The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions

Agata Savary
Université de Tours, France
first.last@univ-tours.fr

Carlos Ramisch
Aix Marseille Université
France

Silvio Ricardo Cordeiro
Aix Marseille Université
France

Federico Sangati
Independent researcher
Italy

Veronika Vincze
University of Szeged
Hungary

Behrang QasemiZadeh
University of Düsseldorf
Germany

Marie Candito
Université Paris Diderot
France

Fabienne Cap
Uppsala University
Sweden

Voula Giouli
Athena Research Center
Athens, Greece

Ivelina Stoyanova
Bulgarian Academy of Sciences
Sofia, Bulgaria

Antoine Doucet
University of La Rochelle
France

Abstract

Multiword expressions (MWEs) are known as a "pain in the neck" for NLP due to their idiosyncratic behaviour. While some categories of MWEs have been addressed by many studies, verbal MWEs (VMWEs), such as *to take a decision*, *to break one's heart* or *to turn off*, have been rarely modelled. This is notably due to their syntactic variability, which hinders treating them as "words with spaces". We describe an initiative meant to bring about substantial progress in understanding, modelling and processing VMWEs. It is a joint effort, carried out within a European research network, to elaborate universal terminologies and annotation guidelines for 18 languages. Its main outcome is a multilingual 5-million-word annotated corpus which underlies a shared task on automatic identification of VMWEs. This paper presents the corpus annotation methodology and outcome, the shared task organisation and the results of the participating systems.

1 Introduction

Multiword expressions (MWEs) are known to be a "pain in the neck" for natural language processing (NLP) due to their idiosyncratic behaviour (Sag et al., 2002). While some categories of MWEs have been addressed by a large number of NLP studies, verbal MWEs (VMWEs), such as *to* ***take*** *a* ***decision***, *to* ***break*** *one's* ***heart*** or *to* ***turn off***[1], have been relatively rarely modelled. Their particularly challenging nature lies notably in the following facts:

1. Their components may not be adjacent (***turn*** *it* ***off***) and their order may vary (*the* ***decision*** *was hard to* ***take***);
2. They may have both an idiomatic and a literal reading (*to* ***take the cake***);
3. Their surface forms may be syntactically ambiguous (*on* is a particle in the verb-particle construction ***take on*** *the task* and a preposition in *to sit on the chair*);
4. VMWEs of different categories may share the same syntactic structure and lexical choices (*to* ***make*** *a* ***mistake*** is a light-verb construction, *to* ***make a meal*** is an idiom),
5. VMWEs behave differently in different languages and are modelled according to different linguistic traditions.

These properties are challenging for automatic identification of VMWEs, which is a prerequisite for MWE-aware downstream applications such as

[1] Henceforth, boldface will be used to highlight the lexicalised components of MWEs, that is, those that are always realized by the same lexemes.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 31–47,
Valencia, Spain, April 4. ©2017 Association for Computational Linguistics

parsing and machine translation. Namely, challenge 1 hinders the use of traditional sequence labelling approaches and calls for syntactic analysis. Challenges 2, 3 and 4 mean that VMWE identification and categorization cannot be based on solely syntactic patterns. Challenge 5 defies cross-language VMWE identification.

We present an initiative aiming at boosting VMWE identification in a highly multilingual context. It is based on a joint effort, carried on within a European research network, to elaborate universal terminologies, guidelines and methodologies for 18 languages. Its main outcome is a 5-million-word corpus annotated for VMWEs in all these languages, which underlies a shared task on automatic identification of VMWEs.[2] Participants of the shared task were provided with training and test corpora, and could present systems within two tracks, depending on the use of external resources. They were encouraged to submit results for possibly many covered languages.

In this paper, we describe the state of the art in VMWE annotation and identification (§ 2). We then present the corpus annotation methodology (§ 3) and its outcome (§ 4). The shared task organization (§ 5), the measures used for system evaluation (§ 6) and the results obtained by the participating systems (§ 7) follow. Finally, we discuss conclusions and future work (§ 8).

2 Related Work

Annotation There have been several previous attempts to annotate VMWEs. Some focus specifically on VMWEs and others include them among the linguistic phenomena to be annotated. Rosén et al. (2015) offer a survey of VMWE annotation in 17 treebanks, pointing out that, out of 13 languages in which phrasal verbs do occur, 8 have treebanks containing annotated phrasal verbs, and only 6 of them contain annotated light-verb constructions and/or verbal idioms. They also underline the heterogeneity of these MWE annotations. Nivre and Vincze (2015) show that this is also the case in the treebanks of Universal Dependencies (UD), despite the homogenizing objective of the UD project (McDonald et al., 2013). More recent efforts (Adalı et al., 2016), while addressing VMWEs in a comprehensive way, still suffer from missing annotation standards.

Heterogeneity is also striking when reviewing annotation efforts specifically dedicated to VMWEs, such as Estonian particle verbs (Kaalep and Muischnek, 2006; Kaalep and Muischnek, 2008), Hungarian light-verb constructions (Vincze and Csirik, 2010), and Arabic verb-noun and verb-particle constructions (Bar et al., 2014). The same holds for English resources, such as the Wiki50 corpus (Vincze et al., 2011), which includes both verbal and non-verbal MWEs. Resources for English also include data sets of selected sentences with positive and negative examples of light-verb constructions (Tu and Roth, 2011), verb-noun combinations (Cook et al., 2008), and verb-particle constructions (Tu and Roth, 2012). While most annotation attempts mentioned so far focus on annotating MWEs in running texts, there also exist lists of MWEs annotated with their degree of idiomaticity, for instance, German particle verbs (Bott et al., 2016) and English noun compounds (Reddy et al., 2011). In contrast to these seminal efforts, the present shared task relies on VMWE annotation in running text according to a unified methodology.

Identification MWE identification is a well-known NLP task. The 2008 MWE workshop proposed a first attempt of an MWE-targeted shared task. Differently from the shared task described here, the goal of participants was to rank provided MWE candidate lexical units, rather than to identify them in context. True MWEs should be ranked towards the top of the list, whereas regular word combinations should be at the end. Heterogeneous datasets containing several MWE categories in English, German and Czech were made available. Two systems participated, using different combinations of features and machine learning classifiers. In addition to the shared task, the MWE 2008 workshop also focused on gathering and sharing lexical resources containing annotated candidate MWEs. This repository is available and maintained on the community website.[3]

The DiMSUM 2016 shared task (Schneider et al., 2016) challenged participants to label English sentences (tweets, service reviews, and TED talk transcriptions) both with MWEs and supersenses for nouns and verbs.[4] The provided dataset is made of approximately 90,000 tokens containing 5,069 annotated MWEs, about 10% of which are

[2] http://multiword.sourceforge.net/sharedtask2017

[3] http://multiword.sf.net/

[4] http://dimsum16.github.io

discontinuous. They were annotated following Schneider et al. (2014b), and thus contain several VMWEs types on top of non-verbal MWEs.

Links between MWE identification and syntactic parsing have also long been an issue. While the former has often been treated as a pre-processing step before the latter, both tasks are now more and more often integrated, in particular for continuous MWE categories (Finkel and Manning, 2009; Green et al., 2011; Green et al., 2013; Candito and Constant, 2014; Le Roux et al., 2014; Nasr et al., 2015; Constant and Nivre, 2016). Fewer works deal with verbal MWEs (Wehrli et al., 2010; Vincze et al., 2013; Wehrli, 2014; Waszczuk et al., 2016).

3 Annotation Methodology

In order to bring about substantial progress in the state of the art presented in the preceding section, the European PARSEME network[5], dedicated to parsing and MWEs, proposed a shared task on automatic identification of VMWEs. This initiative required the construction of a large multilingual VMWE-annotated corpus.

Within the challenging features of linguistic annotation, as defined by Mathet et al. (2015), the VMWE annotation task is concerned by:

- *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 (cf. Section 3.1).
- *Sporadicity*, i.e. the fact that not all text tokens are subject to annotation (unlike in POS annotation for instance);
- *Free overlap* (e.g. ***take*** *a* ***walk*** *and then a long* ***shower***: 2 LVCs with a shared light verb);
- *Nesting*, both at the syntactic level (e.g. ***take*** *the fact that I didn't* ***give up into account***) and at the level of lexicalized components (e.g. ***let the cat out of the bag***).

Two other specific challenges are:

- *Discontinuities* (e.g. ***take*** *this* ***into account***);
- *Multiword token* VMWEs, e.g. separable IReflVs or VPCs: (ES) ***abstener|se*** (lit. *abstain self*) 'abstain', (DE) ***auf|machen*** (lit. *out|make*) 'open'.[6]

This complexity is largely increased by the multilingual nature of the task, and calls for efficient project management. The 21 participating languages were divided into four *language groups* (LGs): *Balto-Slavic*: Bulgarian (BG), Croatian (HR), Czech (CS), Lithuanian (LT), Polish (PL) and Slovene (SL); *Germanic*: English (EN), German (DE), Swedish (SV) and Yiddish (YI); *Romance*: French (FR), Italian (IT), Romanian (RO), Spanish (ES) and Brazilian Portuguese (PT); and *others*: Farsi (FA), Greek (EL), Hebrew (HE), Hungarian (HU), Maltese (MT) and Turkish (TR). Note that the 4 last are non-Indo-European. Corpus release was achieved for 18 of these languages, that is, all except HR, EN and YI, for which no sufficiently available native annotators could be found. 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.

3.1 Annotation Guidelines

The biggest challenge in the initial phase of the project was the development of the annotation guidelines[7] which would be as universal 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 and was followed by enhancements of the guidelines, corpus format and processing tools. In this way, the initial guidelines dramatically evolved, new VMWE categories emerged, and the following 3-level typology was defined:

1. *universal* categories, that is, valid for all languages participating in the task:

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

[6] 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. Sections 4.2 and 6). 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 wished VMWE annotations to rely on the same tokenization as the other annotation layers.

[7] Their final version, with examples in most participating languages, is available at http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/.

   (a) light verb constructions (LVCs), e.g. *to* ***give*** *a* ***lecture***
   (b) idioms (ID), e.g. *to* ***call it a day***
2. *quasi-universal* categories, valid for some language groups or languages, but not all:
   (a) inherently reflexive verbs (IReflVs), e.g. (FR) ***s'évanouir*** 'to faint'
   (b) verb-particle constructions (VPCs), e.g. *to* ***do in*** 'to kill'
3. *other* verbal MWEs, not belonging to any of the categories above (due to not having a unique verbal head) e.g. *to* ***drink and drive***, *to* ***voice act***, *to* ***short-circuit***.

While we allowed for language-specific categories, none emerged during the pilot or final annotations. The guidelines consist of linguistic tests and examples, organised into decision trees, aiming at maximising the level of determinism in annotator's decision making. Most of the tests are generic, applying to all languages relevant to a given category, but some are language-specific, such as those distinguishing particles from prepositions and prefixes in DE, EN and HU. Once the guidelines became stable, language leaders added examples for most tests in their languages using a dedicated interface.

3.2 Annotation Tools

For this large-scale corpus construction, we needed a centralized 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[8], a web platform based on FoLiA[9], a rich XML-based format for linguistic annotation (van Gompel and Reynaert, 2013). In addition to the required criteria, it 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 customize specific settings for different languages. Out of 18 language teams, 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.

3.3 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 system (gitlab) in which language leaders could share questions 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, cf. § 4.2). 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 languages (BG, FR, HU, IT, PL, and PT) kept a list of VMWEs and their classification, agreed on by the annotators and updated over time. Secondly, some languages (DE, ES, FR, HE, IT, PL, PT, and RO) performed a step of homogenisation once the annotation was complete. An in-house script read the annotated corpus and generated an HTML page where all positive and negative examples of a given VMWE were grouped. Entries were sorted so that similar VMWEs appear nearby – for instance occurrences of ***pay*** *a* ***visit*** would appear next to occurrences of ***receive*** *a* ***visit***. 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.

4 Corpora

Tables 4 and 5 provide overall statistics of the training and test corpora created for the shared task. We show the number of sentences and tokens in each language, the overall number of annotated VMWEs and the detailed counts per category. In total, the corpora contain 230,062 sentences for training and 44,314 sentences for testing. These correspond to 4,5M and 900K tokens, with 52,724 and 9,494 annotated VMWEs. The amount and distribution of VMWEs over categories varies considerably among languages.

No category was used in all languages but the two universal categories, ID and LVC, were used in almost all languages. In HU, no ID was annotated due to the genre of the corpus, mainly com-

[8] github.com/proycon/flat, flat.science.ru.nl
[9] http://proycon.github.io/folia

posed of legal texts. In FA, no categorisation of the annotated VMWEs was performed, therefore, the OTH category has special semantics there: it does not mean that a VMWE cannot be categorised because of its linguistic characteristics, but rather that the categorisation tests were not applied.

The most frequent category is IReflV, in spite of it being quasi-universal, mainly due to its prevalence in CS. IReflVs were annotated in all Romance and Slavic languages, and in DE and SV. VPCs were annotated in DE, SV, EL, HE, HU, IT, and SL. No language-specific category was defined. However, the high frequency of OTH in some languages is a hint that they might be necessary, especially for non-Indo-European languages like HE, MT and TR.

Table 6 provides statistics about the length and discontinuities of annotated VMWEs in terms of the number of tokens. The average lengths range between 2.1 (PL) and 2.85 (DE) tokens. DE has the greatest dispersion for lengths: the mean absolute deviation (MAD) is 1.44 while it is less than 0.75 for all other languages. DE is also atypical with more than 10% of VMWEs containing one token only (length=1), mainly separable VPCs, e.g. ***auf/machen*** (lit. *out/make*) 'open'. The right part of Table 6 shows the length of discontinuities. The data sets vary greatly across languages. While for BG, FA and IT, more than 80% of VMWEs are continuous, for DE, 30.5% of VMWEs have discontinuities of 4 or more tokens.

All the corpora are freely available. 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, including additional annotations (§ 5). These are released under the terms of the original corpora.

4.1 Format

The official format of the annotated data is the parseme-tsv format[10], exemplified in Figure 1. 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 multiword tokens such as contractions, (ii) the token surface form, (iii) an optional `nsp` 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 (e.g., `2:ID` if a token starts an idiom which is the second VMWE in the current sentence). In case of nested, coordinated or overlapping VMWEs multiple codes are separated with a semicolon.

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

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.

Note that tokenisation is closely related to MWE identification, and it has been shown that performing both tasks jointly may enhance the quality of their results (Nasr et al., 2015). However, the data we provided consist of pre-tokenised sentences. This implies that we expect typical systems to perform tokenisation prior to VMWE identification, and that we do not allow the tokenisation output to be modified with respect to the ground truth. The latter is necessary since the evaluation measures are token-based (§ 6). This approach may disadvantage systems which expect untokenised raw text on input, and apply their own tokenisation methods, whether jointly with VMWE identification or not. We are aware of this bias, and we did encourage such systems to participate in the shared task, provided that they define re-tokenisation methods so as to adapt their outputs to the tokenisation imposed by us.

4.2 Inter-Annotator Agreement

Inter-annotator agreement (IAA) measures are meant to assess the hardness of the annotation task, as well as the quality of its methodology and of the resulting annotations. Defining such measures is not always straightforward due the challenges listed in Section 3.

To assess unitising, we report the per-VMWE

[10] 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 1: Annotation of two sample sentences containing a contraction (*wouldn't*), a verbal idiom, and two coordinated VPCs.

	#S	#T	$\#A_1$	$\#A_2$	F_{unit}	κ_{unit}	κ_{cat}
BG	608	27491	298	261	0.816	0.738	0.925
EL	1383	33964	217	299	0.686	0.632	0.745
ES	524	10059	54	61	0.383	0.319	0.672
FA	200	5076	302	251	0.739	0.479	n/a
FR	1000	24666	220	205	0.819	0.782	0.93
HE	1000	20938	196	206	0.522	0.435	0.587
HU	308	8359	229	248	0.899	0.827	1.0
IT	2000	52639	336	316	0.417	0.331	0.78
PL	1175	19533	336	220	0.529	0.434	0.939
PT	2000	41636	411	448	0.771	0.724	0.964
RO	2500	43728	183	243	0.709	0.685	0.592
TR	6000	107734	3093	3241	0.711	0.578	0.871

Table 1: IAA scores: #S, and #T show the the number of sentences and tokens in the corpora used for measuring the IAA, respectively. $\#A_1$ and $\#A_2$ refer to the number of VMWE instances annotated by each of the annotators.

F-score (F_{unit})[11], as defined in § 6, and 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 can be a VMWE.[12] Consequently, one can argue that the probability of chance agreement approaches 0, and IAA can be measured simply using the observed agreement, the F-score. However, in order to provide a lower bound for 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 can roughly be estimated by the number of sentences: κ_{unit} is the IAA for unitising based on this assumption. To assess categorisation, we apply the standard κ (κ_{cat}) to the VMWEs for which annotators agree on the span.

All available IAA results are presented in Table 1. For some languages the IAA in this unitising is rather low. We believe that this results from particular annotation conditions. In ES, the annotated corpus is small (cf. Table 4) so the annotators gathered relatively few experience with the task. A similar effect occurs in PL and FA, 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 HE, and especially of IT, should be studied more thoroughly in the future. Note also that in some languages the numbers from Table 1 are a lower bound for the quality of the final corpus, due to post-annotation homogenisation (§ 3.3).

A novel proposal of the holistic γ measure (Mathet et al., 2015) combines unitising and categorization 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 know for their variable degree of non-compositionality, i.e. their idiomaticity is a matter of scale. Current corpus annotation standards and identification tools require the MWEhood, conversely, to be a binary property, which sub-optimally models a large number of grey-zone VMWE candidates. However, once the decision of the status of a VMWE candidate, as valid, has been taken, its categorization appears to be significantly simpler, as shown in the last 2 columns of Table 1 (except for Romanian). Secondly, our annotation guidelines are structured in two main decision trees - an identification and a categorization tree - to be applied mostly sequentially.[13] Therefore, separate evaluation of these two stages may be helpful in enhancing the guidelines.

5 Shared Task Organization

Corpora were annotated for VMWEs by different language teams. Before concluding the annotation of the full corpora, we requested language teams to provide a small annotated sample of 200 sentences. These were released as a trial corpus meant

[11] Note that F-score is symmetrical, so none of the two annotators is prioritized.

[12] Also note that annotated segments can overlap.

[13] Identification hypotheses may be questioned in the categorization process in case of LVCs or IReflVs though.

to help participants develop or adapt their systems to the shared task particularities.

The full corpora were split by the organizers into train and test sets. Given the heterogeneous nature and size of the corpora, the splitting method was chosen individually for each language. As a general rule, we tried to create test sets that (a) contained around 500 annotated VMWEs and (b) did not overlap with the released trial data. When the annotated corpus was small (e.g. in SV), we favoured the size of the test data rather than of the training data, so as to lessen the evaluation bias.

For all languages except BG, HE and LT, we also released companion files in a format close to CONLL-U[14]. They contain extra linguistic information which could be used by systems as features. For CS, FA, MT, RO and SL, the companion files contain morphological data only (lemmas, POS-tags and morphological features). For the other languages, they also include syntactic dependencies. Depending on the language, these files were obtained from existing manually annotated corpora and/or treebanks such as UD, or from the output of automatic analysis tools such as UDPipe[15]. A brief description of the companion files is provided in the README of each language.

The test corpus was turned into a blind test corpus by removing all VMWE annotations. After its release, participants had 1 week to provide the predictions output by their systems in the parseme-tsv format. Predicting VMWE categories was not required and evaluation measures did not take them into account (§ 6). Participants did not need to submit results for all languages and it was possible to predict only certain MWE categories.

Each participant could submit results in the two tracks of the shared task: closed and open. The *closed* track aims at evaluating systems more strictly, independently of the resources they have access too. Systems in this track, therefore, learn their VMWE identification models using *only* the VMWE-annotated training corpora and the companion files, when available. Cross-lingual systems which predict VMWE annotations for one language using files provided for other languages were still considered in the closed track. Systems using other knowledge sources such as raw monolingual corpora, lexicons, grammars or language models trained on external resources were considered in the *open* track. This track includes purely symbolic and rule-based systems. Open track systems can use any resource they have access to, as long as it is described in the abstract and/or in the system description paper.

We published participation policies stating that data providers and organizers are allowed to participate in the shared task. Although we acknowledge that this policy is non-standard and introduces biases to system evaluation, we were more interested in cross-language discussions than in a real competition. Moreover, many languages have only a few NLP teams working on them, so adopting an exclusive approach would actually exclude the whole language from participation. Nonetheless, systems were not allowed to be trained on any test corpus (even if authors had access to it in advance) or to use resources (lexicons, MWE lists, etc.) employed or built during the annotation phase.

6 Evaluation Measures

The quality of system predictions is measured with the standard metrics of precision (P), recall (R) and F$_1$-score (F). VMWE categories are not taken into account in system ranking, and we do not require participant systems to predict them.[16]

Token	Gold	System1	System2	System3
t1	1	1	1	1;4
t2	1	2	3	3
t3	2	2	2	2;4

Table 2: Toy gold corpus with 3 tokens, 2 gold VMWEs, and 3 system predictions. VMWE codes do not include VMWE categories.

Each VMWE annotation or prediction can be represented as a set of token identifiers. Consider Table 2, which presents a toy gold corpus containing 2 VMWEs over 3 tokens[17] and 3 system predictions. If G denotes the set of gold VMWEs and Si the set of VMWEs predicted by system i, then the following holds[18]:

[14] http://universaldependencies.org/format.html While we used this format to define the semantics of each column, language teams were free to use their own tagsets and features, not strictly respecting the CONLL-U format definition.

[15] https://ufal.mff.cuni.cz/udpipe

[16] Per-category results are provided, for discussion, for those systems which did predict them, at http://multiword.sourceforge.net/sharedtaskresults2017/.

[17] Recall that a VMWE can contain one (multiword) token only.

[18] Let A be a set of sets. Then $|A|$ is the size of A and $||A||$

- $G = \{\{\text{t1,t2}\}, \{\text{t3}\}\}$, $|G| = 2$, $||G|| = 3$.
- $S1 = \{\{\text{t1}\}, \{\text{t2,t3}\}\}$, $|S1| = 2$, $||S1|| = 3$.
- $S2 = \{\{\text{t1}\}, \{\text{t2}\}, \{\text{t3}\}\}$, $|S2| = 3$, $||S2|| = 3$.
- $S3 = \{\{\text{t1}\}, \{\text{t2}\}, \{\text{t3}\}, \{\text{t1,t3}\}\}$, $|S3|=4$, $||S3||=5$.

A simple way to obtain P, R and F is to consider every VMWE as an indivisible instance, and calculate the ratio of the VMWEs that were correctly predicted (precision) and correctly retrieved (recall). We call this the **per-VMWE** scoring. The per-VMWE scoring for the sample in Table 2 is calculated as follows, with TPi being the number of true positive VMWEs predicted by system i:

- $TP1 = |G \cap S1| = |\varnothing| = 0$
  $R = TP1/|G| = 0/2$
  $P = TP1/|S1| = 0/2$.
- $TP2 = |G \cap S2| = |\{\{\text{t3}\}\}| = 1$
  $R = TP2/|G| = 1/2$.
  $P = TP2/|S2| = 1/3$.
- $TP3 = |G \cap S3| = |\{\{\text{t3}\}\}| = 1$
  $R = TP3/|G| = 1/2$
  $P = TP3/|S3| = 1/4$.

Per-VMWE scores may be too penalising for large VMWEs or VMWEs containing elements whose lexicalisation is uncertain (e.g. definite or indefinite articles: *a*, *the*, etc.). We define, thus, an alternative **per-token** evaluation measure, which allows a VMWE to be partially matched. Such a measure must be applicable to all VMWEs, which is difficult, given the complexity of possible scenarios allowed in the representation of VMWEs, as discussed in Section 3. This complexity hinders the use of evaluation measures found in the literature. For example, Schneider et al. (2014a) use a measure based on pairs of MWE tokens, which is not always possible here given single-token VMWEs. The solution we adopted considers all possible bijections between the VMWEs in the gold and system sets, and takes a matching that maximizes the number of correct token predictions (true positives, denoted below as TPi_{max} for each system i). The application of this metric to the system outcome in Tab. 2 is the following:

- $TP1_{max} = |\{\text{t1,t2}\} \cap \{\text{t1}\}| + |\{\text{t3}\} \cap \{\text{t2,t3}\}| = 2$
  $R = TP1_{max}/||G|| = 2/3$
  $P = TP1_{max}/||S1|| = 2/3$.
- $TP2_{max} = |\{\text{t1,t2}\} \cap \{\text{t1}\}| + |\{\text{t3}\} \cap \{\text{t3}\}| + |\varnothing \cap \{\text{t2}\}| = 2$
  $R = TP2_{max}/||G|| = 2/3$
  $P = TP2_{max}/||S2|| = 2/3$.
- $TP3_{max} = |\{\text{t1,t2}\} \cap \{\text{t1}\}| + |\{\text{t3}\} \cap \{\text{t3}\}| + |\varnothing \cap \{\text{t2}\}| + |\varnothing \cap \{\text{t1,t3}\}| = 2$
  $R = TP3_{max}/||G|| = 2/3$
  $P = TP3_{max}/||S3|| = 2/5$.

Formally, let $G = \{g_1, g_2, \ldots, g_{|G|}\}$ and $S = \{s_1, s_2, \ldots, s_{|S|}\}$ be the ordered sets of gold and system VMWEs in a given sentence, respectively[19]. Let B be the set of all bijections $b : \{1, 2, .., N\} \rightarrow \{1, 2, .., N\}$, where $N = max(|G|, |S|)$. We define $g_i = \varnothing$ for $i > |G|$, and $s_i = \varnothing$ for $i > |S|$.

We denote by TP_{max} the maximum number of true positives for any possible bijection (we calculate over a set of pairs, taking the intersection of each pair and then adding up the number of matched tokens over all intersections):

$$TP_{max} = max_{b \in B} |g_1 \cap s_{b(1)}| + \\ |g_2 \cap s_{b(2)}| + ... + |g_N \cap s_{b(N)}| \qquad (1)$$

The values of TP_{max} are added up for all sentences in the corpus, and precision/recall values are calculated accordingly. Let TP^j_{max}, G^j, S^j and N^j be the values of TP_{max}, G, S and N for the j-th sentence. For a corpus of M sentences, we define:

$$P = \frac{\sum_{j=1}^{M} TP^j_{max}}{\sum_{j=1}^{M} ||S^j||} \quad R = \frac{\sum_{j=1}^{M} TP^j_{max}}{\sum_{j=1}^{M} ||G^j||} \qquad (2)$$

In any of the denominators above is equal to 0 (i.e. either the corpus contains no VMWEs or the system found no VMWE occurrence) the corresponding measure is defined as equal to 0.

Note that these measures operate both on a micro scale (the optimal bijections are looked for within a given sentence) and a macro scale (the results are summed up for all sentences in the corpus). Alternatively, micro-only measures, i.e. the average values of precision and recall for individual sentences, could be considered. Given that the density of VMWEs per sentence can vary greatly, and in many languages the majority of sentences do not contain any VMWE, we believe that the macro measures are more appropriate.

Note also that the measures in (2) are comparable to the CEAF-M measures (Luo, 2005) used in the coreference resolution task.[20] There, mentions are grouped into entities (clusters) and the best bijection between gold and system entities is searched for. The main difference with our approach resides in the fact that, while coreference

the sum of sizes of the elements in A.

[19] We require an ordering so as to be able to define a bijection where some elements do not match anything.

[20] Notable is also the similarity of CEAF with the holistic γ evoked in section 4.2.

is an equivalence relation, i.e. each mention belongs to exactly one entity, VMWEs can exhibit overlapping and nesting. This specificity (as in other related tasks, e.g. named entity recognition) necessarily leads to counter-intuitive results if recall or precision are considered alone. A system which tags all possibles subsets of the tokens of a given sentence as VMWEs will always achieve recall equal to 1, while its precision will be above 0. Note, however, that precision cannot be artificially increased by repeating the same annotations, since the system results (i.e. S and s_i above) are defined as sets.

Potential overlapping and nesting of VMWEs is also the reason of the theoretical exponential complexity of (2) in function of the length of a sentence. In our shared task, the maximum number of VMWEs in a sentence, whether in a gold corpus or in a system prediction (denoted by $N_{max} = max_{j=1,...,M} N^j$), never exceeds 20. The theoretical time complexity of both measures in (2) is $\mathcal{O}({N_{max}}^3 \times M)$.

7 System Results

Seven systems participated in the challenge and submitted a total of 71 runs. One system (LATL) participated in the open track and six in the closed track. Two points of satisfaction are that (i) each one of the 18 languages was covered and (ii) 5 of the 7 systems were multilingual. Systems were ranked based on their per-token and per-VMWE F-scores, within the open and the closed track. Results and rankings are reported, by language groups, in Tables 7–10.

Most systems used techniques originally developed for parsing: LATL employed Fips, a rule-based multilingual parser; the TRANSITION system is a simplified version of a transition-based dependency parsing system; LIF employed a probabilistic transition-based dependency parser and the SZEGED system made use of the POS and dependency modules of the Bohnet parser. The ADAPT and RACAI systems employed sequence labelling with CRFs. Finally, MUMULS exploited neural networks by using the open source library TensorFlow.

In general, scores for precision are much higher than for recall. This can be explained by the fact that most MWEs occur only once or twice in the corpora, which implies that many of the MWEs of the test data were not observed in the training data. As expected, for most systems their per-VMWE scores are (sometimes substantially) lower than their per-token scores. In some cases, however, the opposite happens, which might be due to frequent errors in long VMWEs.

The most popular language of the shared task was FR, as all systems submitted predictions for French MWEs. Based on the numerical results, FA, RO, CS and PL were the easiest languages, i.e. ones for which the best F-scores were obtained. In contrast, somewhat more modest performance resulted for SV, HE, LT and MT, which is clearly a consequence of the lesser amount of training examples for these languages (see Tab. 4). The results for BG, HE, and LT would probably be higher if companion CONLL-U files with morphological/syntactic data could be provided. This would notably allow systems to neutralize inflection, which is particularly rich in verbs in all of these languages, as well as in nouns and adjectives in the first three of them.

FA is an outstanding case (with F-score of the best system exceeding 0.9) and its results are probably correlated with two factors. Firstly, light verbs are explicitly marked as such in the morphological companion files. Secondly, the density of VMWEs is exceptionally high. If we assume, roughly, one verb per sentence, almost each FA verb is the head of a VMWE, and the system prediction boils down to identifying its lexicalized arguments. Further analysis of this phenomenon should notably include data on the most frequent POS-tags and functions of the lexicalized verbal arguments (e.g. how often is it a nominal direct object) and the average length of VMWEs in this language.

Another interesting case is CS, where the size of the annotated data is considerable. This dataset was obtained by adapting annotations from the Prague Dependency Treebank (PDT) to the annotation guidelines and formats of this shared task (Uresová et al., 2016; Bejček et al., 2017). PDT is a long-standing treebank annotation project with advanced modelling and processing facilities. From our perspective it is as a good representative of a high-quality large-scale MWE modelling effort. In a sense, the results obtained for this language can be considered a benchmark for VMWE identification tools.

The relatively high results for RO, CS and PL might relate to the high ratio of IReflVs in these languages. Since the reflexive marker is most often realised by the same form, (CS) *se*, (PL) *się* and (RO) *se* 'self', the task complexity is reduced to identifying its head verb (often adjacent) and establishing the compositionality status of the bigram. Similar effects would be expected, but are not observed, in SL and BG, maybe due to the smaller sizes of the datasets, and to the missing companion file for BG.

Note also the high precision of the leading systems in RO, PL, PT, FR and HU, which might be related to the high proportion of LVCs in these languages, and with the fact that some very frequent light verbs, such as (RO) *da* 'give', (PL) *prowadzić* 'carry on', (PT) *fazer* 'make', (FR) *effectuer* 'perform' and (HU) *hoz* 'bring', connect with a large number of nominal arguments. A similar correlation would be expected, but is not observed, in EL, and especially in TR, where the size of the dataset is substantial. Typological particularities of these languages might be responsible for this missing correlation.

8 Conclusions and Future Work

We have described a highly multilingual collaborative VMWE-dedicated framework meant to unify terminology and annotation methodology, as well as to boost the development of VMWE identification tools. These efforts resulted in (i) the release of openly available VMWE-annotated corpora of over 5 million words, with generally high quality of annotations, in 18 languages, and (ii) a shared task with 7 participating systems. VMWE identification, both manual and automatic, proved a challenging task, and the performance varies greatly among languages and systems.

Future work includes a fine-grained linguistic analysis of the annotated corpora on phenomena such as VMWE length, discontinuities, variability, etc. This should allow us to discover similarities and peculiarities among languages, language families and VMWE types. We also wish to extend the initiative to new languages, so as to confront the annotation methodology with new phenomena and increase its universality. Moreover, we aim at converging with other universal initiatives such as UD. These advances should further boost the development and enhancement of VMWE identification systems and MWE-aware parsers.

Acknowledgments

The work described in this paper has been supported by the IC1207 PARSEME COST action[21], as well as national funded projects: LD-PARSEME[22] (LD14117) in the Czech Republic, PARSEME-FR[23] (ANR-14-CERA-0001) in France, and PASTOVU[24] in Lithuania.

We are grateful to Maarten van Gompel for his intensive and friendly help with adapting the FLAT annotation platform to the needs of our community. Our thanks go also to all language leaders and annotators 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;
- (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;

[21] http://www.parseme.eu
[22] https://ufal.mff.cuni.cz/grants/ld-parseme
[23] http://parsemefr.lif.univ-mrs.fr/
[24] http://mwe.lt/en_US/

- (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.

References

Kübra Adalı, Tutkum Dinç, Memduh Gokirmak, and Gülşen Eryiğit. 2016. Comprehensive Annotation of Multiword Expressions for Turkish. In *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*, pages 60–66, Konya, Turkey, April.

Kfir Bar, Mona Diab, and Abdelati Hawwari. 2014. Arabic Multiword Expressions. In Nachum Dershowitz and Ephraim Nissan, editors, *Language, Culture, Computation. Computational Linguistics and Linguistics: Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday, Part III*, pages 64–81, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eduard Bejček, Jan Hajič, Pavel Straňák, and Zdeňka Urešová. 2017. Extracting Verbal Multiword Data from Rich Treebank Annotation. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT 15)*, pages 13–24. Indiana University, Bloomington, Indiana University, Bloomington.

Stefan Bott, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. 2016. GhoSt-PV: A Representative Gold Standard of German Particle Verbs. In *CogALex-V: Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon at COLING 2016*, pages 125–133.

Marie Candito and Matthieu Constant. 2014. Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 743–753, Baltimore, Maryland, June. Association for Computational Linguistics.

Matthieu Constant and Joakim Nivre. 2016. A Transition-Based System for Joint Lexical and Syntactic Analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 161–171, Berlin, Germany, August. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-Tokens Dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22, Marrakech, Morocco.

Jenny Rose Finkel and Christopher D. Manning. 2009. Joint Parsing and Named Entity Recognition. In *HLT-NAACL*, pages 326–334. The Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, 39(1):195–227.

Heiki-Jaan Kaalep and 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 Contexts*, pages 57–64, Trento, Italy. ACL.

Heiki-Jaan Kaalep and 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 2008)*, pages 23–26, Marrakech, Morocco.

Joseph Le Roux, Antoine Rozenknop, and Matthieu Constant. 2014. Syntactic Parsing and Compound Recognition via Dual Decomposition: Application to French. In *Proceedings of COLING 2014, the 25th International Conference on Computational*

Lang.	ID	LVC	Quasi-universal / OTH
BG	бълвам змии и гущери (lit. *to spew snakes and lizards*) ’to shower abuse’	държа под контрол ’to keep under control’	усмихвам се (IReflV) (lit. *to smile self*) ’to smile’
CS	***házet klacky pod nohy*** (lit. *to throw sticks under feet*) ’to put obstacles in one’s way’	***vyslovovat nesouhlas*** (lit. *to voice disagreement*) ’to disagree’	***chovat se*** (IReflV) (lit. *to keep SELF*) ’to behave’
DE	***schwarz fahren*** (lit. *to drive black*) ’to take a ride without a ticket’	*eine* ***Rede halten*** (lit. *a speech hold*) ’to give a speech’	***sich enthalten*** (IReflV) (lit. *himself contain*) ’to abstain’
EL	χάνω τα αυγά και τα καλάθια (lit. *loose-1SG the eggs and the baskets*) ’to be at a complete and utter loss’	κάνω μία πρόταση (lit. *make-1SG a proposal*) ’to propose’	μπαίνω μέσα (VPC) (lit. *get-1SG in*) ’to go bankrupt’
ES	***hacer de tripas corazón*** (lit. *make of intestines heart*) ’to pluck up the courage’	***hacer*** *una* ***foto*** (lit. *to make a picture*) ’to take a picture’	***coser y cantar*** (OTH) (lit. *to sew and to sing*) ’easy as pie, a piece of cake’
FA	دسته گل به آب دادن (lit. *give flower bouquet to water*) ’to mess up, to do sth. wrong’	امتحان کردن (lit. *to do exam*) ’to test’	به خود آمدن (lit. *to come to self*) ’to gain focus’
FR	***voir le jour*** (lit. *to see the daylight*) ’to be born’	***avoir*** *du* ***courage*** ’to have courage’	***se suicider*** (IReflV) ’to suicide’
HE	אבד עליו כלח ***’avad*** *‘alav* ***kelax*** (lit. *kelax is lost on him*) ’he is outdated’	הגיע למסקנה ***hgi` lmsqnh*** (lit. *to come to a conclusion*) ’to conclude’	לא הבישן למד ***la hbišn lmd*** ’one who is bashful does not learn’
HU	***kinyír*** (lit. *out.cut*) ’to kill’	***szabályozást ad*** (lit. *control-ACC give*) ’to regulate’	***feltüntet*** (VPC) (lit. *up.strike*) ’to mark’
IT	***entrare in vigore*** (lit. *to enter into force*) ’to come into effect’	***fare*** *un* ***discorso*** (lit. *to give a speech*) ’to give a speech’	***buttare giù*** (VPC) (lit. *throw down*) ’to swallow’
LT	***pramušti dugną*** (lit. *to break the-bottom*) ’to collapse’	***turėti veiklų*** (lit. *to have activities*) ’to be busy, to have side jobs’	
MT	għasfur żgħir qalli (lit. *a bird small told me*) ’to hear something from the grapevine’	ħa deċizjoni ’to take a decision’	iqum u joqgħod (OTH) (lit. *jump and stay*) ’to fidget’
PL	***rzucać grochem o ścianę*** (lit. *throw peas against a wall*) ’to try to convince somebody in vain’	***odnieść sukces*** (lit. *to carry-away a success*) ’to be successful’	***bać się*** (IReflV) (lit. *to fear SELF*) ’to be afraid’
PT	***fazer das tripas coração*** (lit. *transform the tripes into heart*) ’to try everything possible’	***fazer*** *uma* ***promessa*** ’to make a promise’	***se queixar*** (IReflV) ’to complain’
RO	*a* ***trage pe sfoară*** (lit. *to pull on rope*) ’to fool’	*a* ***face*** *o* ***vizită*** (lit. *to make a visit*) ’to pay a visit’	*a* ***se gândi*** (IReflV) ’to think’
SL	***spati kot ubit*** (lit. *sleep like dead*) ’to sleep soundly’	***postaviti vprašanje*** (lit. *to put a question*) ’to pose a question’	***bati se*** (IReflV) ’to be afraid’
SV	*att* ***Plocka russinen ur kakan*** (lit. *to pick the raisins out of the cake*) ’to choose only the best things’	***ta*** *ett* ***beslut*** ’to take a decision’	***det knallar och går*** (OTH) (lit. *it trots and walks*) ’it is OK/as usual’
TR	***yüzüstü bırakmak*** (lit. *facedown to leave (sb)*) ’to forsake’	***engel olmak*** (lit. *obstacle to become*) ’to prevent’	***karar vermek*** (OTH) (lit. *decision to give*) ’to decide’

Table 3: Examples of various categories of VMWEs (IDs, LVCs, quasi-universal or other VMWEs) in all 18 languages.

Language	Sentences	Tokens	VMWE	ID	IReflV	LVC	OTH	VPC
BG	6,913	157,647	1,933	417	1,079	435	2	0
CS	43,955	740,530	12,852	1,419	8,851	2,580	2	0
DE	6,261	120,840	2,447	1,005	111	178	10	1,143
EL	5,244	142,322	1,518	515	0	955	16	32
ES	2,502	102,090	748	196	336	214	2	0
FA	2,736	46,530	2,707	0	0	0	2,707	0
FR	17,880	450,221	4,462	1,786	1,313	1,362	1	0
HE	4,673	99,790	1,282	86	0	253	535	408
HU	3,569	87,777	2,999	0	0	584	0	2,415
IT	15,728	387,325	1,954	913	580	395	4	62
LT	12,153	209,636	402	229	0	173	0	0
MT	5,965	141,096	772	261	0	434	77	0
PL	11,578	191,239	3,149	317	1,548	1,284	0	0
PT	19,640	359,345	3,447	820	515	2,110	2	0
RO	45,469	778,674	4,040	524	2,496	1,019	1	0
SL	8,881	183,285	1,787	283	945	186	2	371
SV	200	3,376	56	9	3	13	0	31
TR	16,715	334,880	6,169	2,911	0	2,624	634	0
Total	230,062	4,536,603	52,724	11,691	17,777	14,799	3,995	4,462

Table 4: Overview of the training corpora: number of sentences, tokens, and annotated VMWEs, followed by broken down number of annotations per VMWE category.

Language	Sentences	Tokens	VMWE	ID	IReflV	LVC	OTH	VPC
BG	1,947	42,481	473	100	297	76	0	0
CS	5,476	92,663	1,684	192	1,149	343	0	0
DE	1,239	24,016	500	214	20	40	0	226
EL	3,567	83,943	500	127	0	336	21	16
ES	2,132	57,717	500	166	220	106	8	0
FA	490	8,677	500	0	0	0	500	0
FR	1,667	35,784	500	119	105	271	5	0
HE	2,327	47,571	500	30	0	127	158	185
HU	742	20,398	500	0	0	146	0	354
IT	1,272	40,523	500	250	150	87	2	11
LT	2,710	46,599	100	58	0	42	0	0
MT	4,635	11,1189	500	185	0	259	56	0
PL	2,028	29,695	500	66	265	169	0	0
PT	2,600	54,675	500	90	81	329	0	0
RO	6,031	100,753	500	75	290	135	0	0
SL	2,530	52,579	500	92	253	45	2	108
SV	1,600	26,141	236	51	14	14	2	155
TR	1,321	27,197	501	249	0	199	53	0
Total	44,314	902,601	9,494	2,064	2,844	2,724	807	1,055

Table 5: Overview of the test corpora: number of sentences, tokens, and annotated VMWEs, followed by broken down number of annotations per VMWE category.

	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	2.3	0.46	0	1.35	1.53	6625	51.5	2357	1465	944	1461	(11.4%)
DE	2.85	1.44	715	2.96	2.94	619	35.7	283	159	142	529	(30.5%)
EL	2.46	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	0.7	2243	82.9	202	103	60	99	(3.7%)
FR	2.29	0.44	1	0.65	0.8	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	4.78	13.27	2205	1.01	1.29	506	63.7	178	34	15	61	(7.7%)
IT	2.59	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	2.66	0.69	7	0.34	0.53	589	77.0	123	33	12	8	(1.0%)
PL	2.11	0.2	0	0.53	0.77	2307	73.3	470	195	90	87	(2.8%)
PT	2.24	0.41	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	2.28	0.44	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%)

Table 6: Length in number of tokens of VMWEs and of discontinuities in the training corpora. Columns 1-3: average and mean absolute deviation (MAD) for length, number of VMWEs with length 1 (=1). Columns 4-10: average and MAD for the length of discontinuities, absolute and relative number of continuous VMWEs, number of VMWEs with discontinuities of length 1, 2 and 3. Last 2 columns: absolute and relative number of VMWEs with discontinuities of length >3.

Lang	System	Track	P-MWE	R-MWE	F-MWE	Rank-MWE	P-token	R-token	F-token	Rank-token
DE	SZEGED	closed	0.5154	0.3340	0.4053	2	0.6592	0.3468	0.4545	1
DE	TRANSITION	closed	0.5503	0.3280	0.4110	1	0.5966	0.3133	0.4109	2
DE	ADAPT	closed	0.3308	0.1740	0.2280	3	0.7059	0.2837	0.4048	3
DE	MUMULS	closed	0.3277	0.1560	0.2114	4	0.6988	0.2286	0.3445	4
DE	RACAI	closed	0.3652	0.1300	0.1917	5	0.6716	0.1793	0.2830	5
SV	ADAPT	closed	0.4860	0.2203	0.3032	2	0.5253	0.2249	0.3149	1
SV	SZEGED	closed	0.2482	0.2966	0.2703	3	0.2961	0.3294	0.3119	2
SV	TRANSITION	closed	0.5100	0.2161	0.3036	1	0.5369	0.2150	0.3070	3
SV	RACAI	closed	0.5758	0.1610	0.2517	4	0.6538	0.1677	0.2669	4

Table 7: Results for Germanic languages.

Lang	System	Track	P-MWE	R-MWE	F-MWE	Rank-MWE	P-token	R-token	F-token	Rank-token
BG	TRANSITION	closed	0.6887	0.5518	0.6127	1	0.7898	0.5691	0.6615	1
BG	MUMULS	closed	0.3581	0.3362	0.3468	2	0.7686	0.4809	0.5916	2
CS	TRANSITION	closed	0.7897	0.6560	0.7167	1	0.8246	0.6655	0.7365	1
CS	ADAPT	closed	0.5931	0.5621	0.5772	3	0.8191	0.6561	0.7286	2
CS	RACAI	closed	0.7009	0.5918	0.6418	2	0.8190	0.6228	0.7076	3
CS	MUMULS	closed	0.4413	0.1028	0.1667	4	0.7747	0.1387	0.2352	4
LT	TRANSITION	closed	0.6667	0.1800	0.2835	1	0.6786	0.1557	0.2533	1
LT	MUMULS	closed	0.0000	0.0000	0.0000	n/a	0.0000	0.0000	0.0000	n/a
PL	ADAPT	closed	0.7798	0.6020	0.6795	2	0.8742	0.6228	0.7274	1
PL	TRANSITION	closed	0.7709	0.6260	0.6909	1	0.8000	0.6312	0.7056	2
PL	MUMULS	closed	0.6562	0.5460	0.5961	3	0.8310	0.6013	0.6977	3
PL	SZEGED	closed	0.0000	0.0000	0.0000	n/a	0.0000	0.0000	0.0000	n/a
SL	TRANSITION	closed	0.4343	0.4300	0.4322	1	0.4796	0.4522	0.4655	1
SL	MUMULS	closed	0.3557	0.2760	0.3108	3	0.6142	0.3628	0.4562	2
SL	ADAPT	closed	0.5142	0.2900	0.3708	2	0.7285	0.3262	0.4506	3
SL	RACAI	closed	0.5503	0.2080	0.3019	4	0.7339	0.2145	0.3320	4

Table 8: Results for Balto-Slavic languages.

Lang	System	Track	P-MWE	R-MWE	F-MWE	Rank-MWE	P-token	R-token	F-token	Rank-token
ES	TRANSITION	closed	0.6122	0.5400	0.5739	1	0.6574	0.5252	0.5839	1
ES	ADAPT	closed	0.6105	0.3480	0.4433	2	0.7448	0.3670	0.4917	2
ES	MUMULS	closed	0.3673	0.3100	0.3362	4	0.6252	0.3995	0.4875	3
ES	SZEGED	closed	0.2575	0.5000	0.3399	3	0.3635	0.5629	0.4418	4
ES	RACAI	closed	0.6447	0.1960	0.3006	5	0.7233	0.1967	0.3093	5
FR	ADAPT	closed	0.6147	0.4340	0.5088	2	0.8088	0.4964	0.6152	1
FR	TRANSITION	closed	0.7484	0.4700	0.5774	1	0.7947	0.4856	0.6028	2
FR	RACAI	closed	0.7415	0.3500	0.4755	3	0.7872	0.3673	0.5009	3
FR	SZEGED	closed	0.0639	0.0520	0.0573	6	0.5218	0.2482	0.3364	4
FR	MUMULS	closed	0.1466	0.0680	0.0929	5	0.5089	0.2067	0.2940	5
FR	LIF	closed	0.8056	0.0580	0.1082	4	0.8194	0.0532	0.1000	6
FR	LATL	open	0.4815	0.4680	0.4746	1	0.5865	0.5108	0.5461	1
IT	TRANSITION	closed	0.5354	0.3180	0.3990	1	0.6134	0.3378	0.4357	1
IT	SZEGED	closed	0.1503	0.1560	0.1531	4	0.4054	0.3064	0.3490	2
IT	ADAPT	closed	0.6174	0.1420	0.2309	2	0.6964	0.1532	0.2511	3
IT	RACAI	closed	0.6125	0.0980	0.1690	3	0.6837	0.1053	0.1824	4
PT	TRANSITION	closed	0.7543	0.6080	0.6733	1	0.8005	0.6370	0.7094	1
PT	ADAPT	closed	0.6410	0.5320	0.5814	2	0.8348	0.6054	0.7018	2
PT	MUMULS	closed	0.5358	0.3740	0.4405	3	0.8247	0.4717	0.6001	3
PT	SZEGED	closed	0.0129	0.0080	0.0099	4	0.6837	0.1987	0.3079	4
RO	MUMULS	closed	0.7683	0.7760	0.7721	2	0.8620	0.8112	0.8358	1
RO	ADAPT	closed	0.7548	0.7140	0.7338	4	0.8832	0.7636	0.8190	2
RO	TRANSITION	closed	0.7097	0.8020	0.7531	3	0.7440	0.8449	0.7912	3
RO	RACAI	closed	0.8652	0.7060	0.7775	1	0.8773	0.7019	0.7799	4

Table 9: Results for Romance languages.

Lang	System	Track	P-MWE	R-MWE	F-MWE	Rank-MWE	P-token	R-token	F-token	Rank-token
EL	TRANSITION	closed	0.3612	0.4500	0.4007	1	0.4635	0.4742	0.4688	1
EL	ADAPT	closed	0.3437	0.2880	0.3134	4	0.5380	0.3601	0.4314	2
EL	MUMULS	closed	0.2087	0.2580	0.2308	5	0.4294	0.4143	0.4217	3
EL	SZEGED	closed	0.3084	0.3300	0.3188	2	0.4451	0.3757	0.4075	4
EL	RACAI	closed	0.4286	0.2520	0.3174	3	0.5616	0.2953	0.3871	5
FA	TRANSITION	closed	0.8770	0.8560	0.8664	1	0.9159	0.8885	0.9020	1
FA	ADAPT	closed	0.7976	0.8040	0.8008	2	0.8660	0.8416	0.8536	2
HE	TRANSITION	closed	0.7397	0.2160	0.3344	1	0.7537	0.1975	0.3130	1
HE	MUMULS	closed	0.0000	0.0000	0.0000	n/a	0.0000	0.0000	0.0000	n/a
HU	SZEGED	closed	0.7936	0.6934	0.7401	1	0.8057	0.6317	0.7081	1
HU	MUMULS	closed	0.6291	0.6152	0.6221	5	0.7132	0.6657	0.6886	2
HU	TRANSITION	closed	0.6484	0.7575	0.6987	2	0.6502	0.7012	0.6747	3
HU	ADAPT	closed	0.7570	0.5992	0.6689	3	0.7846	0.5710	0.6610	4
HU	RACAI	closed	0.8029	0.5471	0.6508	4	0.8208	0.5015	0.6226	5
MT	TRANSITION	closed	0.1565	0.1340	0.1444	1	0.1843	0.1460	0.1629	1
MT	ADAPT	closed	0.2043	0.0380	0.0641	2	0.3084	0.0518	0.0887	2
MT	RACAI	closed	0.2333	0.0280	0.0500	3	0.2481	0.0259	0.0469	3
MT	MUMULS	closed	0.0000	0.0000	0.0000	n/a	0.0000	0.0000	0.0000	n/a
TR	TRANSITION	closed	0.6106	0.5070	0.5540	1	0.6123	0.5039	0.5528	1
TR	ADAPT	closed	0.4541	0.4052	0.4283	3	0.5993	0.4728	0.5285	2
TR	RACAI	closed	0.6304	0.4391	0.5176	2	0.6340	0.4348	0.5159	3
TR	MUMULS	closed	0.4557	0.2774	0.3449	4	0.6452	0.3502	0.4540	4

Table 10: Results for other languages.

Linguistics: Technical Papers, pages 1875–1885, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

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

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and 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)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alexis Nasr, Carlos Ramisch, José Deulofeu, and André Valli. 2015. Joint Dependency Parsing and Multiword Expression Tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1116–1126, Beijing, China, July. Association for Computational Linguistics.

Joakim Nivre and Veronika Vincze. 2015. Light Verb Constructions in Universal Dependencies. In *IC1207 COST PARSEME 5th general meeting*, Iaşi, Romania.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Barbu Mitetelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories conference*, Warsaw, Poland, December.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206, April.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive Annotation of Multiword Expressions in a Social Web Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 546–559, San Diego, California, June. Association for Computational Linguistics.

Yuancheng Tu and 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*, pages 31–39, Portland, Oregon, USA, June. Association for Computational Linguistics.

Yuancheng Tu and 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 Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 65–69, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zdenka Uresová, Eduard Bejcek, and Jan Hajic. 2016. Inherently pronominal verbs in czech: Description and conversion based on treebank annotation. In Valia Kordoni, Kostadin Cholakov, Markus Egg, Stella Markantonatou, and Preslav Nakov, editors, *Proceedings of the 12th Workshop on Multiword Expressions, MWE@ACL 2016, Berlin, Germany, August 11, 2016*. The Association for Computer Linguistics.

Maarten van Gompel and 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, 12/2013.

Veronika Vincze and János Csirik. 2010. Hungarian Corpus of Light Verb Constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1110–1118, Beijing, China. Coling 2010 Organizing Committee.

Veronika Vincze, István Nagy T., and 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*, pages 289–295, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Veronika Vincze, János Zsibrita, and István Nagy T. 2013. Dependency Parsing for Identifying Hungarian Light Verb Constructions. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 207–215, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. 2016. Promoting multiword expressions in A* TAG parsing. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 429–439. ACL.

Eric Wehrli, Violeta Seretan, and Luka Nerima. 2010. Sentence Analysis and Collocation Identification. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 27–35, Beijing, China, August. Association for Computational Linguistics.

Eric Wehrli. 2014. The Relevance of Collocations for Parsing. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 26–32, Gothenburg, Sweden, April. Association for Computational Linguistics.