generalisations, model VMWEs according to common principles, develop cross-language VMWE identifiers and compare results obtained by different authors on different datasets. Such a consistency is, however, largely missing: different authors assign different names to the same phenomena or call different phenomena by the same name, be it from a linguistic or an NLP point

of view. This situation is similar to other areas of linguistic modelling, where universalism-driven efforts have been undertaken – such as the Universal Dependencies (UD) project dedicated to standardising morphological and syntactic annotations for dozens of languages (Nivre et al. 2016), or the normalisation of uncertainty cue annotation across languages, genres and domains (Szarvas et al. 2012).

This chapter describes an initiative taken by the European PARSEME network,[3] towards bringing about substantial progress in modelling and processing MWEs. Its main outcomes include unified definitions and annotation guidelines for several types of VMWEs, as well as a large multilingual openly available VMWE-annotated corpus. Eighteen languages are addressed (note that the last 4 are non-Indo-European):

- *Balto-Slavic*: Bulgarian (BG), Czech (CS), Lithuanian (LT), Polish (PL) and Slovene (SL);
- *Germanic*: German (DE) and Swedish (SV);
- *Romance*: French (FR), Italian (IT), Romanian (RO), Spanish (ES) and Portuguese (PT);[4]
- *Others*: Farsi (FA), Greek (EL), Hebrew (HE), Hungarian (HU), Maltese (MT) and Turkish (TR).

The corpus gave rise to the PARSEME shared task on automatic identification of VMWEs, whose organisation and results are described by Savary et al. (2017). See also Taslimipoor et al. (2018 [this volume]) and Maldonado & QasemiZadeh (2018 [this volume]) who address the use of the PARSEME corpus in VMWE identification and its evaluation, as well as Moreau et al. (2018 [this volume]), Al Saied et al. (2018 [this volume]) and Simkó et al. (2018 [this volume]) who describe 3 of the 7 systems participating in the shared task.

This chapter builds upon those sections of the PARSEME shared task description paper (Savary et al. 2017), presented in the MWE 2017 workshop, which describe the corpus construction. Each of these sections has been substantially extended, except the descriptions of the corpus format and inter-annotator agreement, which required few additions and updates. Many new analyses and examples have been added, conclusions drawn from the PARSEME annotation campaign have been addressed and the state of the art has been thoroughly revised. As a result, the chapter is organised as follows. We give the definitions underly-

[3] http://www.parseme.eu

[4] In this chapter we address the Brazilian dialect of Portuguese. All examples cited here are taken from this dialect.

ing the scope of our work (§2), and the VMWE typology (§3). We describe the annotation principles, including the VMWE identification and categorisation tests, and the deviations from the unified guidelines applied in some languages (§4). We discuss the annotation methodology and tools (§5). We present the resulting corpus and a cross-language quantitative analysis of some phenomena relevant to challenges 1–10 (§6). We describe some language-specific studies based on the corpus (§7) and discuss interesting problems which occurred during the project (§8). We analyse the state of the art in MWE modelling and annotation, and compare it to our approach (§9). We finally conclude and discuss future work (§10).

2 Definitions and scope

While the definition of a MWE inherently relies on the notion of a WORD (i.e. a linguistically motivated unit), identification of VMWEs is performed on pragmatically defined TOKENS. The relation between tokens and words can be threefold:

(1) A token coincides with a word, e.g. (MT) *ferħ* 'happiness', (SV) *förvåning* 'surprise'.

(2) Several tokens build up one MULTITOKEN WORD (MTW), if punctuation marks are considered token boundaries, as in (EN) *Pandora's*, (PL) *SMS-ować* 'to write an SMS'. Note that the latter example is not a VMWE as it contains only one word.

(3) One MULTIWORD TOKEN (MWT) contains several words, as in contractions, e.g. (IT) *della* 'of.the', or detachable pre-verbal particles, e.g. (DE) ***ausmachen*** 'PART.make' ⇒ 'to turn off'. Note that the latter example is a (one-token) VMWE. A MWT is not always a simple concatenation of words, e.g. (IT) *della* is a contraction of *di* 'of' and *la* 'the.FEM'.

In this work, MULTIWORD EXPRESSIONS (MWEs) are understood as (continuous or discontinuous) sequences of words which:

- contain at least two component words which are **lexicalised**, i.e. always realised by the same lexemes (see below for a more precise definition), including a head word and at least one other syntactically related word,
- display some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy, formalised by the annotation procedures in §4.1–§4.2.

This definition relatively closely follows the one by Baldwin & Kim (2010). Two notable exceptions are that we impose syntactic constraints on the lexicalised components (one of them must be the head word), and that Baldwin & Kim (2010)

include pragmatic and statistical idiosyncrasy in the set of the MWE definition criteria. For us, conversely, COLLOCATIONS, i.e. word co-occurrences whose idiosyncrasy is of pragmatic or statistical nature only (e.g. *all aboard*, *the graphic shows*, *drastically drop*) are disregarded.

Note that there is no agreement on the understanding of the border between the scopes of MWEs and collocations. For Sag et al. (2002), collocations are any statistically significant word co-occurrences, i.e. they include all forms of MWEs. For Baldwin & Kim (2010), collocations form a proper subset of MWEs. According to Mel'čuk (2010), collocations are binary, semantically compositional combinations of words subject to lexical selection constraints, i.e. they intersect with what is here understood as MWEs. This chapter puts forward yet another point of view: MWEs and collocations are seen as disjoint sets of linguistic objects.

Our definition of a MWE is also relatively close to the notion of non-compositional semantic phrasemes in Mel'čuk (2010), but we include light-verb constructions in our scope. It is compatible as well with the one by Sag et al. (2002), where a MWE is seen as an "idiomatic interpretation that crosses word boundaries". The major differences between our approach and these seminal works are its multilingual context and the fact that, within the restricted scope of verbal MWEs (see below), we delimit the MWE phenomenon by a relatively precise and complete MWE identification and categorisation procedure, given in the form of decision trees built upon linguistic tests (§4). Note that this approach does not focus on another salient property of MWEs which is their variable degree of idiosyncrasy (Gross 1988), that is, the fact that various MWEs exhibit more or less unexpected lexical, syntactic and semantic properties. A scale-wise modelling of MWEs is hard to implement in the task of MWE annotation, which is our major operational objective. Instead, we assume that decisions on MWE-hood are binary, and the decision trees are designed so as to make them reproducible.

VERBAL MWEs (VMWEs) are multiword expressions whose canonical form (see below) is such that: (i) its syntactic head is a verb V, (ii) its other lexicalised components form phrases directly dependent on V. Boundary cases for condition (i) include at least two types of VMWEs. Firstly, those with irregular syntactic structures may hinder the identification of the headword as in (EN) ***short-circuited***, where the verb is atypically prefixed by an adjective. Secondly, for those with two coordinated lexicalised verbs there is no consensus as to which component – the conjunction or the first verb – should be considered the head, as in (5). Condition (ii) requires that the lexicalised components of a VMWE form a connected dependency graph. For instance, in (EN) *to* ***take on*** *the task* 'to agree to be in charge of the task' the particle *on* directly depends on the verb, thus ***take***

on fulfils the syntactic requirements to be a VMWE. Conversely, if the lexicalist hypothesis in syntax is followed (de Marneffe et al. 2014),[5] the preposition *on* in (EN) *to rely on someone* does not directly depend on the verb, thus, *rely on* cannot be considered a VMWE.

(5) *wo man **lebt** und **leben lässt*** (DE)
where one lives and live lets
where one lives and lets live 'where one is tolerant'

Just like a regular verb, the head verb of a VMWE may have a varying number of arguments. For instance, the direct object and the prepositional complement are compulsory in (EN) *to **take** someone **by surprise***. Some components of such compulsory arguments may be LEXICALISED, that is, always realized by the same lexemes. Here, *by surprise* is lexicalised while *someone* is not.

Note that lexicalisation is traditionally defined as a diachronic process by which a word or a phrase acquires the status of an autonomous lexical unit, that is, "a form which it could not have if it had arisen by the application of productive rules" (Bauer 1983 apud Lipka et al. 2004). In this sense all expressions considered VMWEs in this work are lexicalized. Our notion of lexicalisation extends this standard terminology, as it applies not only to VMWEs but to their components as well. The reason is that, in the context of the annotation task, we are in need of specifying the precise span of a VMWE, i.e. pointing at those words which are considered its inherent, lexically fixed components. Precisely these components are referred to as lexicalized within the given VMWE. Throughout this chapter, the lexicalised components of VMWEs are highlighted in bold.

A prominent feature of VMWEs is their rich morpho-syntactic variability. For instance, the VMWE (EN) *to **take** someone **by surprise*** can be inflected (*they **took** him **by surprise***), negated (*they did not **take** him **by surprise***), passivised (*he will be **taken by surprise***), subject to extraction (*the **surprise** by which I was **taken***), etc. Neutralizing this variation is needed when applying the linguistic tests defined in the annotation guidelines (§4), which are driven by the syntactic structure of the VMWE candidates. We define a PROTOTYPICAL VERBAL PHRASE as a minimal sentence in which the head verb *V* occurs in a finite non-negated form and all its arguments are in singular and realized with no extraction. For instance, (EN) *Paul made/makes a pie* is a prototypical verbal phrase while *Paul did not make a pie*, *the pie which Paul made* and *the pie was made by Paul* are

[5]The lexicalist hypothesis strongly inspired the PARSEME annotation guidelines, and is expected to be even more thoroughly followed in the future versions of the corpus.

not. If a VMWE can occur as a prototypical verbal phrase while keeping its idiomatic meaning, then such a phrase is its CANONICAL FORM. Otherwise, its least marked variation is considered canonical (a non-negated form is less marked than a negated one, active voice is less marked then passive, and a form with an extraction is more marked than one without it). For instance, a canonical form of (EN) *a bunch of* ***decisions*** *which were* ***made*** *by him* is (EN) *he* ***made*** *a* ***decision***. But since (6) and (7) lose their idiomatic readings in active voice – (PL) *#wszyscy rzucili kości* 'everyone threw dies' – and with no negation – (BG) *#tya* ***iska i da chue*** 'she wants to also hear' – their canonical forms are passive and negated, respectively. Whenever a VMWE candidate is identified in a sentence, the linguistic tests are to be applied to one of its canonical forms (whether it is a prototypical verbal phrase or not).

(6) ***Kości zostały rzucone***. (PL)
dies were cast
The dies were cast. 'The point of no-return has been passed.'

(7) *Тя* ***не иска и да чуе***. (BG)
Tya **ne iska i da chue**
she not want and to hear
She does not even want to hear. 'She opposes strongly.'

(8) ***Пиле не може да прехвръкне***. (BG)
pile ne mozhe da prehvrakne
Bird not can to PART.fly
A bird cannot fly across something. 'Something is very strictly guarded.'

Throughout this chapter examples of VMWEs will always be given in their canonical forms, possibly accompanied by adjuncts, if the subject is lexicalised as in (8). Otherwise, their canonical forms may alternate – for brevity – with infinitive forms, or – rarely – with other variants when particular phenomena are to be illustrated.

MWEs containing verbs but not functioning as verbal phrases or sentences are excluded from the scope of annotation, e.g. (FR) *peut-être* 'may-be' ⇒ 'maybe', *porte-feuille* 'carry-sheet' ⇒ 'wallet'.

Let us finally comment on the notion of universalism. Formally, this term should only be used when a property or a phenomenon has been proven relevant to all languages, which is practically out of range of any endeavour, however multilingual and inclusive. Therefore, in this chapter we use the adjective 'universal' in the sense of a scientific hypothesis rather than of a proven fact. When

we speak about a universal category or property, it is to be understood that we deem them universal, based on the evidence from the languages currently in our scope. Since our framework is meant to continually evolve by including new languages and MWE types, we hope our definitions and findings to approximate the truly universal properties increasingly well.

3 VMWE typology

The typology of VMWEs, as well as linguistic tests enabling their classification, were designed so as to represent properties deemed universal in a homogeneous way, while rendering language-specific categories and features at the same time. The 3-level typology consists of:

1. *Universal* categories, valid for all languages participating in the task:
   a) light-verb constructions (LVCs), as in (9):

      (9) *Eles* ***deram*** *uma* ***caminhada****.* (PT)
      they gave a walk
      They gave a walk. 'They took a walk.'

   b) idioms (ID), as in (10):

      (10) به قدر کافی برای من خواب دیده است. (FA)
      ast **dide khab** man **baraye** kafi qadre be
      is seen sleep me for enough quantity to
      He had enough sleep for me. 'He has many plans for me.'

2. *Quasi-universal* categories, valid for some language groups or languages, but not all:
   a) inherently reflexive verbs (IReflVs), as in (11):

      (11) *Ils ne* ***s'apercevront*** *de rien.* (FR)
      they not REFL.3.PL'perceive.3.PL.FUT of nothing
      They will REFL-perceive nothing. 'They will not realise anything.'

   b) verb-particle constructions (VPCs), as in (12):

      (12) *Sie* ***macht*** *die Tür* ***auf****.* (DE)
      she makes the door PART
      She makes PART the door. 'She opens the door.'

3. *Other* verbal MWEs (OTH), not belonging to any of the categories above (due to not having a unique verbal head) e.g. (EN) *he never **drinks and drives**, she **voice acted**, the radio **short-circuited**.*

Table 1: Examples of various categories of VMWEs in four non-Indo-European languages.

Lang.	ID	LVC	Quasi-universal / OTH
HE	**אבד עליו כלה** 'Kelax is lost on him.' 'He is outdated.'	**הגיע למסקנה** 'to come to a conclusion' 'to conclude'	**לא הבישן לומד** 'the bashful does not learn' 'one should dare ask questions'
HU	**kinyír** 'to out-cut' 'to kill'	**szabályozást ad** 'to give control' 'to regulate'	**feltüntet** (VPC) 'to PART-strike' 'to mark'
MT	**Għasfur żgħir qalli**. 'A small bird told me.' 'I learned it informally.'	**ħa deċizjoni** 'to take a decision' 'to make a decision'	**iqum u joqgħod** (OTH) 'to jump and stay' 'to fidget'
TR	**yüzüstü bırakmak** 'to leave (sb) face down' 'to forsake'	**engel olmak** 'to become obstacle' 'to prevent'	**karar vermek** (OTH) 'to give a decision' 'to make a decision'

While we allowed for language-specific categories, none emerged so far. Table 1 and Table 2 show examples of VMWEs of different categories in the 18 languages in our scope (4 non-Indo-European and 14 Indo-European). None of those languages seems to possess VMWEs of all 5 terminal categories (LVC, ID, IReflV, VPC and OTH).

We thoroughly considered introducing another universal category of inherently prepositional verbs (IPrepVs), such as (EN) *to rely on*, *to refer to*, or *to come across*. However, the IPrepV-related linguistic tests used in the pilot annotation proved not sufficiently reliable to distinguish such expressions from compositional verb-preposition combinations, such as (EN) *to give something to someone*. Therefore, we abandoned this category, considering that prepositions belong to the area of verb valency and should be handled by a regular grammar (combined with a valency lexicon). Reconsidering this category experimentally belongs to future work (§10).

Table 2: Examples of various categories of VMWEs in 14 Indo-European languages.

Lang.	ID	LVC	Quasi-universal / OTH
BG	**бълвам змии и гущери** 'to spew snakes and lizards' 'to shower abuse'	**държа под контрол** 'to keep under control' 'to keep under control'	**усмихвам се** (IReflV) 'to smile REFL' 'to smile'
CS	**házet klacky pod nohy** 'to throw sticks under feet' 'to put obstacles in one's way'	**vyslovovat nesouhlas** 'to voice disagreement' 'to disagree'	**chovat se** (IReflV) 'to keep REFL' 'to behave'
DE	**schwarz fahren** 'to drive black' 'to take a ride without a ticket'	eine **Rede halten** 'a hold a speech' 'to give a speech'	**sich enthalten** (IReflV) 'to contain REFL' 'to abstain'
EL	**χάνω τα αυγά και τα καλάθια** 'to lose the eggs and the baskets' 'to be at a complete and utter loss'	**κάνω** μία **πρόταση** 'to make a proposal' 'to propose'	**μπαίνω μέσα** (VPC) 'to get PART' 'to go bankrupt'
ES	**hacer de tripas corazón** 'to make heart of intestines' 'to pluck up the courage'	**hacer** una **foto** 'to make a picture' 'to take a picture'	**coser y cantar** (OTH) 'to sew and to sing' 'as easy as pie'
FA	دست گل به آب دادن 'to give a flower bouquet to water' 'to mess up, to do sth. wrong'	امتحان کردن 'to do an exam' 'to test'	به خود آمدن 'to come to REFL' 'to gain focus'
FR	**voir le jour** 'to see the daylight' 'to be born'	**avoir** du **courage** 'to have courage' 'to have courage'	**se suicider** (IReflV) 'to suicide REFL' 'to commit suicide'
IT	**entrare in vigore** 'to enter into force' 'to come into effect'	**fare** un **discorso** 'to make a speech' 'to give a speech'	**buttare giù** (VPC) 'to throw PART' 'to swallow'
LT	**pramušti dugną** 'to break the bottom' 'to collapse'	**priimti sprendimą** 'to take on a decision' 'to make a decision'	
PL	**rzucać grochem o ścianę** 'to throw peas against a wall' 'to try to convince somebody in vain'	**odnieść sukces** 'to carry-away a success' 'to be successful'	**bać się** (IReflV) 'to fear REFL' 'to be afraid'
PT	**fazer das tripas coração** 'make the tripes into heart' 'to try everything possible'	**fazer** uma **promessa** 'to make a promise' 'to make a promise'	**se queixar** (IReflV) 'to complain REFL' 'to complain'
RO	a **trage pe sfoară** 'to pull on rope' 'to fool'	a **face** o **vizită** 'to make a visit' 'to pay a visit'	a **se gândi** (IReflV) 'to think REFL' 'to think'
SL	**spati kot ubit** 'to sleep like killed' 'to sleep soundly'	**postaviti vprašanje** 'to put a question' 'to ask a question'	**bati se** (IReflV) 'to fear REFL' 'to be afraid'
SV	att **plocka russinen ur kakan** 'to pick raisins out of the cake' 'to choose only the best things'	**ta** ett **beslut** 'to take a decision' 'to make a decision'	**det knallar och går** (OTH) 'it trots and walks' 'it is OK/as usual'

4 Annotation guidelines

Given the definitions in §2 and a text to annotate, each iteration of the annotation process starts with: (i) selecting a candidate sequence, i.e. a combination of a verb with at least one other word which could form a VMWE, (ii) establishing the precise list of its lexicalised components and its canonical forms. These steps are largely based on the annotator's linguistic knowledge and intuition.

Once a candidate sequence has been selected, its status as a VMWE is tested in two steps: identification and categorisation. Each step is based on linguistic tests and examples in many languages, organised into decision trees, so as to maximise the determinism in decision making.

4.1 Identification tests

Five generic non-compositionality tests were defined in order to identify a VMWE (of any category):

> **Test 1** [CRAN]: Presence of a cranberry word, e.g. (EN) *it* ***goes astray***;
>
> **Test 2** [LEX]: Lexical inflexibility, e.g. (EN) *they #allowed the feline out of the container* (*they* ***let the cat out of the bag***); **to give a stare* (*to* **give** a **look**);
>
> **Test 3** [MORPH]: Morphological inflexibility, e.g. (EN) *to #take a turn* (*to* ***take turns***);
>
> **Test 4** [MORPHOSYNT]: Morpho-syntactic inflexibility, e.g. (EN) *#I give you his word for that* (*I* ***give*** *you my* ***word*** *for that*);
>
> **Test 5** [SYNT]: Syntactic inflexibility, e.g. (EN) *#Bananas are gone* (*he* ***went bananas***).

If none of these tests apply, an additional hypothesis covers the LVC candidates, which usually fail Tests 1 and 3–5 and for which Test 2 is hard to apply due to their relatively high, although restricted, productivity.

> [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. (EN) ***makes a speech***.

Passing any of Tests 1–5 is sufficient for a candidate sequence to be identified as a VMWE, while the LVC hypothesis has to be confirmed by the LVC-specific tests.[6]

4.2 Decision tree for categorisation

Once a VMWE has been identified or hypothesised following the tests in the preceding section, its categorisation follows the decision tree shown in Figure 1. Tests 6–8 are structural, the others are category-specific.

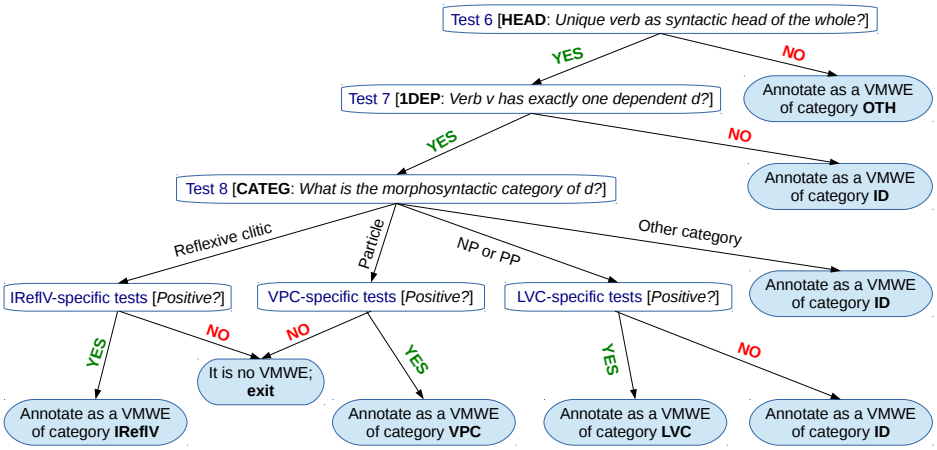

Figure 1: Decision tree for VMWE categorisation.

4.2.1 Structural tests

Categorisation of a VMWE depends on the syntactic structure of its canonical form determined by the following three tests:

> **Test 6** [HEAD]: Presence of a unique verb functioning as the syntactic head of the whole expression, like in (13) and unlike in (14).

[6] As explained in §10, feedback from the large-scale annotation of version 1.0 of the corpus led us to questioning the correctness of the two-stage VMWE annotation. In edition 1.1 we transformed the identification tests into ID-specific tests and performed VMWE identification simultaneously to their categorisation.

(13) *Je* ***laisse tomber***. (FR)
I let fall
I let fall. 'I let go, I abandon.'

(14) *wo man* ***lebt*** *und* ***leben lässt*** *(DE)*
where one lives and live lets
where one lives and lets live 'where one is tolerant'

Test 7 [1DEP]: Among the phrases dependent on the head verb exactly one contains lexicalised components, as in (EN) ***made*** *it* ***up***, and unlike in (EN) ***made up*** *her* ***mind***.

Test 8 [CATEG]: Morphosyntactic category of the verb's dependent. Contrary to most other tests, the result of this test is not binary but taken from a closed list of values: (i) reflexive clitic (REFL), as in (15), (ii) particle (PART), as in (16); (iii) nominal or prepositional phrase, as in (17); (iv) other (including a verb, an adverb, a non-reflexive pronoun, etc.), as in (18).

(15) *Той* ***се*** ***страхува***. (BG)
toy **se** **strahuva**
he REFL fears
He fears REFL. 'He is afraid.'

(16) *Der Film* ***fängt*** ***an***. (DE)
the film catches PART
The film catches PART. 'The film begins.'

(17) *Mój bratanek* ***buja*** ***w*** ***obłokach***. (PL)
my nephew swings in clouds
My nephew swings in the clouds. 'My nephew fantasizes.'

(18) *Uma ajudinha* ***cai*** *muito* ***bem***. (PT)
a help.DIM falls very well
A little help falls very well. 'A little help comes at the right moment.'

When a VMWE fails Test 6 or 7, it is automatically classified as OTH and ID, respectively. This means that we do not allow cumulative categories. For instance,

in (20) the reflexive clitic considerably changes the meaning of the base VPC from (19), which might qualify the whole as an IReflV. However, due to the presence of two lexicalised syntactic arguments of the verb, such cases are necessarily classified as IDs (here: with a nested VPC).

(19) *Er **stellte** mir seine Freundin **vor**.* (DE)
he put me his friend PART
He put his friend PART to me. 'He presented his friend to me.'

(20) *Er **stellte sich** die Reise **vor**.* (DE)
he put REFL.3.SG the travel PART
He put the travel PART to REFL. 'He imagined the travel.'

Test 8, with return values (i)-(iii), triggers the category-specific tests for IReflVs, VPCs and LVCs, respectively. For other categories the candidate automatically qualifies as an ID.

4.2.2 Light-verb constructions

Light-verb constructions (LVCs) gave rise to a vast literature since first introduced by Jespersen (1965), possibly because there is no consensus on their exact definition and scope. We consider a candidate sequence an LVC if it consists of a verb *V* and a nominal complement *N*, possibly introduced by a preposition, provided that it passes all of the following tests:

Test 9 [N-EVENT]: *N* denotes an event or a state, as in (21);

(21) *Οι συσκευές **έχουν** τη **δυνατότητα** σύνδεσης.* (EL)
I siskieves eχun ti δinatotita sinδesis
the devices have the ability connection.SG.GE.
The devices have the ability to connect. 'The devices can connect.'

Test 10 [N-SEM]: *N* has one of its original senses, as in (22) and unlike in (23);

(22) *Steffi **rend** **visite** à Monica.* (FR)
Steffi returns visit to Monica
Steffi returns a visit to Monica. 'Steffi pays a visit to Monica.'

(23) *Je **jette** l'**éponge**.* (FR)
I throw the'sponge
I throw the sponge. 'I give up.'

Test 11 [V-LIGHT]: *V* only contributes morphological features (tense, mood, person, number, etc.) but adds no semantics that is not already present in *N*, other than the semantic role of *V*'s subject with respect to *N*, as in (24);

(24) *Gydytojai **padarė išvadą**, kad gijimo procesas vyksta*
Doctors made conclusion, that recovery process happens
sėkmingai. (LT)
successfully.
The doctors made the conclusion that the recovery process is successful. 'The doctors came to the conclusion that the recovery process is successful.'

Test 12 [V-REDUC]: An NP headed by *N* can be formed containing all of *V*'s syntactic arguments, and denoting the same event or state as the LVC, e.g. (EN) *Paul **had** a nice **walk*** denotes the same event as (EN) *the nice walk of Paul.*

Test 13 [N-PROHIBIT-ARG]: A semantic argument of the same type cannot be syntactically realised twice – both for *N* and for *V*, e.g. (EN) **Paul **made** the **decision** of the committee* is meaningless, while (EN) *Paul leads the discussion of the committee* is acceptable. Therefore, *to lead a discussion* is not an LVC.

Tests 12 and 13 are syntactic tests approximating the property that one of *V*'s syntactic arguments (generally its subject) is *N*'s semantic argument.

Note that our definition of an LVC does not fully overlap with the state of the art. On the one hand, we are more restrictive than some approaches in that we do not include cases in which the verb does add some (even bleached) semantics to the noun. For instance, inchoative verbs combined with non-inchoative nouns such as (PL) *objąć patronat* 'to embrace patronage' ⇒ 'to take on patronage' fail Test 11 and are therefore not classified as LVCs, although their fully bleached counterparts are, as (PL) ***sprawować patronat*** 'to perform patronage' ⇒ 'to dispense patronage'. On the other hand, we include in LVCs those combinations

in which a semantically void verb selects a large class of action/state nouns so that its lexical non-compositionality is hard to establish, e.g. (FR) ***commettre*** *un* ***crime/délit/meurtre***/… ‘to commit a crime/offence/murder/…’.

The latter reason makes LVCs belong to the grey area of (non-)compositionality. They are mostly morphologically and syntactically regular. They can also be seen as semantically compositional in the sense that the semantically void light verb is simply omitted in the semantic calculus. However, this omission may itself be seen as an irregular property. This confirms the observation of Kracht (2007) that compositionality is a property of linguistic analyses rather than of language items.

4.2.3 Idioms

A verbal idiomatic expression (ID) comprises a head verb *V* (possibly phrasal) and at least one of its arguments. Following the decision tree from Figure 1, a VMWE is classified as an ID in one of the 3 cases:

1. *V* has more than one lexicalised argument, as in (25) and (26)

   (25) ***Srce*** *mu je* ***padlo v hlače***. (SL)
   heart him is fallen in pants
   His heart fell into his pants. ‘He lost courage.’

   (26) رسید. لبم به جانم (FA)
   resid **lab**am **be jan**am
   arrived lips-my to soul-my
   My soul arrived at my lips. ‘I am frustrated.’

2. *V*’s single lexicalised argument is of any category other than a reflexive clitic, a particle or a nominal phrase (possibly introduced by a preposition), as in (27), (28) and (29);

   (27) *Platforma* ***dopięła*** ***swego***. (PL)
   Platform PART-buttoned own
   The Platform buttoned PART her own. ‘The Platform fulfilled its plans.’

(28) ***Es gibt*** *kein Zurück.* (DE)
it gives no back
It gives no retreat. 'There is no retreat.'

(29) *Ele* ***sabe onde pisar***. (PT)
he knows where step
He knows where to step. 'He knows how to succeed.'

3. *V*'s single lexicalised argument is a nominal phrase (possibly introduced by a preposition), at least one of the LVC-specific Tests 9–13 fails but at least one of the identification Tests 1–5 applies, as in (30).

(30) *Artık kimsenin* ***aklına gelmeyecek***. (TR)
anymore of-anyone to-his-mind it-will-not-come
It will not come to the mind of anyone anymore. 'No one will remember it anymore.'

Distinguishing an ID from an LVC in case 3 is one of the hardest and most frequent annotation challenges. In case 1, care must be taken to identify and also annotate nested VMWEs (if any), e.g. the VMWE in (31) contains a nested ID (RO) ***dă pe față*** 'gives on face' ⇒ 'reveals'.

(31) *El* ***dă*** ***cărțile pe față***. (RO)
he gives cards on face
He gives the cards on the face. 'He reveals his intentions.'

Idioms whose head verb is the copula (*to be*) pose special challenges because their complements may be (nominal, adjectival, etc.) MWEs themselves. In this task, we consider constructions with a copula to be VMWEs only if the complement does not retain the idiomatic meaning when used without the verb. For instance, (PL) *on* ***jest jedną nogą na tamtym świecie*** 'he is with one leg in the other world' ⇒ 'he is close to death' is an ID because (PL) *jedna noga na tamtym świecie* 'one leg in the other world' loses the idiomatic meaning, while (PL) to stwierdzenie *jest do rzeczy* 'this statement is to the thing' ⇒ 'this statement is relevant' is not a VMWE since (PL) *do rzeczy* 'to the thing' ⇒ 'relevant' keeps the idiomatic reading.

4.2.4 Inherently reflexive verbs

Pronominal verbs, sometimes also called reflexive verbs, are formed by a verb combined with a reflexive clitic (REFL). They are very common in Romance and Slavic languages, and occur in some Germanic languages such as German and Swedish. Clitics can be highly polysemous and sometimes have an idiomatic rather than a reflexive meaning, in which case we call them inherently reflexive verbs (IReflVs). To distinguish regular from idiomatic uses of reflexive clitics, we rely on an IReflV-specific decision tree[7] containing 8 tests, which are meant to capture an idiosyncratic relation between a verb with a reflexive clitic and the same verb alone. The first 3 of these tests are sufficient to identify most of the actual IReflVs:

Test 14 [INHERENT]: *V* never occurs without *C*, as in (32);

(32) *Jonas har* ***försovit sig*** *idag.* (SV)
Jonas has overslept REFL.3.SG today
Jonas overslept REFL today. 'Jonas overslept today.'

Test 15 [DIFF-SENSE]: *C* markedly changes the meaning of *V*, as in (33);

(33) *kar* ***se tiče*** *Kosova* (SL)
what REFL touches Kosovo
what REFL touches Kosovo 'as far as Kosovo is concerned'

Test 16 [DIFF-SUBCAT]: *C* changes the subcategorisation frame of *V*, as in (34) vs. (PT) *você me esqueceu* 'you forgot me'.

(34) *Você* ***se*** ***esqueceu*** *de mim.* (PT)
you REFL.3.SG forgot of me
You forgot REFL about me. 'You forgot about me.'

IReflVs are hard to annotate because pronominal clitics have several different uses. For example, (IT) *si* 'REFL' can occur not only in IReflVs such as (IT) ***riferirsi*** 'to report.REFL' ⇒ 'to refer', but also in the following non-idiomatic cases: reflexive (IT) *lavarsi* 'to wash.REFL', possessive reflexive (IT) *grattarsi la testa* 'to scratch.REFL head' ⇒ 'to scratch one's head', reciprocal (IT) *baciarsi* 'to

[7] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/?page=ireflv

kiss.REFL' ⇒'to kiss each other', impersonal (IT) *si dorme molto* 'REFL sleeps much' ⇒ 'people sleep a lot', middle alternation (IT) *si affittano case* 'REFL rent houses' ⇒ 'houses are rented' or inchoative (IT) *la porta si apre* 'the door REFL opens' ⇒ 'the door opens'. The IReflV category was reported as the most challenging to annotate by some teams, notably the Spanish and the Romanian ones.

4.2.5 Verb-particle constructions

Verb-particle constructions (VPCs) are pervasive notably in Germanic languages and Hungarian, but virtually non-existent in Romance or Slavic languages. They are formed by a lexicalised head verb *V* and a lexicalised particle *P* dependent on *V*, whose joint meaning is non-compositional. The latter property is approximated by a unique syntactic test:

> **Test 22** [V+PART-DIFF-SENSE] A sentence without *P* does not refer to the same event/state as the sentence with *P*. For example, the sentence in (35) does not imply (HU) *nekem jött ez a koktél* 'this cocktail bumped into me', while (DE) *er legt das Buch auf dem Tisch ab* 'he puts the book on the table PART' implies (DE) *er legt das Buch auf dem Tisch* 'he puts the book on the table'.

(35) ***Be-jött*** *ez a koktél nekem.* (HU)
PART-bumped this the cocktail for.me
This cocktail bumped PART into me. 'I like this cocktail.'

The first challenge in identifying a VPC is to distinguish a particle, as in (EN) *to* ***get*** *<u>**up**</u> a party*, from a homographic preposition, as in (EN) *to get <u>up</u> the hill*. Language-specific tests were designed for German and English to this aim.

In some Germanic languages and also in Hungarian, verb-particle constructions can be spelled either as one (multiword) token, as in (36), or separated, as in (37). Both types of occurrences are to be annotated.

(36) *ő* ***be-rúgott***. (HU)
he PART-kicked
He kicked PART. 'He got drunk.'

(37) *Nem ő* ***rúgott be***. (HU)
not he kicked PART
He did not kick PART. 'He did not get drunk.'

Special care must be taken with polysemous constructions having both a compositional and a non-compositional reading, as in (DE) *ein Schild aufstellen* 'to put up a sign' vs. (DE) *einen Plan* ***aufstellen*** 'to put up a plan' ⟹ 'to draw up a plan'.

4.2.6 Other VMWEs

This category gathers the VMWEs which do not have a single verbal head (cf. Test 6 in Figure 1 and §4.2.1). Those include:

- Coordinations like in example (14) p. 101, or (38)

(38) בריטניה ***נשאה ונתנה*** עם מצרים. (HE)
micrayim 'im **ve-natna nas'a** britanya
Egypt with and-gave carried Britain
Britain carried and gave with Egypt. 'Britain negotiated with Egypt.'

- Compound verbs, resulting usually from conversion of nominal compounds, and therefore having no regular verbal structure, as in (39) or in (EN) *to* ***pretty-print***.

(39) *On* ***court-circuite*** *le réseau terrestre.* (FR)
one short-circuits the network terrestrial
One short-circuits the terrestrial network. 'One bypasses the terrestrial network.'

4.3 Language-specific interpretation of the guidelines

Despite huge efforts put into setting up generic terminologies and methodologies, as well as into the pilot annotations and the project coordination, language-specific interpretation of the final guidelines could not be avoided. This was mainly due to different linguistic sensitivities and traditions, language-specific challenges and incompleteness or imprecision of the guidelines.

The most notable deviation occurred in Farsi, where no categorisation was performed, and the OTH label was used for all identified VMWEs instead. The main reason is the particularly challenging nature of the VMWE phenomenon in this language. There are less than 200 actively used simple (single-word) verbs, and

a large majority of events and processes are expressed by multiword combinations, many of which are potential VMWEs. The implications on our annotation process are at least threefold. Firstly, verbs are extremely polysemous, so Test 11 (§4.2.2) is very difficult to apply. In particular, the highly frequent light verb کردن */kardan/* 'to do/make' is ambiguous in its passive form شدن*/šodan/* 'done/made' with the semi-copula equivalent roughly to 'become'. Only the former interpretation should yield a VMWE annotation but the difference is hard to capture. Secondly, rephrasing an LVC by a single verb, often used to approximate Test 9 in other languages (*to **make** a **decision** = to decide*), is rarely feasible in Farsi. Thirdly, VMWEs are extremely pervasive, which is easily visible in Table 3: the number of annotated VMWEs is roughly the same as the number of sentences, i.e. almost every main verb is the head of a VMWE. As a result, the VMWE phenomenon is particularly hard to capture in Farsi since it can rarely be contrasted with verbal constructions deemed compositional.

Another notable deviation occurred in Slovene, where the VPC category, as defined by the generic guidelines, hardly or never occurs, however it was used instead to annotate idiomatic verb-preposition combinations, such as (SL) ***prišlo** je **do** nesreče* 'it came to an accident' ⟹ 'an accident occurred'.

The status of VPCs in Italian is interesting. As a Romance language, Italian was expected not to exhibit VPCs, but several dozens of VPC annotations do occur in the Italian corpus, e.g. (IT) ***volata via*** 'flew PART' ⟹ 'slipped away', ***tira fuori*** 'pulls PART' ⟹ 'shows', or ***va avanti*** 'goes PART' ⟹ 'goes on'. This shows the possibly ambiguous status of *via* 'by/away', *avanti* 'on/forward', *fuori* 'out/outside', etc. as either adverbs or particles, triggering the ID or the VPC category, respectively. The semantic compositionality of some of these constructions might also be examined more closely.

In Bulgarian and Czech, the auxiliaries accompanying the head verbs were annotated as VMWE components, e.g. in (CS) *on **se** bude **bavit*** 'he REFL will play' ⟹ 'he will play', in (BG) *te ne sa **dali saglasie*** 'they not are given consent' ⟹'they have not given consent'. This is in contrast with the guidelines, which stipulate that only the lexicalised components should be annotated. The motivation for this deviation was to always include a finite verb in the annotated expression, so as to e.g. easily study the tense and mood restrictions in VMWEs. Since such studies are enabled by the accompanying morpho-syntactic data (currently existent in Czech and to be provided in Bulgarian in the future), these divergences should be eliminated in new editions of the corpus.

In German, a deviation was observed with respect to VMWEs containing both a reflexive clitic and a particle such as (DE) *sie **bringen sich ein*** 'they bring REFL

PART' ⟹ 'they contribute'. Such cases were annotated as IReflVs with nested VPCs, which does not conform to Test 7 (§4.2.1) stipulating that, whenever the VMWE has more than one lexicalised dependent of the head verb, it should be classified as an ID (here: with a nested VPC). Good reasons exist for each of these strategies and more discussion is needed to arbitrate for future releases of the guidelines.

Lithuanian seems to have a surprisingly low number of LVCs, despite the large size of the annotated corpus. It would be worthwhile to study in more detail if this phenomenon is inherent to the language or results from a more restrictive understanding of the LVC scope.

In Hebrew, a relatively large number of VMWEs of type OTH was observed (cf. Table 3), and a necessity of defining a new category (specific to non-Indo-European languages) was hypothesised. A more detailed study revealed that most OTH annotations were spurious: they concerned statistical collocations or VMWEs of the ID or LVC types. Some idiomatic verb-preposition combinations were also annotated in Hebrew, despite the fact that we had abandoned the IPrepV category in the earlier stages of the project (§3). There, the annotators faced a particular challenge from prepositions which often attach to the governed noun and annotating them as separate lexicalised tokens was mostly impossible. Thus, in the following sequence: (HE) *sovel me.achuz avtala* 'suffers from.a.percentage of.unemployment' the free complement *achuz* 'percentage' had to be annotated as lexicalised together with its governing preposition *me* 'from'. This problem will be dealt with in the future, when inherently adpositional verbs will be addressed (§10).

In Turkish, the LVC and OTH types also had their language-specific interpretation. Namely, the Turkish PARSEME corpus resulted from adapting a pre-existing MWE typology and dataset (Adalı et al. 2016). There, the definition of a light verb, based on Turkish linguistic works (Siemieniec-Gołaś 2010), was context-independent, i.e. restricted to a closed list of 6 verbs: *olmak* 'to be', *etmek* 'to do', *yapmak* 'to make', *kılmak* 'to render', *eylemek* 'to make' and *buyurmak* 'to order'. Verb-noun combinations with other operator verbs, such as ***söz vermek*** 'promise to give' ⟹ 'to promise', were then classified as OTH. A closer look at the existing OTH annotations reveals, indeed, that most of them can be re-classified as LVC in future releases of the corpus.

Czech is another language in which a pre-existing MWE-annotated corpus (Hajič et al. 2017) was adapted to the needs of the PARSEME initiative. There, complex identification and conversion procedures had to be designed (Bejček et al. 2017). The resulting mapping procedure could be fully automatic, which suggests

that the understanding of the VMWE phenomenon is similar in both annotation projects. It would still be interesting to compare both annotation guidelines more thoroughly and look for possible divergences.

5 Annotation methodology and tools

Mathet et al. (2015) mention several challenging features of linguistic annotation, some of which are relevant to the VMWE annotation task:

- *Unitising*, i.e. identifying the boundaries of a VMWE in the text;
- *Categorisation*, i.e. assigning each identified VMWE to one of the pre-defined categories (§3);
- *Sporadicity*, i.e. the fact that not all text tokens are subject to annotation (unlike in part-of-speech annotation, for instance);
- *Free overlap*, e.g. in (CS) ***ukládal*** *různé* ***sankce*** *a* ***penále*** 'put various sanctions and penalties', where two LVCs share a light verb;
- *Nesting*,
  - at the syntactic level, as in (40), where an IReflV (PL) ***skarżyć się*** 'to complain REFL' ⇒ 'to complain' occurs in a relative clause modifying the predicative noun of the LVC (PL) ***popełnić oszustwo*** 'to commit a fraud'.

    (40) ***Oszustwa****, na jakie* ***skarżą*** ***się*** *Cyganie,* ***popełniły***
    frauds, on which complain REFL Gypsies, committed
    grupy zorganizowane. (PL)
    groups organised
    Organised groups committed frauds about which the Gypsies REFL complain. 'Frauds which Gipsies complain about were committed by organised groups.'

  - at the level of lexicalised components, as in (41), where the ID (PT) ***fazer justiça*** 'to make justice' ⇒ 'to do justice' is nested within a larger ID.

    (41) *Ales* ***fizeram justiça com as próprias mãos****.* (PT)
    they made justice with their own hands
    They made justice with their own hands. 'They took the law into their own hands.'

Two other specific challenges are:

- *Discontinuities*, e.g. (CS) *on* ***ukládal*** *různé* ***sankce*** 'he put various sanctions';
- *Multiword token* VMWEs, e.g. separable IReflVs or VPCs:[8]
  (ES) ***abstener.se*** 'to abstain.REFL' ⇒ 'to abstain',
  (HU) ***át.ruház*** 'to PART.dress' ⇒ 'to transfer'.

This complexity is largely increased by the multilingual nature of the task, and calls for efficient project management and powerful annotation tools.

5.1 Project management

The list of language teams having initially expressed their interest in this initiative included those mentioned in p. 91, as well as English, Croatian and Yiddish, for which no corpus release could be achieved due to the lack of sufficiently available native annotators. All languages were divided into four language groups (LGs) - Balto-Slavic, Germanic, Romance and others - as also described in p. 91. The coordination of this large project included the definition of roles – project leaders, technical experts, language group leaders (LGLs), language leaders (LLs) and annotators – and their tasks.

The biggest challenge in the initial phase of the project was the development of the annotation guidelines[9] which would be as unified as possible but which would still allow for language-specific categories and tests. To this end, a two-phase pilot annotation in most of the participating languages was carried out. Some corpora were annotated at this stage not only by native but also by near-native speakers, so as to promote cross-language convergences. Each pilot annotation phase provided feedback from annotators, triggered discussions among language (group) leaders and organisers, and led to enhancements of the guidelines, corpus format and tools.

We also defined strategies for selecting the final corpora. They should: (i) be written in the original, in order to avoid MWE-related translationese issues; (ii)

[8]Note that annotating separate syntactic words within such tokens would be linguistically more appropriate, and would avoid bias in inter-annotator agreement and evaluation measures – cf. §6.2 and (Savary et al. 2017). However, we preferred to avoid token-to-word homogenising mainly for the reasons of compatibility. Namely, for many languages, pre-existing corpora were used, and we would like VMWE annotations to rely on the same tokenisation as the other annotation layers.

[9]Their final version, with examples in many participating languages, is available under the CC BY 4.0 license at http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/.

correspond to the same genre: newspaper texts or Wikipedia articles;[10] (iii) consist of longer text fragments (rather than isolated sentences), so as to enable disambiguation and coreference resolution; (iv) not be automatically pre-selected in view of a higher density of VMWEs (so as to provide both positive and negative examples); (v) be free from copyright issues, i.e. compatible with open licenses.

5.2 Annotation platform

For this large-scale corpus construction, we needed a centralised web-based annotation tool. Its choice was based on the following criteria: (i) handling different alphabets; (ii) accounting for right-to-left scripts; and (iii) allowing for discontinuous, nested and overlapping annotations. We chose FLAT,[11] a web platform which, in addition to the required criteria, enables token-based selection of text spans, including cases in which adjacent tokens are not separated by spaces. It is possible to authenticate and manage annotators, define roles and fine-grained access rights, as well as customise specific settings for different languages.

FLAT is implemented as a web-based frontend with support for multiple users, user groups, and with configurable access rights. The frontend communicates with the FoLiA document server backend,[12] which loads and holds documents in memory as they are being edited, writes them to disk again at convenient times, and unloads them when they are not used anymore. The document server has Git version control support,[13] allowing changes to be tracked. In addition, for each individual FoLiA annotation, e.g. each VMWE, information such as who made the annotation, and when, is automatically registered.

FLAT is document-centric, i.e. it supports annotation of full documents together with their structure (headers, bulleted lists, figures, etc.). This contrasts with tools which take a more corpus-based approach with keyword-in-context visualisation. FLAT does allow for various other *perspectives* on the document; for the PARSEME annotation task a sentence-based perspective was chosen, presenting users with one or more pages of clearly delimited sentences to annotate. An example is shown in Figure 2.

FLAT is based on FoLiA,[14] a rich XML-based format for linguistic annotation (van Gompel & Reynaert 2013), and is compatible with a wide variety of linguis-

[10] Deviations from this rule occurred in some languages due to the choice of pre-existing corpora, e.g. in Hungarian legal texts were used.

[11] https://github.com/proycon/flat

[12] https://github.com/foliadocserve

[13] https://git-scm.com/

[14] https://proycon.github.io/folia

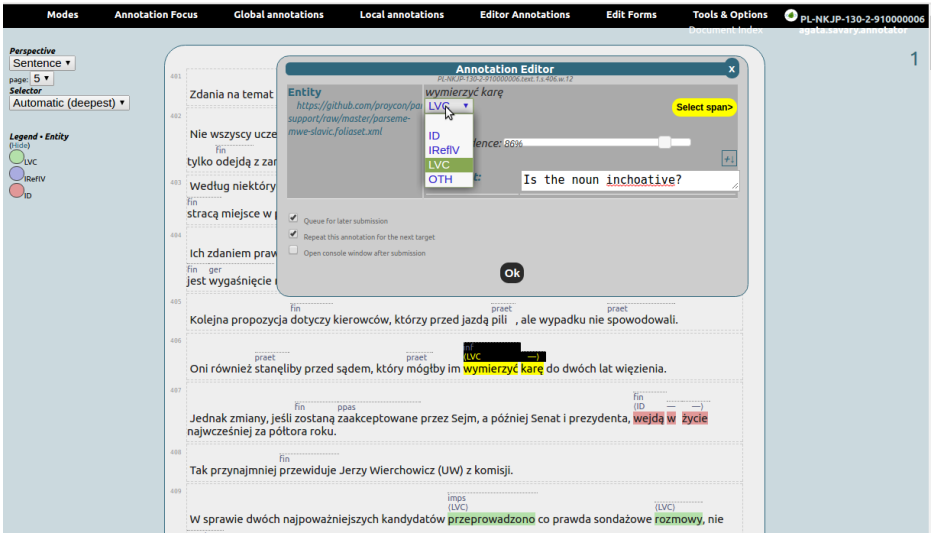

Figure 2: FLAT annotation interface with a Polish text. The VMWEs are coloured according to their categories. POS tags (*fin*, *ger*, *imps*, *ppas*, and *praet*) are displayed above all verbal tokens. Some attributes (VMWE category, confidence level and a comment) of the highlighted VMWE (PL) ***wymierzyć karę*** 'to PART.measure a punishment' ⟹ 'to mete out a punishment' are edited in the annotation editor.

tic annotation types. VMWEs, or entities as they are called more generically in FoLiA, constitute the most important annotation type for PARSEME. Still, certain language teams worked on documents enriched with more linguistic annotations, such as part-of-speech tags, to aid the annotation process, as shown in Figure 2. The underlying aspiration of both FoLiA and FLAT is to provide a single unified solution for multiple annotation needs, with respect to the encoding format and the annotation environment, respectively.

While the FoLiA format specifies possible linguistic annotation types and structural types, it does not commit to any particular tagset/vocabulary nor language. Instead, tagsets are defined externally in *FoLiA set definitions*, which can be published anywhere online by anyone and are deliberately separate from the annotation format itself. A dozen of set definitions for PARSEME, based on the VMWE categories relevant to different languages or language groups (§3) are likewise published in a public repository.[15] All FoLiA documents declare which particular set definitions to use for which annotation types. FLAT uses these set definitions to populate various selection boxes, as shown in Figure 2.

[15] https://github.com/proycon/parseme-support

All software discussed here is available under an open-source license.[16] It is part of a wider and growing infrastructure of FoLiA-capable NLP tools (van Gompel et al. 2017), developed and funded in the scope of the CLARIAH[17] project and its predecessor CLARIN-NL.

Although FLAT has been in use for various other annotation projects, the PARSEME initiative, currently with over 80 active FLAT users, is the biggest use case to date, and as such has had a very positive influence in terms of the maturity of the software, fixing bugs, attaining improved performance and scalability, and compiling appropriate documentation. Various features were added to accommodate PARSEME specifically: (i) uploading documents in non-FoLiA formats, needed for the parseme-tsv format (6.1); (ii) right-to-left support necessary for Farsi and Hebrew; (iii) a metadata editor; (iv) enhanced file and user management; (v) confidence level and free-text comments as part of the editable attributes (Figure 2).

Out of 18 language teams which achieved a corpus release, 13 used FLAT as their main annotation environment. The 5 remaining teams either used other (generic or in-house) annotation tools, or converted existing VMWE-annotated corpora.

5.3 Automatic VMWE pre-annotation

Automatic pre-annotation of corpora is a current practice in many annotation tasks. In the PARSEME corpus project, it was applied by the Bulgarian and Hungarian teams, on the basis of manually compiled lists of VMWEs. All texts were then manually checked and corrected.

More precisely, pre-annotation in Bulgarian included automatic annotation of: (a) verb forms (triggers for VMWEs), (b) IReflV candidates consisting of a verb and a reflexive particle, and (c) VMWEs from a large dictionary of Bulgarian MWEs (Koeva et al. 2016). Cases of false positives included: (i) literal uses of existing VMWEs, (ii) false IReflVs which are true reflexive or passive constructions instead (§4.2.4), or (iii) coincidental co-occurrence of VMWE components. All annotations were manually verified and such cases were eliminated. False negatives could also be efficiently tracked thanks to the highlighted verb forms.

Automatic pre-annotation is known to introduce a task-dependent bias (Marcus et al. 1993; Fort & Sagot 2010) which may be both positive (simple repetitive tasks are handled uniformly and speeded up) and negative (annotators may tend

[16] GNU Public License v3

[17] https://www.clariah.nl

to rely too much on the automatic pre-annotation and fail to detect false negatives). We are not aware of any studies about biases related to VMWE annotation. We expect a minor risk of bias to stem from a possibly unbalanced VMWE dictionary: if one category (e.g. LVCs) is better represented than others, annotators may become more attentive to it. A bias might also be introduced by relatively productive constructions, when a large majority, but not all, of their occurrences belong to a unique category. For instance, the verb (BG) *davam* 'to give' occurs often and in many different LVCs, e.g. with *saglasie* 'consent', *razreshenie* 'permission' *obyasnenie* 'explanation', etc. The annotators could, therefore, tend to wrongly assign the LVC category to other expressions containing the same verb, such as ***davam duma*** 'to give word' (ID), or *davam prizovka* 'to give subpoena' (non-VMWE or borderline case).

5.4 Consistency checks and homogenisation

Even though the guidelines heavily evolved during the two-stage pilot annotation, there were still questions from annotators at the beginning of the final annotation phase. We used an issue tracker (on Gitlab)[18] in which language leaders and annotators could discuss issues with other language teams.

High-quality annotation standards require independent double annotation of a corpus followed by adjudication, which we could not systematically apply due to time and resource constraints. For most languages, each text was handled by one annotator only (except for a small corpus subset used to compute inter-annotator agreement, see §6.2). This practice is known to yield inattention errors and inconsistencies between annotators, and since the number of annotators per language varies from 1 to 10, we used consistency support tools.

Firstly, some language teams (Bulgarian, French, Hungarian, Italian, Polish, and Portuguese) kept a list of VMWEs and their classification, agreed upon by all annotators and updated collaboratively over time.[19] Secondly, for some languages (German, French, Hebrew, Italian, Polish, Portuguese, Romanian and Spanish) the annotation was followed by homogenisation. An in-house tool extracted the annotated VMWEs from a given corpus and rescanned the corpus to find all potential occurrences of the same VMWEs, whether already annotated or not. It then generated an HTML page where all positive and negative examples of a given VMWE were grouped, and could be accepted or rejected manually. En-

[18] https://gitlab.com/parseme/sharedtask-guidelines/issues

[19] Like automatic pre-annotation, this practice increases the consistency and speed of the annotator's work, but it also introduces a risk of bias, since collective decisions may override linguistic intuition. Therefore, such instruments should always be used with special care.

tries were sorted so that similar VMWEs, such as (EN) ***payed** a **visit*** and ***received** a **visit***, appeared next to each other. In this way, noise and silence errors could easily be spotted and manually corrected. The tool was mostly used by language leaders and/or highly committed annotators. The resulting gain in precision and recall was substantial. For instance, in Spanish the number of the annotated MWEs increased by 40% (from 742 to 1248), most notably in the IReflV category. Figure 3 shows the interface used to correct consistency problems.

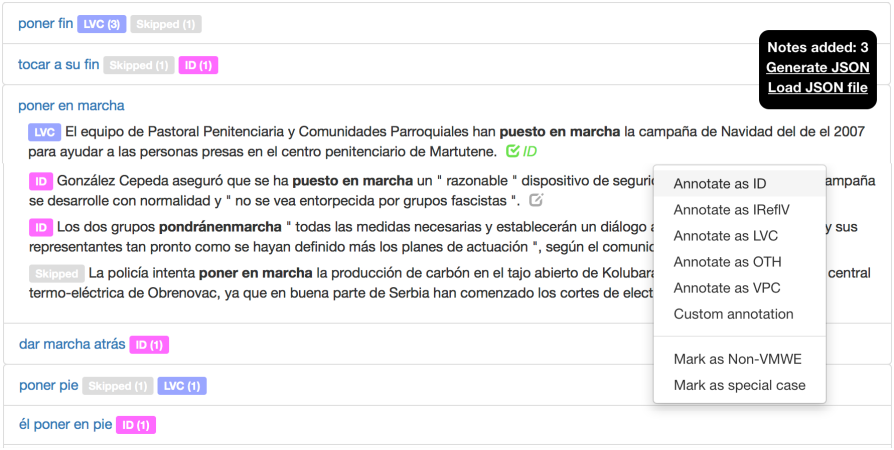

Figure 3: Consistency-check tool at work. Here, (ES) ***poner en marcha*** 'to put in march' ⟹ 'to start' was annotated once as LVC, twice as ID and once skipped. The clickable icon next to each example allows the user to add, correct or delete an annotation. VMWEs with the same noun, e.g. (ES) ***poner fin*** 'to put end' ⟹ 'to terminate' and ***tocar a su fin*** 'to touch to its end' ⟹ 'to come to its end' on the top of the screen, are gathered so as to enhance annotation consistency, especially for LVCs.

6 Properties of the annotated corpus

Table 3 provides overall statistics of the corpus annotated for the shared task.[20] In total, it contains almost 5,5 million tokens, 274 thousand sentences and 62 thousand VMWE annotations. The amount and distribution of VMWEs over categories varies considerably across languages.

No category was used in all languages, but the two universal categories, ID and LVC, were used in almost all languages. In Hungarian, no ID was annotated

[20]The split into training and test corpora is indicated in Savary et al. (2017).

Table 3: Overview of the annotated corpora in terms of the number of sentences, of tokens (whether belonging to the annotated VMWEs or not), and of the annotated VMWEs occurrences (overall and per category).

Language	Sentences	Tokens	VMWE occurrences					
			All	IDs	IReflVs	LVCs	OTHs	VPCs
BG	8,860	200,128	2,406	517	1,376	511	2	0
CS	49,431	833,193	14,536	1,611	10,000	2,923	2	0
DE	7,500	144,856	2,947	1,219	131	218	10	1,369
EL	8,811	226,265	2,018	642	0	1,291	37	48
ES	4,634	159,807	1,248	362	556	320	10	0
FA	3,226	55,207	3,207	0	0	0	3,207	0
FR	19,547	486,005	4,962	1,905	1,418	1,633	6	0
HE	7,000	147,361	1,782	116	0	380	693	593
HU	4,311	108,175	3,499	0	0	730	0	2,769
IT	17,000	427,848	2,454	1,163	730	482	6	73
LT	14,863	256,235	502	287	0	215	0	0
MT	10,600	152,285	1,272	446	0	693	133	0
PL	13,606	220,934	3,649	383	1,813	1,453	0	0
PT	22,240	414,020	3,947	910	596	2,439	2	0
RO	51,500	879,427	4,540	599	2,786	1,154	1	0
SL	11,411	235,864	2,287	375	1,198	231	4	479
SV	1,800	29,517	292	60	17	27	2	186
TR	18,036	362,077	6,670	3,160	0	2,823	687	0
Total	274,376	5,439,204	62,218	13,755	20,621	17,523	4,802	5,517

due to the genre of the corpus, mainly composed of legal texts. In Farsi, no categorisation was performed (§4.3), and all annotated VMWEs are marked as OTH instead.

The most frequent category is IReflV, in spite of it being quasi-universal, mainly due to its prevalence in Czech. IReflVs were annotated in all Romance and Slavic languages, and in German and Swedish. VPCs were annotated in German, Swedish, Greek, Hungarian, Hebrew, Italian, and Slovene. In the three last languages this category had a language-specific interpretation, as was the case of OTH in Hebrew and Turkish (§4.3). No language-specific categories have been defined.

All the corpora are freely available on the LINDAT/CLARIN platform.[21] The VMWE annotations are released under Creative Commons licenses, with constraints on commercial use and sharing for some languages. Some languages use data from other corpora (notably from the UD project), including additional annotations. These are released under the terms of the original licenses.

6.1 Format

The official format of the annotated data is the parseme-tsv format,[22] exemplified in Figure 4. It is adapted from the CoNLL format, with one token per line and an empty line indicating the end of a sentence. Each token is represented by 4 tab-separated columns featuring (i) the position of the token in the sentence, or a range of positions (e.g. `1–2`) in case of MWTs such as contractions; (ii) the token surface form; (iii) an optional `nsp` (no space) flag indicating that the current token is adjacent to the next one; and (iv) an optional VMWE code composed of the VMWE's consecutive number in the sentence and – for the initial token in a VMWE – its category, for example, `2:ID` if a token is the first one in an idiom which is the second VMWE in the current sentence. In case of nested, coordinated or overlapping VMWEs, multiple codes are separated with a semicolon.

Formatting of the final corpus required a language-specific tokenisation procedure, which can be particularly tedious in languages presenting contractions. For instance, (FR) *du* 'of-the' is a contraction of the preposition (FR) *de* 'of' and the article (FR) *le* 'the.MASC'.

Some language teams resorted to previously annotated corpora which have been converted to the parseme-tsv format automatically (or semi-automatically if some tokenisation rules were revisited). Finally, scripts for converting the parseme-tsv format into the FoLiA format and back were developed to ensure corpus compatibility with FLAT (5.2).

6.2 Inter-annotator agreement

Inter-annotator agreement (IAA) measures are meant to assess the hardness of the annotation task, as well as the quality of the annotation guidelines, of the annotation methodology, and of the resulting annotations. Defining such measures is not always straightforward due to the challenges listed in §5.

To assess unitising, two annotators double-annotated an extract of the corpus in each language. We then calculated the MWE-based F-score ($F1_{unit}$) of one

[21] http://hdl.handle.net/11372/LRT-2282

[22] http://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation

1-2	Wouldn't		
1	Would		
2	not		
3	questioning		
4	colonial		
5	boundaries		
6	open		1:ID
7	a		
8	dangerous		
9	Pandora	nsp	1
10	'	nsp	1
11	s		1
12	box	nsp	1
13	?		
1	They		
2	were		
3	letting		1:VPC;2:VPC
4	him		
5	in		1
6	and		
7	out		2
8	.	nsp	

Figure 4: Annotation of two sample sentences containing a contraction (*wouldn't*), a verbal idiom, and two overlapping VPCs.

annotator with respect to the other.[23] MWE-based F-score is defined in Savary et al. (2017) and was used to evaluate the systems submitted to the shared task.

We also report an estimated Cohen's κ (κ_{unit}). Measuring IAA, particularly κ, for unitising is not straightforward due to the absence of negative examples, that is, spans for which both annotators agreed that they are not VMWEs. From an extreme perspective, any combination of a verb with other tokens (of any length) in a sentence is a potential VMWE.[24] Consequently, as the density of VMWEs in most languages is rather low, one can argue that the probability of chance agreement approaches 0, and IAA can be measured simply using the observed agreement $F1_{unit}$. However, in order to provide a possibly less biased measure

[23] That is, we suppose that one annotator represents the system, and the other one represents the gold standard. Note that F-score is symmetrical (depending on the order, recall and precision are inverted), so none of the two annotators is prioritised.

[24] Also note that annotated segments can overlap.

to the reported F-scores, we assume that the total number of stimuli in the annotated corpora is approximately equivalent to the number of verbs, which is slightly higher than the number of sentences. We roughly estimate this quantity as the number of sentences plus the number of VMWEs annotated by at least one annotator.[25] Finally, to assess categorisation, we apply the standard κ (κ_{cat}) to the VMWEs for which annotators agree on the span.

Due to time and resource constraints, the majority of the corpus for most languages was annotated by a single annotator. Only small fractions were double-annotated for the purpose of the IAA calculation. All available IAA results are presented in Table 4. For some languages the IAA in unitising is rather low. We believe that this results from particular annotation conditions. In Spanish, the annotated corpus is small (Table 3), so the annotators did not become sufficiently accustomed to the task. A similar effect occurs in Polish and Farsi, where the first annotator performed the whole annotation of the train and test corpora, while the second annotator only worked on the IAA-dedicated corpus. The cases of Hebrew, and especially of Italian, should be studied more thoroughly in the future. Note also that in some languages the numbers from Table 4 are a lower bound for the quality of the final corpus, due to post-annotation homogenisation (§5.4).

A novel proposal of the holistic γ measure (Mathet et al. 2015) combines unitising and categorisation agreement in one IAA score, because both annotation subtasks are interdependent. In our case, however, separate IAA measures seem preferable both due to the nature of VMWEs and to our annotation methodology. Firstly, VMWEs are known for their variable degree of non-compositionality. In other words, their idiomaticity is a matter of scale. But this fact is not accounted for in current corpus annotation standards and identification tools, which usually rely on binary decisions, i.e. a candidate is seen as a VMWE or a non-VMWE, with no gradation of this status. Such a binary model is largely sub-optimal for a large number of grey-zone VMWE candidates. However, once a VMWE has been considered valid, its categorisation appears to be significantly simpler, as shown in the last 2 columns of Table 4 (except for Romanian and Hebrew). Secondly, as described in §4.1 – §4.2, our annotation guidelines are structured in two main decision trees – an identification and a categorisation tree – to be applied mostly sequentially. Therefore, separate evaluation of these two stages may be helpful in enhancing the guidelines.

[25] In other words, the number of items on which both annotators agree as being no VMWEs is estimated as the number of sentences. This assumption ignores the fact that some verbs may be part of more than one VMWE, since this is rare.

Table 4: IAA scores: #S, and #T show the the number of sentences and tokens in the double-annotated sample used to measure IAA, respectively. $\#A_1$ and $\#A_2$ refer to the number of VMWE instances annotated by each of the annotators.

	#S	#T	$\#A_1$	$\#A_2$	$F1_{unit}$	κ_{unit}	κ_{cat}
BG	608	27491	298	261	81.6	0.738	0.925
EL	1383	33964	217	299	68.6	0.632	0.745
ES	524	10059	54	61	38.3	0.319	0.672
FA	200	5076	302	251	73.9	0.479	n/a
FR	1000	24666	220	205	81.9	0.782	0.93
HE	1000	20938	196	206	52.2	0.435	0.587
HU	308	8359	229	248	89.9	0.827	1.0
IT	2000	52639	336	316	41.7	0.331	0.78
PL	1175	19533	336	220	52.9	0.434	0.939
PT	2000	41636	411	448	77.1	0.724	0.964
RO	2500	43728	183	243	70.9	0.685	0.592
TR	6000	107734	3093	3241	71.1	0.578	0.871

6.3 Cross-language analysis

The common terminology and annotation methodology achieved in this endeavour enable cross-language observations. In this section we offer a comparative quantitative analysis of several phenomena relevant to the challenges VMWEs pose in NLP, as discussed in §1. Namely, we analyse the lengths, discontinuities, coverage, overlapping and nesting of VMWEs across languages and VMWE types.

Table 5 provides statistics about the length and discontinuities of annotated VMWEs in terms of the number of tokens.[26] The average lengths range between 1.27 (in Hungarian) and 2.71 (in Hebrew) tokens, but the dispersion varies across languages: the mean absolute deviation (MAD) is 0.75 for Hebrew, while it is 0.11 for Turkish. Single-token VMWEs (length=1) are frequent in Hungarian and German (63% and 24% of all VMWEs, respectively) but rare or non-existent in other languages. The right part of Table 5 shows the lengths of discontinuities (gaps). This factor is measured in terms of the total number of tokens not belonging to

[26]Since the version published in Savary et al. (2017), we corrected a bug in the length average and MAD calculation, which impacted the results for languages containing VMWEs with one token only (especially DE and HU).

Table 5: Length and discontinuities of VMWE occurrences in number of tokens in the training corpora. Col. 2–3: average and mean absolute deviation (MAD) for length. Col. 4: number of single-token VMWEs. Col. 5–6: average and MAD for the length of discontinuities. Col. 7–8: number and percentage of continuous VMWEs. Col. 9–11: number of VMWEs with discontinuities of length 1, 2 and 3. Col. 12–13: number and percentage of VMWEs discontinuities of length > 3.

	Length of VMWE			Length of discontinuities (excl. VMWEs of length 1)								
Lang.	Avg	MAD	=1	Avg	MAD	0	%0	1	2	3	>3	%>3
BG	2.45	0.63	1	0.64	1.05	1586	82.1	206	33	25	82	(4.2%)
CS	2.30	0.46	0	1.35	1.53	6625	51.5	2357	1465	944	1461	(11.4%)
DE	2.02	0.61	715	2.96	2.94	619	35.7	283	159	142	529	(30.5%)
EL	2.45	0.61	3	0.94	1.08	870	57.4	389	124	50	82	(5.4%)
ES	2.24	0.39	0	0.47	0.66	523	69.9	162	33	14	16	(2.1%)
FA	2.16	0.27	0	0.42	0.70	2243	82.9	202	103	60	99	(3.7%)
FR	2.29	0.44	1	0.65	0.80	2761	61.9	1116	336	125	123	(2.8%)
HE	2.71	0.75	0	0.47	0.74	1011	78.9	129	54	43	45	(3.5%)
HU	1.27	0.39	2205	1.01	1.29	506	63.7	178	34	15	61	(7.7%)
IT	2.58	0.64	2	0.28	0.46	1580	80.9	278	56	22	16	(0.8%)
LT	2.35	0.53	0	0.72	0.94	261	64.9	79	36	9	17	(4.2%)
MT	2.64	0.68	7	0.34	0.53	589	77.0	123	33	12	8	(1.0%)
PL	2.11	0.20	0	0.53	0.77	2307	73.3	470	195	90	87	(2.8%)
PT	2.19	0.37	76	0.67	0.78	1964	58.3	1016	223	82	86	(2.6%)
RO	2.15	0.25	1	0.55	0.72	2612	64.7	689	693	32	13	(0.3%)
SL	2.27	0.43	14	1.47	1.54	787	44.4	445	221	118	202	(11.4%)
SV	2.14	0.25	0	0.38	0.59	44	78.6	7	3	1	1	(1.8%)
TR	2.06	0.11	3	0.57	0.57	3043	49.4	2900	162	33	28	(0.5%)

a VMWE but appearing between its left- and right-most lexicalised components. For instance, a gap of length 3 is counted in (DE) *jetzt **bin** ich bestimmt **aus** dem **Alter heraus*** 'now am I certainly out-of the age PART' ⇒ 'now I am too old'. The discontinuities vary greatly across languages. While for Bulgarian, Farsi and Italian more than 80% of VMWEs are continuous, only 35.7% of German VMWEs do not have any gaps, and 30.5% of them contain discontinuities of 4 or more tokens.

Figure 5 and Figure 6 show a breakdown of the length and discontinuity scores per VMWE category (Farsi, where categorisation was not performed, is not included). Not surprisingly, IDs are longer on average than all other categories (OTHs are omitted due to their rarity), and the average ID length ranges roughly between 2.5 and 3 components. The average lengths for the other categories are closer to 2, which is expected given their definitions. Note though that VPCs are

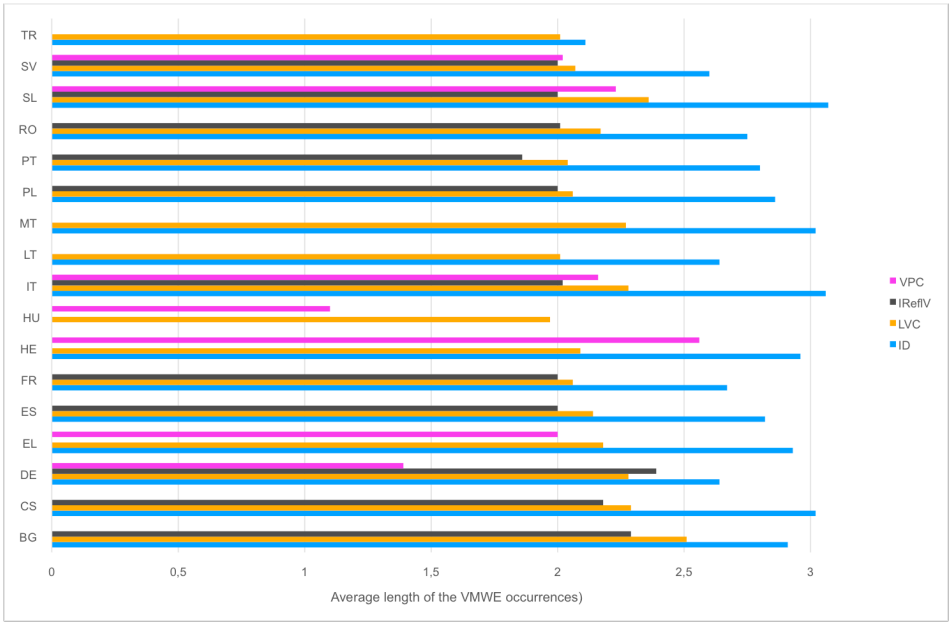

Figure 5: Average lengths of VMWE occurrences per category, in number of components. Single-token VMWEs (frequent for Hungarian and German) are included.

more contrasted across languages, with a low average length for German and Hungarian, due to the massive presence of single-token VMWEs. As far as IReflVs are concerned, a similar effect can be observed for some languages depending on morphological and tokenisation rules, due to the presence of IReflVs of length 1, for instance (ES) ***referir.se*** 'to refer.REFL' ⟹ 'to refer'. IReflVs of length greater than 2 in Czech, Bulgarian and German result from language-specific interpretations of the guidelines (§4.3).

When comparing the lengths of discontinuities across languages (Figure 6), German stands clearly out in all categories and so does Slovene to a smaller extent (probably due to the language-specific interpretation of the VPC category, §4.3), whereas Italian, Hebrew or Maltese show very few discontinuities. Note the difference for LVCs within Romance languages, which should be studied in more detail. LVCs are clearly the category showing the longest discontinuities overall, mainly due to the presence of non-lexicalised determiners and pre-modifiers of the predicative nouns, although extraction of the nouns also comes into play.

While regularities do exist in the formation of MWEs, it essentially remains an idiosyncratic and lexical phenomenon. Hence, it is very likely that the annotated

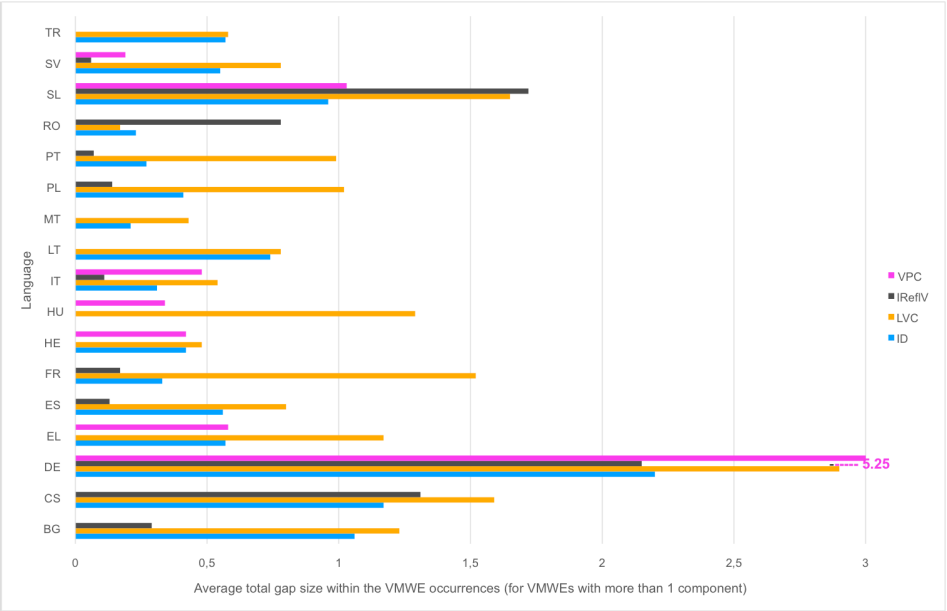

Figure 6: Size of discontinuities in VMWEs. The gap size is the total number of tokens not belonging to a VMWE but appearing between its left- and right-most lexicalised components. VMWEs of length 1 are not considered. For German the VPC average gap size is 5.25.

datasets cover only a small fraction of all the VMWEs existing in each of the 18 languages. In order to evaluate this coverage, we propose to measure the ratio of unknown VMWEs considering a corpus split into training and test sets, similar to the split used in the shared task (Savary et al. 2017). In other words, we arbitrarily split the corpus into a training and a test set, and study the proportion of VMWEs present in the test but absent in the training set.[27]

Ideally, we should perform this estimation on an intra- and inter-domain basis. Unfortunately, we do not know the domain of the source text for each annotated sentence.[28] To circumvent this limitation, we can still provide a lower bound of the unknown VMWE ratios by considering different splits that use continuous portions of the corpus, as shown in Figure 7. For each language for which the morphological companion files were provided, we show the average rate of un-

[27] See also Maldonado & QasemiZadeh (2018 [this volume]) and Taslimipoor et al. (2018 [this volume]) for in-depth considerations on how the training vs. test corpus split influences the results of automatic VMWE identification.

[28] For instance the French dataset contains the UD corpus, whose sentences come from various untraced sources and are mixed.

known VMWEs[29] computed over 5 cross-validation splits, plotted against the total number of VMWE occurrences. For instance for Italian we get an average unknown rate of 66.2%, with roughly 2,000 annotated VMWE tokens, which means that, on average, in a fraction of 400 VMWEs, two thirds are not present in the remaining 1,600 VMWEs. The ratios are rather high, except for Hungarian and Romanian. Although we would expect these scores to have negative correlation with the size of the annotated data, the plot shows great differences even among languages with comparable numbers of annotated VMWEs. We can hypothesise that other factors come into play, such as cross-language variability of domains, text genres and annotation quality.

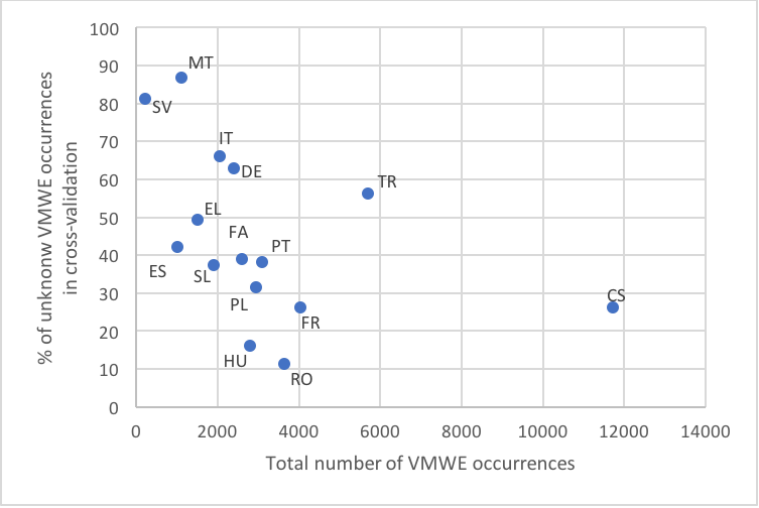

Figure 7: Ratios of unknown VMWEs in the different language datasets. X-axis: the total number of VMWEs tokens in the train+test corpus. Y-axis: average proportion of unknown VMWEs (present in the test but not in the train set) when performing cross-validation with 5 different train/test splits.

We also investigated two other challenging phenomena: overlapping and nesting of VMWEs. The former was measured in terms of the frequency of tokens belonging to at least 2 VMWEs. It occurs – most often due to ellipsis in coordinated VMWEs – in most of the languages but rarely concerns more than two VMWEs at a time, as shown in Table 6. The highest number of overlapping VMWEs was

[29]Matching of VMWEs in train and test sets is performed on lemmatised forms, and with limited normalisation of the order of components (in particular verb-noun for LVCs, and clitic-verb for IReflVs). Note that better normalisation should be performed in order to match multitoken VMWEs against their single-token variants.

five, as seen in (42), where the light verb (PL) *wykonywać* 'perform' is shared by five LVCs.

(42) *Piloci* ***wykonywali*** *podstawowe* ***manewry*** *i serie* ***wznoszeń***,
pilots performed basic maneuvers and series climbs.GEN,
nurkowań, ***pętli*** *i* ***zwrotów***. (PL)
dives.GEN, rolls.GEN and turns.GEN
'The pilots performed basic maneuvers and series of climbs, dives, rolls and turns.'

As far as nesting is concerned, measuring this phenomenon precisely, as defined in §5, would require the availability of syntactic annotations for all languages. Since this is not the case, we approximated nesting at the syntactic level by pairs of VMWEs E_1 and E_2 such that all lexicalised components of E_2 are placed between the left- and right-most lexicalised components of E_1. Single-token VMWEs were disregarded. As the last line of Table 6 shows, such configurations occur very rarely in the data. This might be due to the fact that large gaps introduced within the outer-most VMWEs by the nested structure are harder to process for the human mind.

Table 6: Overlapping and nested VMWEs. Overlap >=2 and >2: the token belongs to at least 2 or more than 2 VMWEs, respectively. Only percentages above 0.49% are indicated. They are counted wrt. all tokens belonging to VMWEs.

	BG	CS	DE	EL	ES	FA	FR	HE	HU	IT	LT	MT	PL	PT	RO	SL	SV	TR
Overlap >= 2	0	520	122	5	22	1	60	235	30	73	0	1	44	65	53	0	1	19
		(1.6%)	(2%)					(5%)		(1.2%)			(0.6%)	(0.5%)	(0.5%)			
Overlap > 2	0	11	0	1	0	0	5	9	0	0	0	0	1	6	0	0	0	0
Nested VMWEs	4	29	1	0	0	0	1	0	1	3	0	0	4	1	0	2	0	0

7 Language-specific studies based on the corpus

Since its publication in January 2017, the PARSEME VMWE-annotated corpus has enabled studies in corpus linguistics in several languages.

The French corpus was addressed by Pasquer (2017), who focuses on the variability of the most frequent VMWEs. Three aspects are studied: (i) morphological

variability of VMWE components, (ii) length and nature of discontinuities between the VMWE components, (iii) syntactic dependencies between the VMWE components and their dependents/governors. The results show a distinctly higher variability in LVCs than in IDs. Namely, nouns inflect and govern external modifiers, respectively, 8 and 1.7 times more often in LVCs (*il* ***rend*** *les derniers* ***hommages*** 'he pays the last tributes') than in IDs. IDs include a lexicalised determiner (*elle* ***tourne la page*** 'she turns the page' vs. *elle* ***joue un role*** 'she plays a role') and a compulsory negation (*ça* ***ne paye pas de mine*** 'it does not pay a face' ⟹ 'it is not much to look at'), 20 and 10 times more often than LVC, respectively. LVCs exhibit discontinuities and passivise 1.5 and 29 times more often than IDs, respectively. Additionally, types of syntactic variants are listed and quantified for the 3 most variable VMWEs. Interesting types of morphological variants, such as prefixations (***redonner raison*** 'to re-give reason' ⟹ 'to admit again that someone is right'), are also revealed.

In Maltese, investigations on LVCs were also carried out in the PARSEME corpus extended with the Maltese UD corpus. The annotated LVCs were extracted and proofread, and the 20 most frequent light verbs (LVs) were listed. Those were used to find other candidate LVCs in a larger raw corpus (not annotated for VMWEs). For each LV the number of unique predicative nouns they combine with could be established. The results show that some LVs are inherently light (e.g. *ta* 'to give', *ħa* 'to take' and *għamel* 'to make/do') and combine with large numbers of nouns (here: 60, 48, and 46, respectively), while others are light only when combined with a few nouns (e.g. *ġarr* 'to carry', *laħaq* 'to reach/achieve', *talab* 'to request/ask'). An analogous experiment, performed for nouns, shows that most of them occur with two LVs (*ta* 'to give' and *ħa* 'to take'), while only few (*appoġġ* 'support', *kura* 'care/treatment' and *kenn* 'shelter') combine with many LVs. Other interesting findings are of etymological nature. Maltese is a language with influences from Semitic and Romance languages, as well as English. The inspected LVCs were mostly of Romance origin (70%), some of Semitic (25%) and some of English (5%). Interestingly, some LVCs accommodate borrowings and Semitic elements that are no longer productive, for example, ***ħa nifs*** 'to take a breath' is ten times more frequent than the Semitic *niffes* 'to breathe'.

LVC-specific analyses were also performed in Lithuanian. Two groups of verbs were identified based on their frequencies in LVCs: (i) 4 high-connectivity verbs i.e. those that combine with large numbers of nouns: *vykdyti* 'to carry out' connects with 19 nouns, *atlikti* 'to perform' – 14, *turėti* 'to have' – 12, *daryti* 'to do/to make' – 10; (ii) 17 low-connectivity verbs i.e. those combining with less than 10 nouns, e.g. *teikti* 'to deliver' – 6, *surengti* 'to arrange' – 4, *imtis* 'to undertake' – 3,

priimti 'to accept' – 3, *patirti* 'to experience' – 3, *duoti* 'to give' – 3, *sudaryti* 'to make' – 3, etc. The numbers of the LVCs containing the verbs from (i) and (ii) are comparable – 55 and 38, respectively – but the diversity of the verbs is significantly higher in (ii) than in (i). The LVCs containing the verbs from group (i) seem to be the most prototypical ones, e.g. ***vykdyti patikrinimus*** 'to carry out inspections', ***atlikti analizę*** 'to perform an analysis', ***daryti spaudimą*** 'to put pressure', etc. These findings pave the way towards developing a comprehensive list of light verbs for Lithuanian.

8 Interesting problems

The considerable collective PARSEME corpus effort led us to confront various phenomena across different language families, various linguistic traditions, and annotation practices. As a result, some interesting findings allow us to view the VMWE phenomenon more globally, which should enable further cross-language generalisations.

Since semantic non-compositionality is the most pervasive property of MWEs, it should possibly be captured by generic definitions and tests in a multilingual endeavour like ours. However, semantic properties show up in different languages via different morphological, syntactic and semantic means. As a result, some semantic non-compositionality phenomena cross word boundaries in some languages, and are therefore relevant to MWEs, and others do not. This distinction can also vary from language to language for the same phenomenon.

For instance, particles in Germanic and Finno-Ugric VPCs, like (EN) *to* ***turn off***, have similar roles as prefixes in Slavic verbs, like (PL) *wy.łączyć* 'to PART.connect' ⇒'to turn off'. The former are traditionally considered separate lexemes, and can therefore form VMWEs with their governing verbs. The latter, conversely, are considered inherent components of verbs, and therefore cannot trigger MWE-related considerations.

Similarly, aspect can be realised by various lexical, morphological and syntactic means, and can therefore be seen as either a semantic or a morphological feature (or both). For instance, perfective or continuous aspect can be introduced by inflection and analytical tenses: (EN) *is doing*, *has done*. Starting, continuation, completion and perfective aspect can also be expressed by specific verbs modifying other verbs: (EN) *to start/continue/stop/complete the action*. Finally, in Slavic languages each verbal lexeme (i.e. independently of its inflected form), has inherent aspect, either perfective or imperfective, and is marked as a morphological feature (recognisable either by a prefix or by an ending): (PL) *robić* 'to do.IMPERF'

vs. *z.robić* 'to PART.do.PERF'; *wy.łączać* 'to PART.connect.IMPERF' ⟹ 'to turn off' vs. *wy.łączyć* 'to PART.connect.PERF' ⟹ 'to turn off'. Therefore, in Slavic languages the verb in an LVC necessarily adds aspect to the predicate, so its status in Test 11 (§4.2.2) should be examined along slightly different lines than in Romance and Germanic languages. Additionally, if adding any aspectual semantics to the predicate should necessarily block the LVC classification in Test 11, then (EN) *to* ***take*** *a* ***decision*** should be annotated as an LVC, while (EN) ***taking*** *a* ***decision*** might not. These observations led us to revise the LVC tests for future editions of the guidelines.

Another finding concerns productivity. Some verbs admit arguments from large semantic classes, and, conversely, some nouns select various verbal operators. More precisely, we observed the hardness of delimiting productive from non-productive cases in VMWE categories: (i) whose semantic non-compositionality is weak, or (ii) whose components are not content words. The former mainly concerns LVCs. We found no effective and reproducible way to distinguish lexical selection from selection of large semantic classes. For instance, (EN) *to deliver* is often used with the class of nouns expressing formal speech acts such as *speech*, *lecture*, *verdict*, etc. However, we can also use the verb *to give* instead of *to deliver* with the same class of nouns, which likely shows a productive rather than a strict lexical selection. Problem (ii) concerns VPCs, IReflVs and prepositional verbs. Namely, as the semantics of particles is hard to establish, we could come up with only one VPC-related test (§4.2.5), which should clearly evolve in future work. Also, the ambiguity of various uses of the reflexive clitic, and the resulting hardness of the IReflV annotation, was stressed by many language teams. Finally, the non-compositionality of prepositional verbs was so hard to establish in the pilot annotation that we abandoned them in the final annotation.

We also underestimated the importance of modelling not only the semantic non-compositionality of idioms but their conventionalisation as well. As a result, we currently have no efficient way to distinguish MWEs from metaphors. The resemblance is strong since many idioms are metaphors, e.g. (PT) *ele* ***abre mão*** 'he opens hand' ⟹ 'he gives up', but non-idiomatic metaphors, created for the need of a particular text, do occur, e.g. (PL) *podpisanie tej umowy to stryczek założony na szyję Polski* 'signing this treaty is a noose put around Poland's neck'. The difference is hard to tackle, and especially to test, since it seems to lie precisely in the fact that MWEs are conventionalised while metaphors are not necessarily so. A partial solution to this problem may probably stem from statistical estimations, although the "long tail" of conventionalised and still infrequent MWEs may largely resemble non-conventionalised metaphors. We put forward the MWE vs. metaphor distinction as a future research issue.

9 Related work

In this section we contextualise our work with respect to existing MWE typologies, annotation methodologies and annotated corpora.

9.1 MWE typologies

In previous approaches to modelling MWEs, various classifications of MWEs were put forward. Here, we focus on several proposals, summarised in Table 7, which seem relevant to our work in that they: (i) have been particularly influential in the NLP community (Sag et al. 2002; Baldwin & Kim 2010; Mel'čuk 2010) (ii) were tested against a representative data set (Mel'čuk 2010), notably in corpus annotation (Schneider et al. 2014), (iii) use MWE flexibility, which is a pervasive feature of verbal MWEs, as a major classification criterion (Sag et al. 2002), (iv) focus exclusively on verbal MWEs (Sheinfux et al. forthcoming), (v) put a verbal component in the heart of the classification criterion (Laporte 2018).

Sag et al. (2002) is a highly influential seminal work whose MWE classification implements the hypothesis put forward by Nunberg et al. (1994) about the correlation between the semantic decomposability of an idiom and its syntactic flexibility. According to this theory, it is because *pull* can be rephrased as *use* and *strings* as *one's influence* that the idiom *to* ***pull strings*** admits variations like *to* ***pull*** *all the (political)* ***strings***, *the* ***strings*** *he* ***pulled***, etc. The hypothesis has been criticised, e.g. by Sheinfux et al. (forthcoming) and Laporte (2018), notably by demonstrating non-decomposable MWEs which still exhibit flexibility. The Sag et al. (2002) classification also calls for adjustments in inflectionally rich and free-word-order languages. Still, it remains widely used, notably due to its usefulness for NLP applications. Namely, MWE flexibility is a major obstacle in MWE identification since it prohibits seeing a MWE as a "word with spaces" and using sequence labelling approaches.

Baldwin & Kim (2010) assume the flexibility-driven classification by Sag et al. (2002) and they additionally introduce an orthogonal typology based on purely syntactic criteria, that is, on the syntactic structure of the MWE. There, verbal subcategories are both English-specific and non-exhaustive since verb-noun idioms are considered, but not, for example, verb-adjective ones.

The typology of Mel'čuk (2010) is based, conversely, on mainly semantic criteria. Different types of semantic compositionality are defined, and non-compositional subtypes are those where the semantic head is missing. The latter further subdivide into: (i) *quasi-locutions* in which the meanings of the components are combined, as in (FR) ***donner le sein*** 'to give the breast' ⟹ 'to breastfeed', (ii)

Table 7: Various MWE classifications compared.

Reference	Language	Scope	Classes	# classified expressions	Defining criteria
Sag et al. (2002)	EN	MWEs and collocations	I. Lexicalised: 1. Fixed (*by and large*); 2. Semi-fixed: non-decomposable idioms (*shoot the breeze* 'chat'), compound nominals (*part of speech*), proper names (*San Francisco 49ers*); 3. Syntactically-flexible: VPCs (*break up*), decomposable idioms (*spill the beans*), LVCs (*make a decision*); II. Institutionalised (*traffic lights*)	unknown	lexicalisation, morphological and syntactic flexibility, semantic decomposability
Baldwin & Kim (2010)	EN	MWEs and collocations	I. Nominal (*golf club*, *connecting flight*); II. Verbal: 1. VPCs (*take off*, *cut short*, *let go*); 2. Prepositional verbs (*come accross*); 3. LVCs (*take a walk*); 4. Verb-noun idioms (*shoot the breeze*); III. Prepositional: 1. Determinerless prepositional phrases (*on top*, *by car*); 2. Complex prepositions (*on top of*, *in addition to*)	unknown	syntactic structure
Mel'čuk (2010)	FR	MWEs and collocations	I. Pragmatic (*emphasis mine*); II. Semantic: 1. Semantically compositional: clichés (*in other words*), collocations (*busy as a bee*, *award a prize*); 2. Semantically non-compositional: quasi-locutions ((FR) *donner le sein* 'give the breast' ⇒ 'breastfeed'), 2. Semi-locutions ((FR) *fruits de mer* 'sea fruit' ⇒ 'seafood'), 3. Complete locutions ((FR) *en tenue d'Adam et Eve* 'in Adam's and Eve's dress' ⇒ 'naked')	4,400 collocations, 3,200 locutions (Pausé 2017)	selection constraints, semantic non-compositionality
Schneider et al. (2014)	EN	all MWEs	I. Strong (*close call*); II. Weak (*narrow escape*)	3,500 occurrences	strength of association between words
Sheinfux et al. (forthcoming)	HE	verbal idioms	I. Transparent figurative (*saw logs* 'snore'); II. Opaque figurative (*shoot the breeze* 'chat'); III. Opaque non-figurative (*take umbrage* 'feel offended')	15 VMWEs, 400 occurrences	transparency, figuration
Laporte (2018)	FR	MWEs and collocations	I. Lexicalised: 1. MWEs without support verbs: verbal (*take stock*), nominal (*traffic lights*), adverbial (*for instance*); 2. Support-verb constr.: a. Vsup is not copula (*have an aim*, *get loose*), b. Vsup in copula (*be a genius*, *be angry*, *be on time*); II. Non-lexicalised (*salt and pepper*)	dozens of thousands of (lexicalised) MWEs	lexicalisation, presence of a support verb
This chapter	BG,CS,DE,EL, ES,FA,FR,HE, HU,IT,LT,MT, PL,PT,RO,SL, SV,TR	verbal MWEs	I. Universal: LVCs (*make a decision*), IDs (*spill the beans*); II. Quasi-universal: IReflVs ((FR) *s'avérer* 'REFL'reveal' ⇒ 'prove (to be)'), VPCs (*take off*); III. OTH (*drink and drive*, *to voice act*)	62,000 occurrences	universalism, syntactic structure, lexical, syntactic and semantic idiosyncrasy

semi-locutions which include the meaning of only a part of their components, as in (FR) *fruits de mer* 'sea fruit' ⇒ 'seafood', (iii) *complete locutions*, which include the meaning of none of their components, as in (FR) *en tenue d'Adam et Eve* 'in Adam's and Eve's dress' ⇒ 'naked'.

Schneider et al. (2014) propose a rather shallow typology with only two types based on the strength of association between component words. Strong MWEs are those whose meaning is not readily predictable from component words, as in (EN) *close call* 'a situation in which something bad almost happened but could be avoided'. Weak MWEs are those with more transparent semantics and more flexibility, like (EN) *narrow escape* 'a situation in which something bad almost happened but could be avoided'. This typology was applied to annotate a large publicly available corpus, underlying the DiMSUM[30] shared task on identification of minimal semantic units and their supersenses.

In Sheinfux et al. (forthcoming) the hypothesis of Nunberg et al. (1994) is questioned on a sample of verbal Hebrew idioms, and a novel classification is put forward which relies on figuration (the degree to which the idiom can be assigned a literal meaning) and transparency (the relationship between the literal and idiomatic reading). In *transparent figurative* idioms the relationship between the literal and the idiomatic reading is easy to recover (*to* ***saw logs*** 'snore'). In *opaque figurative* idioms the literal picture is easy to imagine but its relationship to the idiomatic reading is unclear (*to* ***shoot the breeze*** 'chat'). Finally, in *opaque non-figurative* idioms no comprehensible literal meaning is available, notably due to cranberry words which have no status as individual lexical units (*to* ***take umbrage*** 'to feel offended'). The study further tests VMWEs of the 3 categories against 4 types of lexical and syntactic flexibility, and stresses the fact that flexibility is a matter of scale rather than a binary property.

Laporte (2018) formalises a MWE classification emerging from the lexicon-grammar theory and encoding practice (Gross 1986; 1994). Its specificity is to put the notion of support verb (roughly equivalent to light verb) in the heart of the classification, and push the MWE frontier far beyond what is admitted in other approaches. Namely, with the copula support verb *to be*, large classes of nouns, adjectives and PPs are seen as predicates of support-verb constructions, which should, thus, be lexically described.

Comparing our classification (§3) to the above ones (Table 7), several facts are striking: (i) we restrict ourselves to verbal MWEs only, (ii) we perform a large-scale multilingual evaluation and enhancement of the classification via corpus annotation in 18 languages, (iii) we assess semantic non-compositionality via

[30]https://dimsum16.github.io/

mostly syntactic tests, (iv) we define a novel VMWE category of IReflVs and linguistic tests delimiting its borders, we also display the quantitative importance of this category, mainly in Romance and Slavic languages, (v) we give access to detailed annotation guidelines organised as decision trees, with linguistic tests illustrated in many languages. As far as the scope of the MWE-related phenomena are concerned, recall that we exclude statistical collocations and retain only lexically, syntactically or semantically idiosyncratic expressions. This fact seemingly contrasts with other approaches shown in Table 7. Note, however, that some of these authors understand collocations differently, as discussed in §2.

9.2 MWE annotation practices

Modelling the behaviour of MWEs in annotated corpora, and prominently in treebanks, has been undertaken in various languages and linguistic frameworks. Rosén et al. (2015) offer a survey of MWE annotation in 17 treebanks for 15 languages, collaboratively documented according to common guidelines.[31] According to this survey, multiword named entities constitute by far the most frequently annotated category (Erjavec et al. 2010), sometimes with elaborate annotation schemes accounting for nesting and coordination (Savary et al. 2010). Continuous MWEs such as compound nouns, adverbs, prepositions and conjunctions are also covered in some corpora (Abeillé et al. 2003; Laporte et al. 2008; Branco et al. 2010). Verbal MWEs have been addressed for fewer languages. The survey also shows the heterogeneity of MWE annotation practices. For instance, VPCs are represented in dependency treebanks by dedicated relations between head verbs and particles. In constituency treebanks, particles constitute separate daughter nodes of sentential or verbal phrases and are assigned categories explicitly indicating their status of selected particles. Additionally, in an LFG (Lexical Functional Grammar) treebank, verbs and their particles are merged into single predicates appearing in functional structures.

Similar conclusions about the heterogeneity of MWE annotation were drawn concerning UD (McDonald et al. 2013), an initiative towards developing syntactically full-fledged and cross-linguistically consistent treebank annotation for many languages. Nivre & Vincze (2015) show that LVCs annotation in UD treebanks is threefold: (i) some treebanks lack or do not distinguish LVCs from regular verb-object pairs, (ii) some distinguish them by their structure (the direct object is dependent on the light verb rather than on the predicative noun), (iii) some account for them explicitly by the dependency labels between the noun

[31] http://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme

and the verb. Furthermore, De Smedt et al. (2015) point out that 3 different dependency relations in UD[32] can be used to describe MWEs - `compound`, `mwe` and `name` (with possible sub-relations, e.g. `compound:prt` for verb-particle constructions) - and that these are used across different UD treebanks in a largely inconsistent way. More recent efforts (Adalı et al. 2016), while addressing VMWEs in a comprehensive way, still suffer from missing annotation standards.

As compared to this state of the art, the PARSEME effort aims at developing annotation guidelines and practices which would be universal but would leave room for language-dependent specificities. Our scope covers all types of VMWEs.

9.3 Corpora and datasets with VMWEs

As seen in the previous section, most efforts towards anotating MWEs were either language- or MWE category-specific. The same holds for verbal MWEs in particular. In this section we mention some outcomes of the previous VMWE annotation initiatives.

The Wiki50 (Vincze et al. 2011) corpus contains 50 English Wikipedia articles annotated for MWEs, including several VMWEs types. The dataset of Tu & Roth (2011) consists of 2,162 sentences from the British National Corpus in which verb-object pairs formed with *do*, *get*, *give*, *have*, *make*, and *take* are marked as positive and negative examples of LVCs. Tu & Roth (2012) built a crowdsourced corpus in which VPCs are manually distinguished from compositional verb-preposition combinations, again for six selected verbs. Baldwin (2005) presents another dataset of English VPCs. Finally, SZPFX (Vincze 2012) is an English-Hungarian parallel corpus with LVC annotations in both languages. For German, idiomatic combinations of verbs and prepositional phrases were described in a database by Krenn (2008) and annotated in the TIGER corpus by Brants et al. (2005).

In Slavic languages, a notable effort was made with the Prague Dependency Treebank of Czech (Hajič et al. 2017), annotated at 3 layers: morphological, analytical (accounting for syntax) and tectogrammatical (accounting for functional relations). MWEs, including some VMWEs, are annotated by identifying monosemic subtrees in the 3rd layer and replacing them by single nodes (Bejček & Straňák 2010), which unifies different morphosyntactic variants of the same MWE (Bejček et al. 2011). Each MWE occurrence is linked to its entry in an associated MWE lexicon. It is also argued that elements elided in MWEs (e.g. due to coordination) should be restored in deep syntactic trees. The Czech PARSEME corpus results from a mostly automatic (although challenging) transformation of the PDT annotations into the parseme-tsv format (Bejček et al. 2017).

[32] This analysis concerns UD v1 - these labels evolved in UD v2.

Kaalep & Muischnek (2006; 2008) and Vincze & Csirik (2010) present databases and corpora of VMWEs for Estonian particle verbs and Hungarian LVCs, respectively. VMWE annotations are available in several Turkish treebanks. In Eryiğit et al. (2015) various MWEs are labeled with a unique dependency label independently of their category, while in Adalı et al. (2016) they are classified as either strong or weak, similarly to Schneider et al. (2014). Finally, QasemiZadeh & Rahimi (2006) provide annotations for Farsi LVCs in the framework of the MULTEXT-East initiative, and in the Uppsala Persian Dependency Treebank (Seraji et al. 2014) the *lvc* dependency relationship is used for annotating non-verbal component of Farsi LVCs that are not in any other type of syntactic relationship.

The PARSEME corpus initiative builds upon these previous efforts by incorporating and extending some pre-existing datasets and annotation experiences. In some languages it is novel in that: (i) it constitutes the first attempt to annotate and analyse VMWEs in running text, e.g. in Greek and Maltese, (ii) it pays special attention, for the first time, to certain VMWE categories, e.g. to VPCs in Greek, to LVCs in Lithuanian, to IReflVs in most Slavic and Romance languages, and to distinguishing VMWEs from semi-copula-based expressions in Farsi (§4.3). But the most notable achievement going beyond the state of the art is to offer the first large highly multilingual VMWE corpus annotated according to unified guidelines and methodologies.

10 Conclusions and future work

We described the results of a considerable collective effort towards setting up a common framework for annotating VMWEs in 18 languages from 9 different language families. Unlike McDonald et al. (2013), our methodology is not English-centred. We draft the guidelines and test them on many languages in parallel, without giving priority to any of them (except for communication purposes). We offer a classification of VMWEs where properties hypothesised as universal or quasi-universal are treated in a homogeneous way, while leaving room to language-specific categories and features at the same time. Additionally to its importance for language modelling, and contrastive linguistic studies, this typology may be useful for various language technology tasks, notably because different VMWE types show different degrees of semantic decomposability, which influences their interpretation and translation. For instance, in LVCs nouns may translate literally and verbs may be omitted in the semantic calculus, but the same usually does not hold for IDs. Our annotation guidelines are organised in decision trees, so as to maximise the replicability of the annotators' decisions.

Our efforts also pave the way towards unified terminology and notation conventions. In particular, we stress the relations between words and tokens, which are crucial for defining the scope of the MWE phenomenon. We formalise the notion of a canonical form of a VMWE. Moreover, the notational conventions used in this volume for citing, glossing and translating multilingual examples of VMWEs largely result from our documentation work.

The PARSEME VMWE corpus[33] and its annotation guidelines,[34] both available under open licenses, are meant as dynamic resources, subject to continuous enhancements and updates. The size of the corpus is still modest for many languages and should be progressively increased. Adopting higher annotation standards, including a double annotation and adjudication, would lead to more reliable guidelines, increase the quality of the data, and strengthen our claims and findings. Since the publication of version 1.0 of the corpus, rich feedback was gathered from language teams, several dozens of issues were formulated and were discussed in a dedicated Gitlab space[35] and version 1.1[36] of the guidelines was elaborated. The most important evolutions include:

- Abandoning the category-neutral identification stage, since the annotation practice showed that VMWE identification is virtually always done in a category-specific way. The previous identification tests become ID-specific tests.
- Abandoning the OTH category due to its very restricted use. VMWEs classified previously as OTH now enter the ID category (except when the interpretation of the OTH category was language-specific).
- Introducing the multiverb construction (MVC) category to account for idiomatic serial verbs in Asian languages such as Hindi, Indonesian, Japanese and Chinese.
- Redesigning the tests and the decision trees for the LVC and VPC category, so as to increase the determinism in the annotation of these two categories.
- Introducing – optionally and experimentally – the category of inherently adpositional verbs (IAVs), roughly equivalent to the previously abandoned

[33] http://hdl.handle.net/11372/LRT-2282
[34] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/
[35] https://gitlab.com/parseme/sharedtask-guidelines/issues (restricted access, new users are welcome upon registration with the project leaders)
[36] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/

inherently prepositional verbs (IPrepVs). The IAV should be addressed in the post-annotation step, i.e. once the VMWEs of all other categories have been identified.

- Renaming the IReflV category by IRV, for an easier pronunciation.
- Renaming the ID category to VID (verbal idiom), to explicitly account for the verbal-only scope.

Adjustments of the previously annotated corpus to the guidelines version 1.1 are ongoing. The corpus should also significantly grow, as new portions of data are being annotated and new language teams (Arabic, Basque, Croatian, English and Hindi) are joining the project. Edition 1.1 of the PARSEME shared task (cf. Savary et al. 2017 for edition 1.0), based on the enhanced guidelines and corpus, is taking place as this volume is being edited.

In the long run, we intend to include other categories of MWEs (nominal, adjectival, adverbial, prepositional, named entities, etc.) under the annotation scope, as well as pave the way towards consistent representation and processing of both MWEs and syntax.

Acknowledgments

The work described in this chapter was supported by: (i) the IC1207 PARSEME COST action,[37]; (ii) national funded projects: LD-PARSEME[38] (LD14117) in the Czech Republic, PARSEME-FR[39] (ANR-14-CERA-0001) in France, and PASTO-VU[40] in Lithuania; (iii) European Union's Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant No 713567); (iv) Science Foundation Ireland in the ADAPT Centre[41] (Grant 13/RC/2106) at Dublin City University.

We are grateful to all language teams for their contributions to preparing the annotation guidelines and the annotated corpora. The full composition of the annotation team is the following.

Balto-Slavic languages:

- (BG) Ivelina Stoyanova (LGL, LL), Tsvetana Dimitrova, Svetla Koeva, Svetlozara Leseva, Valentina Stefanova, Maria Todorova;

[37] http://www.parseme.eu
[38] https://ufal.mff.cuni.cz/grants/ld-parseme
[39] http://parsemefr.lif.univ-mrs.fr/
[40] http://mwe.lt/en_US/
[41] www.adaptcentre.ie

- (CS) Eduard Bejček (LL), Zdeňka Urešová, Milena Hnátková;
- (LT) Jolanta Kovalevskaitė (LL), Loic Boizou, Erika Rimkutė, Ieva Bumbulienė;
- (SL) Simon Krek (LL), Polona Gantar, Taja Kuzman;
- (PL) Agata Savary (LL), Monika Czerepowicka.

Germanic languages:

- (DE) Fabienne Cap (LGL, LL), Glorianna Jagfeld, Agata Savary;
- (EN) Ismail El Maarouf (LL), Teresa Lynn, Michael Oakes, Jamie Findlay, John McCrae, Veronika Vincze;
- (SV) Fabienne Cap (LL), Joakim Nivre, Sara Stymne.

Romance languages:

- (ES) Carla Parra Escartín (LL), Cristina Aceta, Itziar Aduriz, Uxoa Iñurrieta, Carlos Herrero, Héctor Martínez Alonso, Belem Priego Sanchez;
- (FR) Marie Candito (LGL, LL), Matthieu Constant, Ismail El Maarouf, Carlos Ramisch (LGL), Caroline Pasquer, Yannick Parmentier, Jean-Yves Antoine;
- (IT) Johanna Monti (LL), Valeria Caruso, Manuela Cherchi, Anna De Santis, Maria Pia di Buono, Annalisa Raffone;
- (RO) Verginica Barbu Mititelu (LL), Monica-Mihaela Rizea, Mihaela Ionescu, Mihaela Onofrei;
- (PT) Silvio Ricardo Cordeiro (LL), Aline Villavicencio, Carlos Ramisch, Leonardo Zilio, Helena de Medeiros Caseli, Renata Ramisch;

Other languages:

- (EL) Voula Giouli (LGL,LL), Vassiliki Foufi, Aggeliki Fotopoulou, Sevi Louisou;
- (FA) Behrang QasemiZadeh (LL);
- (HE) Chaya Liebeskind (LL), Yaakov Ha-Cohen Kerner (LL), Hevi Elyovich, Ruth Malka;
- (HU) Veronika Vincze (LL), Katalin Simkó, Viktória Kovács;
- (MT) Lonneke van der Plas (LL), Luke Galea (LL), Greta Attard, Kirsty Azzopardi, Janice Bonnici, Jael Busuttil, Ray Fabri, Alison Farrugia, Sara Anne Galea, Albert Gatt, Anabelle Gatt, Amanda Muscat, Michael Spagnol, Nicole Tabone, Marc Tanti;
- (TR) Kübra Adalı (LL), Gülşen Eryiğit (LL), Tutkum Dinç, Ayşenur Miral, Mert Boz, Umut Sulubacak.

We also thank Mozhgan Neisani from University of Isfahan and Mojgan Seraji from the Uppsala Universitet for their contribution to the inter-annotator agreement calculation.

Abbreviations

FUT	future	MTW	multitoken word
GEN	genitive	MWT	multiword token
IAA	inter-annotator-agreement	NLP	natural language processing
ID	idiom	OTH	other VMWEs
IREFLV	inherently reflexive verb	PART	particle
LGL	language group leader	REFL	reflexive clitic
LL	language leader	SG	singular
LV	light verb	UD	Universal Dependencies
LVC	light-verb construction	VID	verbal idiom
MAD	mean absolute deviation	VMWE	verbal multiword expression
MASC	masculine	VPC	verb-particle construction
MWE	multiword expression	1, 2, 3	first, second, third person

References

Abeillé, Anne, Lionel Clément & François Toussenel. 2003. Building a treebank for French. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 165–187. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Adalı, Kübra, Tutkum Dinç, Memduh Gokirmak & Gülşen Eryiğit. 2016. Comprehensive annotation of multiword expressions for Turkish. In *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*.

Al Saied, Hazem, Marie Candito & Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 209–226. Berlin: Language Science Press. DOI:10.5281/zenodo.1469561

Baggio, Giosuè, Michiel van Lambalgen & Peter Hagoort. 2012. The processing consequences of compositionality. In *The Oxford handbook of compositionality*, 655–672. New York: Oxford University Press.

Baldwin, Timothy. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language* 19(4). 398–414. DOI:10.1016/j.csl.2005.02.004

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing, Second edition*, 267–292. Boca Raton: CRC Press.

Bauer, Laurie. 1983. *English word-formation* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press. https://books.google.pl/books?id=yGfUHs6FCvIC.

Bejček, Eduard, Jan Hajič, Pavel Straňák & Zděnska Urěnsová. 2017. Extracting verbal multiword data from rich treebank annotation. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories* (TLT 15), 13–24.

Bejček, Eduard & Pavel Straňák. 2010. Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation* 44(1–2). 7–21.

Bejček, Eduard, Pavel Straňák & Daniel Zeman. 2011. Influence of treebank design on representation of multiword expressions. In Alexander F. Gelbukh (ed.), *Computational Linguistics and intelligent text processing - 12th International Conference,* (CICLing 2011), Tokyo, Japan, February 20-26, 2011. *Proceedings, Part I*, vol. 6608 (Lecture Notes in Computer Science), 1–14. Springer. DOI:10.1007/978-3-642-19400-9_1

Branco, António, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto & João Graça. 2010. Developing a deep linguistic databank supporting a collection of treebanks: The CINTIL DeepGramBank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th conference on international language resources and evaluation* (LREC 2010). European Language Resources Association (ELRA).

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit. 2005. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation* 2(4). 597–620. DOI:10.1007/s11168-004-7431-3

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), 4585–4592. European Language Resources Association (ELRA).

De Smedt, Koenraad, Victoria Rosén & Paul Meurer. 2015. *MWEs in universal dependency treebanks*. IC1207 COST PARSEME 5th general meeting. Iaşi, Ro-

mania. http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015.

Erjavec, Tomaz, Darja Fiser, Simon Krek & Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC 2010), 1806–1809. European Language Resources Association (ELRA).

Eryiğit, Gülşen, Kübra Adali, Dilara Torunoğlu-Selamet, Umut Sulubacak & Tuğba Pamay. 2015. Annotation and extraction of multiword expressions in Turkish treebanks. In *Proceedings of the 11th Workshop on Multiword Expressions* (MWE '15), 70–76. Association for Computational Linguistics. http://www.aclweb.org/anthology/W15-0912.

Fort, Karën & Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the 4th Linguistic Annotation Workshop* (LAW IV '10), 56–63. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1868720.1868727.

Gross, Gaston. 1988. Degré de figement des noms composés. *Langages* 90. 57–71.

Gross, Maurice. 1986. Lexicon-grammar: The representation of compound words. In *Proceedings of the 11th coference on computational linguistics* (COLING '86), 1–6. Association for Computational Linguistics. DOI:10.3115/991365.991367

Gross, Maurice. 1994. The lexicon-grammar of a language: Application to French. In Ashley R. E. (ed.), *The encyclopedia of language and linguistics*, 2195–2205. Oxford: Oxford/NewYork/Seoul/Tokyo: Pergamon. https://hal-upec-upem.archives-ouvertes.fr/hal-00621380.

Hajič, Jan, Eva Hajičová, Marie Mikulová & Jiří Mírovský. 2017. Prague Dependency Treebank. In *Handbook on Linguistic Annotation* (Springer Handbooks), 555–594. Berlin, Germany: Springer Verlag.

Janssen, Theo M. V. 2001. Frege, contextuality and compositionality. *Journal of Logic, Language and Information* 10(1). 115–136. DOI:10.1023/A:1026542332224

Jespersen, Otto. 1965. *A Modern English grammar on historical principles, Part VI, Morphology*. London: Allen & Unwin.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2006. Multi-word verbs in a flective language: The case of Estonian. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Context* (MWE '06), 57–64. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W06/W06-2400.pdf.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2008. Multi-word verbs of Estonian: A database and a corpus. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions* (MWE '08), 23–26. Association for Computational Linguistics. http : / / www . lrec - conf . org / proceedings / lrec2008 / workshops/W20_Proceedings.pdf.

Kim, Su Nam. 2008. *Statistical modeling of multiword expressions*. Melbourne: University of Melbourne dissertation.

Koeva, Svetla, Ivelina Stoyanova, Maria Todorova & Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of Bulgarian multiword expressions. In *Proceedings of the Workshop on Lexicographic Resources for Human Language Technology* (GLOBALEX 2016), 86–95.

Kracht, Marcus. 2007. Compositionality: The very idea. *Research on Language and Computation* 5(3). 287–308. DOI:10.1007/s11168-007-9031-5

Krenn, Brigitte. 2008. Description of evaluation resource – German PP-verb data. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions* (MWE '08), 7–10. Association for Computational Linguistics.

Laporte, Éric. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective* (Phraseology and Multiword Expressions). Language Science Press.

Laporte, Éric, Takuya Nakamura & Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Proceedings of the 2nd Linguistic Annotation Workshop*, 48–51. https : / / halshs . archives - ouvertes.fr/halshs-00286541.

Lipka, Leonhard, Susanne Handl & Wolfgang Falkner. 2004. Lexicalization & institutionalization. The state of the art in 2004. *SKASE Journal of Theoretical Linguistics* 1(1). 2–19. http://www.skase.sk/Volumes/JTL01/lipka.pdf.

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

Marcus, Mitchell P., Mary Ann Marcinkiewicz & Beatrice Santorini. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics* 19(2). 313–330. http://dl.acm.org/citation.cfm?id=972470.972475.

Mathet, Yann, Antoine Widlöcher & Jean-Philippe Métivier. 2015. The unified and holistic method Gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics* 41(3). 437–479. DOI:10.1162/COLI_a_00227

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)*, 92–97. Association for Computational Linguistics. http://www.aclweb.org/anthology/P13-2017.

Mel'čuk, Igor A. 2010. La phraséologie en langue, en dictionnaire et en TALN. In *Actes de la 17ème conférence sur le traitement automatique des langues naturelles 2010*.

Moreau, Erwan, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel & Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 177–207. Berlin: Language Science Press. DOI:10.5281/zenodo.1469559

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Nivre, Joakim & Veronika Vincze. 2015. *Light verb constructions in universal dependencies*. IC1207 COST PARSEME 5th general meeting. Iaşi, Romania. http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015.

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

Pagin, Peter & Dag Westerståhl. 2001. Compositionality II: Arguments and problems. *Philosophy Compass* 5. 250–264. DOI:10.1111/j.1747-9991.2009.00228.x

Partee, Barbara H., Alice ter Meulen & Robert E. Wall. 1990. *Mathematical methods in linguistics* (Studies in Linguistics and Philosophy 30). Dordrecht: Kluwer.

Pasquer, Caroline. 2017. Expressions polylexicales verbales : étude de la variabilité en corpus. In *Actes de la 18e rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues* (TALN-RÉCITAL 2017).

Pausé, Marie-Sophie. 2017. *Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire*. Nancy, France: Université de Lorraine dissertation.

QasemiZadeh, Behrang & Saeed Rahimi. 2006. Persian in MULTEXT-East framework. In *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August 23-25, 2006, Proceedings*, 541–551. DOI:10.1007/11816508_54

Rosén, Victoria, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova & Verginica Barbu Mitetelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories Cconference*. https://hal.archives-ouvertes.fr/hal-01226001.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Savary, Agata, Jakub Waszczuk & Adam Przepiórkowski. 2010. Towards the annotation of named entities in the National Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th conference on international language resources and evaluation* (LREC 2010). European Language Resources Association (ELRA).

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Nora Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Smith Noah A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evalua-*

tion (LREC 2014), 455–461. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf.

Seraji, Mojgan, Carina Jahani, Beáta Megyesi & Joakim Nivre. 2014. A Persian treebank with Stanford typed dependencies. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014). European Language Resources Association (ELRA).

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. Forthcoming. Verbal MWEs: Idiomaticity and flexibility. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions* (Phraseology and Multiword Expressions), 5–38. Berlin: Language Science Press.

Siemieniec-Gołaś, Ewa. 2010. On some Turkish auxiliary verbs in giovanni molino's dittionario della lingua italiana, turchesca (1641). *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 127(1). 57–77.

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 227–243. Berlin: Language Science Press. DOI:10.5281/zenodo.1469563

Szarvas, György, Veronika Vincze, Richárd Farkas, György Móra & Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics – Special Issue on Modality and Negation* 38(2). 335–367.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 299–317. Berlin: Language Science Press. DOI:10.5281/zenodo.1469569

Tu, Yuancheng & Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE '11), 31–39. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0807.

Tu, Yuancheng & Dan Roth. 2012. Sorting out the most confusing English phrasal verbs. In *Proceedings of the First Joint Conference on Lexical and Computational*

Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval '12), 65–69. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2387636.2387648.

van Gompel, Maarten & Martin Reynaert. 2013. FoLia: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal* 3. 63–81.

van Gompel, Maarten, Ko van der Sloot, Martin Reynaert & Antal van den Bosch. 2017. FoLiA in practice: The infrastructure of a linguistic annotation format. In, (To appear). Ubiquity Press.

Vincze, Veronika. 2012. Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC-2012), 2381–2388. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/177_Paper.pdf.

Vincze, Veronika & János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING '10), 1110–1118. Association for Computational Linguistics. http://www.aclweb.org/anthology/C10-1125.

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 289–295. RANLP 2011 Organising Committee. http://aclweb.org/anthology/R11-1040.