Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions

Carlos Ramisch
Aix Marseille Univ.
France

Silvio Ricardo Cordeiro
Aix Marseille Univ.
France

Agata Savary
Univ. of Tours
France

Veronika Vincze
Univ. of Szeged
Hungary

Verginica Barbu Mititelu
Romanian Academy

Archna Bhatia
Florida IHMC
USA

Maja Buljan
University of Stuttgart,
Germany

Marie Candito
Univ. Paris Diderot
France

Polona Gantar
Faculty of Arts
Slovenia

Voula Giouli
Athena Research Center
Greece

Tunga Güngör
Bogazici Univ.
Turkey

Abdelati Hawwari
Gorge Washington
Univ., USA

Uxoa Iñurrieta
University of the
Basque Country

Jolanta Kovalevskaitė
Vytautas Magnus Univ.
Lithuania

Simon Krek
Jožef Stefan Institute
Slovenia

Timm Lichte
University of Düsseldorf
Germany

Chaya Liebeskind
Jerusalem College
of Technology, Israel

Johanna Monti
"L'Orientale" Univ.
of Naples, Italy

Carla Parra Escartín
Dublin City University
Ireland

Behrang QasemiZadeh
University of Düsseldorf
Germany

Renata Ramisch
Interinstitutional Center
for Computational Linguistic, Brazil

Nathan Schneider
Georgetown University
USA

Ivelina Stoyanova
Bulgarian Academy of Sciences

Ashwini Vaidya
IIT Delhi, India

Abigail Walsh
Dublin City University, Ireland

Abstract

This paper describes the PARSEME Shared Task 1.1 on automatic identification of verbal multiword expressions. We present the annotation methodology, focusing on changes from last year's shared task. Novel aspects include enhanced annotation guidelines, additional annotated data for most languages, corpora for some new languages, and new evaluation settings. Corpora were created for 20 languages, which are also briefly discussed. We report organizational principles behind the shared task and the evaluation metrics employed for ranking. The 17 participating systems, their methods and obtained results are also presented and analysed.

1 Introduction

Across languages, multiword expressions (MWEs) are widely recognized as a significant challenge for natural language processing (NLP) (Sag et al., 2002; Baldwin and Kim, 2010). An international and highly multilingual research community, forged via regular workshops and initiatives such as the PARSEME network (Savary et al., 2015), has rallied around the goals of characterizing MWEs in lexicons, grammars and corpora and enabling systems to process them. Recent shared tasks, namely DiMSUM (Schneider et al., 2016) and the first edition of the PARSEME Shared Task on automatic identification of verbal multiword expressions in 2017 (Savary et al., 2017), have helped drive MWE research forward, yielding new corpora and testbeds for MWEs identification systems.

This paper describes edition 1.1 of the PARSEME Shared Task, which builds on this momentum. We amalgamated organizational experience from last year's task, a more polished version of the annotation methodology and an extended set of linguistic data, yielding an event that attracted 12 teams from 9 countries. Novel aspects in this year's task include additional annotated data for most of the languages, some new languages with annotated datasets and enhanced annotation guidelines.

The structure of the paper is the following. First, related work is presented, then details on the annotation methodology are described, focusing on changes from last year's shared task. We have annotated corpora for 20 languages, which are briefly discussed. Main organizational principles behind the shared task, as well as the evaluation metrics are reported next. Finally, participating systems are introduced and their results are discussed before we draw our conclusions.

2 Related Work

In the last few years, there have been several evaluation campaigns for MWE identification. First, the 2008 MWE workshop contained an MWE-targeted shared task. However, the goal of participants was to rank the provided MWE candidates instead of identifying them in raw texts. The recent DiMSUM 2016 shared task (Schneider et al., 2016) challenged participants to label English sentences in tweets, user reviews of services, and TED talks both with MWEs and supersenses for nouns and verbs. Last, the 1.0 edition of the PARSEME Shared Task in 2017 (Savary et al., 2017) provided annotated datasets for 18 languages, where the goal was to identify verbal MWEs in context. Our current shared task is similar in vein to the previous edition. However, the annotation methodology has been enhanced (see Section 3) and the set of languages covered has also been changed.

Rosén et al. (2015) reports on a survey of MWE annotation in 17 treebanks for 15 languages, collaboratively documented according to common guidelines. They highlight the heterogeneity of MWE annotation practices. Similar conclusions have been drawn for Universal Dependencies (McDonald et al., 2013). With regard to these conclusions, we intended to provide unified guidelines for all the participating languages, in order to avoid heterogeneous, hence incomparable, datasets.

MWE identification in syntactic parsing has also gained some popularity in recent years. While often treated as a pre-processing step for parsing, both tasks are now more and more integrated (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). Although fewer works deal with verbal MWEs, there are some notable exceptions (Wehrli et al., 2010; Vincze et al., 2013; Wehrli, 2014; Waszczuk et al., 2016). Some systems that participated in edition 1.0 of the PARSEME Shared Task are also based on parsing (Al Saied et al., 2017; Nerima et al., 2017; Simkó et al., 2017). Other approaches to MWE identification include sequence labeling using CRFs (Boroş et al., 2017; Maldonado et al., 2017) and neural networks (Klyueva et al., 2017).

3 Enhanced Annotation Methodology

The first PARSEME annotation campaign (Savary et al., forthcoming) generated a rich feedback from annotators and language team leaders. It also attracted the interest of new teams, working on languages not covered by the previous version of the PARSEME corpora. About 80 issues were raised and discussed among dozens of contributors.[1] This boosted our efforts towards a better understanding of VMWE-related phenomena, and towards a better synergy of terminologies across languages and linguistic traditions. The annotation guidelines were gradually enhanced, so as to achieve more clear-cut distinctions among categories, and make the decision process easier and more reliable. As a result, we expected higher-quality annotated corpora and better VMWE identification systems learned on them.

3.1 Definitions

We maintain all major definitions (unified across languages) introduced in edition 1.0 of the annotation campaign (Savary et al., forthcoming, Sec. 2). In particular, we understand *multiword expressions* as expressions with at least two *lexicalized components* (i.e. always realised by the same lexemes), including a head word and at least one other syntactically related word. Thus, lexicalized components of MWEs must form a connected dependency graph. Such expressions must display some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy, formalised by the annotation procedures.

[1] The issues can be found at Gitlab: `https://gitlab.com/parseme/sharedtask-guidelines/issues`

As previously, syntactic variants of MWE candidates are normalised to their least marked form (called the *canonical form*) maintaining the idiomatic reading, before it is submitted to linguistic tests. A *verbal MWE* is defined as a MWE whose head in a canonical form is a verb, and which functions as a verbal phrase, unlike e.g. FR *peut-être* 'may-be'⇒'maybe' (which is always an adverbial). As in edition 1.0, we account for single-token VMWEs with multiword variants, e.g. ES ***hacerse*** 'make-self'⇒'become' vs. ***se hace*** 'self makes'⇒'becomes'.

3.2 Typology

Major changes in the annotation guidelines between edition 1.0 and 1.1 include redesigning the VMWE typology, which is now defined as follows:[2]

1. Two *universal* categories, that is, valid for all languages participating in the task:
   (a) LIGHT VERB CONSTRUCTIONS (LVC), divided into two subcategories:
      i. LVCs in which the verb is semantically totally bleached (LVC.full), DE *eine* ***Rede halten*** 'hold a speech'⇒'give a speech',
      ii. LVCs in which the verb adds a causative meaning to the noun (LVC.cause),[3] e.g. PL ***narazić na straty*** 'expose to losses'
   (b) VERBAL IDIOMS (VID),[4] grouping all VMWEs not belonging to other categories, and most often having a relatively high degree of semantic non-compositionality, e.g. LT *našta* ***gula ant*** *savivaldybių* ***pečių*** 'the burden lies on the shoulders of the municipality'⇒'the municipality is in charge of the burden'
2. Three *quasi-universal* categories, valid for some language groups or languages, but not all:
   (a) INHERENTLY REFLEXIVE VERBS (IRV)[5] – pervasive in Romance and Slavic languages, and present in Hungarian and German – in which the reflexive clitic (REFL) either always co-occurs with a given verb, or markedly changes its meaning or subcategorisation frame, e.g. PT ***se formar*** 'REFL form'⇒'graduate'
   (b) VERB-PARTICLE CONSTRUCTIONS (VPC) – pervasive in Germanic languages and Hungarian, rare in Romance and absent in Slavic languages – with two subcategories:
      i. fully non-compositional VPCs (VPC.full),[6] in which the particle totally changes the meaning of the verb, e.g. HU ***berúg*** 'in-kick'⇒'get drunk'
      ii. semi non-compositional VPCs (VPC.semi),[7] in which the particle adds a partly predictable but non-spatial meaning to the verb, e.g. EN ***wake up***
   (c) MULTI-VERB CONSTRUCTIONS (MVC)[8] – close to semantically non-compositional serial verbs in Asian languages like Chinese, Hindi, Indonesian and Japanese (but also attested in Spanish), e.g. HI ***kar le*** 'do take'⇒'do (for one's own benefit)', ***kar de*** 'do give'⇒'do (for other's benefit)'
3. One *language-specific* category, introduced for Italian:
   (a) INHERENTLY CLITIC VERBS (LS.ICV),[9] in which at least one non-reflexive clitic (CLI) either always accompanies a given verb or markedly changes its meaning or its subcategorisation frame, e.g. IT ***prenderle*** 'take-them'⇒'get beaten up'

[2] In-line examples contain a two-letter language code, a literal translation into English, and an idiomatic translation. The lexicalized components are highlighted in bold.

[3] This subcategory is new in edition 1.1. It absorbs some verb-noun combinations previously annotated as IDs, but also includes many previously non-annotated ones.

[4] This category largely overlaps with IDs introduced in edition 1.0. Major changes include: (i) shifting some verb+noun combinations into the LVC.cause category, (ii) absorbing the previously used OTH category (covering verbs not having a single verbal head) due to its very restricted use.

[5] In edition 1.0 the acronym IReflV was used for this category. It was changed to IRV for easier pronunciation.

[6] This subcategory corresponds to the VPC category from edition 1.0.

[7] This subcategory is new in edition 1.1.

[8] This subcategory is new in edition 1.1. It absorbs some rare cases of previously annotated verb-verb combinations like FR ***laisser tomber*** 'let fall'⇒'abandon'.

[9] This subcategory is new in edition 1.1. It absorbs some cases of previously annotated IDs in Italian.

4. One *optional experimental* category, to be considered in the post-annotation step:
   (a) INHERENTLY ADPOSITIONAL VERBS (IAV) - they include idiomatic combinations of verbs with prepositions or post-positions, depending on the language, e.g. HR *ne* ***dođe do*** *usporavanja* 'it will not come to delay'⇒'no delay will occur'[10]

3.3 Decision tree for annotation

Edition 1.0 featured a two-stage annotation process, according to which VMWEs were supposed to be first identified in a category-neutral fashion, then classified into one of the VMWE categories. Since the annotation practice showed that VMWE identification is virtually always done in a category-specific way, for this year's task we constructed a unified decision tree, shown in Fig. 1.[11] Note that the first 4 tests are structural. They first hypothesize as VIDs those candidates which: (S.1) do not have a unique verb as head, e.g. HE *britanya* ***nas'a ve-natna*** *'im micrayim* 'Britain carried and gave with Egypt'⇒'Britain negotiated with Egypt', (S.2) have more than one lexicalized dependent of the head verb, EL ρίχνω λάδι στη φωτιά 'pour oil to-the fire'⇒'make a bad or negative situation feel worse', (S.3) have a lexicalized subject, e.g. EU ***deabruak eraman*** 'devil-the.ERG[12] take'⇒'be taken by the devil, go to hell'. The remaining candidates, i.e. those having exactly one head verb and one lexicalized non-subject dependent, trigger category specific tests depending on the part-of-speech of this dependent (S.4).

- ↳ Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
  - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
    - ↳ **NO ⇒** It is not a VMWE, **exit**
  - ↳ **YES ⇒** Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
    - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
      - ↳ **NO ⇒** It is not a VMWE, **exit**
    - ↳ **YES ⇒** Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
      - ↳ **YES ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
        - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
        - ↳ **NO ⇒** It is not a VMWE, **exit**
      - ↳ **NO ⇒** Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
        - ↳ **Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **IRV**
          - ↳ **NO ⇒** It is not a VMWE, **exit**
        - ↳ **Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **VPC.full** or **VPC.semi**
          - ↳ **NO ⇒** It is not a VMWE, **exit**
        - ↳ **Verb with no lexicalized dependent** ⇒ Apply MVC-specific tests ⇒ *MVC tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **MVC**
          - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
            - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
            - ↳ **NO ⇒** It is not a VMWE, **exit**
        - ↳ **Extended NP** ⇒ Apply LVC-specific decision tree ⇒ *LVC tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **LVC**
          - ↳ **NO ⇒** Apply the VID-specific tests ⇒ *VID tests positive?*
            - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
            - ↳ **NO ⇒** It is not a VMWE, **exit**
        - ↳ **Another category** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
          - ↳ **YES ⇒** Annotate as a VMWE of category **VID**
          - ↳ **NO ⇒** It is not a VMWE, **exit**

Figure 1: Decision tree for joint VMWE identification and classification.

[10] This category is considered experimental since, so far, we did not manage to come up with satisfactory tests clearly distinguishing such cases from regular verbal valency.
[11] For Italian and Hindi, this tree is slightly modified to account for: (i) the Italian-specific LS.ICV category, (ii) Hindi MVCs in which an adjective is morphologically identical to an eventive noun.
[12] ERG: ergative case, which is generally attached to the subject of transitive verbs in Basque.

3.4 Consistency checks

Due to manpower constraints, we could not perform double annotation followed by adjudication. For most languages, only small fractions of the corresponding corpus were double-annotated (Sec. 4.2). Therefore, in order to increase the consistency of the annotations, we applied the consistency checking tool developed for edition 1.0 (Savary et al., forthcoming, Sec. 5.4). The tool provides an "orthogonal" view of the corpus, where all annotations of the same VMWE are grouped and can be corrected interactively. Previous experience showed that the use of this tool greatly reduced noise and silence errors. This year, almost all language teams completed the consistency check phase (with the exception of Arabic).

4 Corpora

For edition 1.1, we prepared annotated corpora for 20 languages divided into four groups:

- Germanic languages: German (DE), English (EN)
- Romance languages: Spanish (ES), French (FR), Italian (IT), Portuguese (PT), Romanian (RO)
- Balto-Slavic languages: Bulgarian (BG), Croatian (HR), Lithuanian (LT), Polish (PL), Slovene (SL)
- Other languages: Arabic (AR), Greek (EL), Basque (EU), Farsi (FA), Hebrew (HE), Hindi (HI), Hungarian (HU), Turkish (TR)

Arabic, Basque, Croatian, English and Hindi were additional languages, compared to the first edition of the shared task. However, the Czech, Maltese and Swedish corpora were not updated and hence were not included in edition 1.1 of the shared task. The Basque corpus comprises texts from the whole UD corpus (Aranzabe et al., 2015) and part of the Elhuyar Web Corpora.[13] The Bulgarian corpus comprises news articles from the Bulgarian National Corpus (Koeva et al., 2012). The Croatian corpus contains sentences from the Croatian version of the *SETimes* corpora: mostly running text but also selected fragments, such as introductory blurbs and image descriptions characteristic of newswire text. The English corpus consists of 7,437 sentences taken from three of the UD: the Gold Standard Universal Dependencies Corpus for English, the LinES parallel corpus and the Parallel Universal Dependencies treebank. The Farsi corpus is built on top of the MULTEXT-East corpora (QasemiZadeh and Rahimi, 2006) and VMWE annotations are added to a portion of Orwell's 1984 novel. The French corpus contains the Sequoia corpus (Candito and Seddah, 2012) converted to UD, the GDS French UD treebank, the French part of the Partut corpus, and part of the Parallel UD (PUD) corpus. The German corpus contains shuffled sentences crawled from online news, reviews and wikis, derived from the WMT16 shared task data (Bojar et al., 2016), and Universal Dependencies v2.0. The Greek corpus comprises Wikipedia articles and newswire texts from various on-line newspaper editions and news portals. The Hebrew corpus contains news and articles from *Arutz 7* and *HaAretz* news websites, collected by the MILA Knowledge Center for Processing Hebrew. The Hindi corpus represents the news genre sentences selected from the test section of the Hindi Treebank (Bhat et al., 2015). The Hungarian corpus contains legal texts from the Szeged Treebank (Csendes et al., 2005). The Italian corpus is a selection of texts from the PAISÁ corpus of web texts (Lyding et al., 2014), including Wikibooks, Wikinews, Wikiversity, and blog services. The Lithuanian corpus contains articles from a Lithuanian news portal DELFI. The Polish corpus builds on top of the National Corpus of Polish (Przepiórkowski et al., 2011) and the Polish Coreference Corpus (Ogrodniczuk et al., 2015). These are balanced corpora, from which we selected mainly daily and periodical press extracts. The Portuguese corpus contains sentences from the informal Brazilian newspaper *Diário Gaúcho* and from the training set of the *UD_Portuguese-GSD* v2.1 treebank. The Romanian corpus is a collection of articles from the concatenated editions of the *Agenda* newspaper. The Slovenian corpus contains parts of the ssj500k 2.0 training corpus (Krek et al., 2017), which consists of sampled paragraphs from the Slovenian reference FidaPLUS corpus (Arhar Holdt et al., 2007), including literary novels, daily newspapers, web blogs and social media. The Spanish corpus consists of newspaper texts from the the Ancora corpus (Taulé et al., 2016), the UD version of Ancora, a corpus compiled by the IXA group in the University of the Basque country, and parts of the training set of the UD v2.0 treebank. The Turkish corpus consists of 18,611 sentences of newswire texts in several genres.

[13] http://webcorpusak.elhuyar.eus/

As shown in Table 2, most languages provided corpora containing several thousand VMWEs, totalling 79,326 VMWEs across all languages. The smallest corpus is in English, containing around 7,437 sentences and 832 VMWEs, and the largest one is in Hungarian, with 7,760 VMWEs. All corpora, except the Arabic one, are available under different flavours of the Creative Common license.[14]

4.1 Format

Edition 1.1 of the shared task saw a major evolution of the data format, motivated by a quest for synergies between PARSEME (Savary et al., forthcoming) and Universal Dependencies (Nivre et al., 2016), two complementary multilingual initiatives aiming at unified terminologies and methodologies. The new format called `cupt`, combines in one file the `conllu` format[15] and the `parsemetsv` format[16], both used in the previous edition of this shared task.

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
# source_sent_id = . . corola-35693
# text = Lidia se stingea pe picioare.
1  Lidia     Lidia   NOUN   Ncfsry      Case=Acc|Definite=Def...       3  nsubj    _  _              *
2  se        sine    PRON   Px3–a———w   Case=Acc|Person=3|...          3  expl:pv  _  _              1:IRV;2:VID
3  stingea   stinge  VERB   Vmii3s      Mood=Ind|Number=Sing|...       0  root     _  _              1;2
4  pe        pe      ADP    Spsa        AdpType=Prep|Case=Acc          5  case     _  _              2
5  picioare  picior  NOUN   Ncfp-n      Definite=Ind|Gender=Fem|...    3  obl      _  SpaceAfter=No  2
6  .         .       PUNCT  PERIOD      _                              3  punct    _  _              *
```

Figure 2: First sentence of a corpus, with a nested VMWE, in the `cupt` format: RO *Lidia* ***se stingea pe picioare*** 'Lidia Refl.Cl.3.Sg.Acc. was_extinguishing on legs'⇒'Lidia was going into decline'.

As seen in Fig. 2, each token in a sentence is now represented by 11 columns: the 10 columns compatible with the `conllu` specification (notably: rank, token, lemma, part-of-speech, morphological features, and syntactic dependencies), and the 11th column containing the VMWE annotations, according to the same conventions as `parsemetsv` but with the updated set of categories (cf. Sec. 3.2). Note the presence of an IRV (tokens 2–3) embedded in a VID (tokens 2–5). The underscore '_', when it occurs alone in a field, is reserved for underspecified annotations. It can be used in incomplete annotations or in blind versions of the annotated files. The star '*', when it occurs alone in a field, is reserved for empty annotations, which are different from underspecified. This concerns sporadic annotations, typical for VMWEs (where not necessarily all words receive an annotation, as opposed to e.g. part-of-speech tags).

Besides adding a new column to `conllu`, `cupt` also introduces additional conventions concerning comments (lines starting with '#'). The first line of each file must indicate the ordered list of columns (with standardized names) that this file contains, i.e. the same format can be used for any subset of standard columns, in any order. Each sentence is then preceded by the identifier of the source sentence (`source_sent_id`) which consists of three fields: (i) the persistent URI of the original corpus (e.g. of a UD treebank), (ii) the path of the source file in the original corpus, (iii) the sentence identifier, unique within the whole corpus. Items (i) and (ii) contain '.' if there is no external source corpus, as in the example of Figure 2. The following comment line contains the text of the current sentence. Validation scripts and converters were developed for `cupt`, and published before the shared task.

4.2 Inter-Annotator Agreement

Contrary to standard practice in corpus annotation, most corpora were not double-annotated due to lack of human resources. Nonetheless, each language team has double-annotated a sample containing at least 100 annotated VMWEs.[17] The number of sentences (S), number of VMWEs annotated by the first (A_1) and by the second annotator (A_2) are shown in Table 1. The last three columns report two measures to assess span agreement (tokens belonging to a VMWE) and one measure to assess the agreement on

[14] At https://gitlab.com/parseme/sharedtask-data/tree/master/1.1.
[15] http://universaldependencies.org/format.html
[16] https://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation
[17] The Lithuanian team double-annotated a sample from the Lithuanian Treebank treebank.

	S	A_1	A_2	F_{span}	κ_{span}	κ_{cat}		S	A_1	A_2	F_{span}	κ_{span}	κ_{cat}
AR	200	205	207	0.961	0.923	1.000	**HI**	300	188	162	0.634	0.553	0.766
BG	1237	472	459	0.917	0.899	0.957	**HR**	272	270	204	0.515	0.359	0.792
DE	696	305	265	0.673	0.601	0.604	**HU**	308	274	329	0.892	0.831	1.000
EL	1617	428	462	0.694	0.665	0.673	**IT**	1000	341	379	0.586	0.550	0.882
EN	804	153	176	0.529	0.487	0.625	**LT**	2343	157	103	0.469	0.460	0.788
ES	1508	197	103	0.253	0.227	0.573	**PL**	2079	759	707	0.619	0.568	0.882
EU	871	327	355	0.859	0.820	0.859	**PT**	1000	275	241	0.713	0.684	0.837
FA	402	416	336	0.606	0.470	1.000	**RO**	2503	529	556	0.533	0.491	0.823
FR	803	329	363	0.766	0.729	0.960	**SL**	800	214	220	0.811	0.795	0.982
HE	1800	290	291	0.806	0.794	0.932	**TR**	187	154	150	0.987	0.984	0.955

Table 1: Per-language inter-annotator agreement on a sample of S sentences, with A_1 and A_2 VMWEs annotated by each annotator. F_{span} is the F-measure between annotators, κ_{span} is the agreement on the annotation span and κ_{cat} is the agreement on the VMWE category. EL, EN and HI provided corpora annotated by more than 2 annotators. We report the highest scores among all possible annotator pairs.

VMWE categories (Sec. 3.2). The F_{span} score is the MWE-based F-measure when considering that one of the annotators tries to predict the other one's annotations.[18] This is identical to the F1-MWE score used to evaluate participating systems (Sec. 6). F_{span} is an optimistic estimator which ignores chance agreement. On the other hand, κ_{span} and κ_{cat} estimate to what extent the observed agreement P_O exceeds the expected agreement P_E, that is, $\kappa = \frac{P_O - P_E}{1 - P_E}$.

Observed and expected agreement for κ_{span} are based on the number of verbs V in the sample, assuming that a simplification of the task consists of deciding whether each verb belongs to a VMWE or not.[19] If annotators perfectly agree on $A_{1=2}$ annotated VMWEs, then we estimate that they agree on $N = V - A_1 - A_2 + A_{1=2}$ verbs not belonging to a VMWE, so $P_O = \frac{A_{1=2}+N}{V}$ and $P_E = \frac{A_1}{V} \times \frac{A_2}{V}$. As for κ_{cat}, we consider only the $A_{1=2}$ VMWEs on which both annotators agree on the span, and calculate P_O and P_E based on the proportion of times both annotators agree on the VMWE's category label.

Inter-annotator agreement scores can give an idea of the quality of the guidelines and of the training procedures for annotators. We observe a high variability among languages, especially for determining the span of VMWEs, with κ_{span} ranging from 0.227 for Spanish to 0.984 for Turkish. Macro-averaged κ_{span} is 0.691, which is superior to the macro-averaged κ_{unit} reported in 2017, which was of 0.58 (Savary et al., 2017).[20] Categorization agreement results are much more homogeneous, with a macro-average κ_{cat} of 0.836, which is also slightly higher than the one obtained in 2017, which was of 0.819.

The variable agreement values observed could be explained by language and corpus characteristics (e.g. web texts are harder to annotate than newspapers). They could also be explained by the fact that the double-annotated samples are quite small. Finally, they could indicate that the guidelines are still vague and that annotators do not always receive appropriate training. In reality, probably a mixture of all these factors explains the low agreement observed for some languages. In short, Table 1 strongly suggests that there is still room for improvement in (a) guidelines, (b) annotator training, and (c) annotation team management, best practices, and methodology. It should also be noted that lower agreement values may correlate with the results obtained by participants: the lower the IAA for a given language (i.e. the more difficult the task is for humans), the lower the results of automatic MWE identification. Nevertheless, we believe that the systematic use of our in-house consistency checks tool helped homogenizing some of these annotation disagreements (Sec. 3.4).

[18]Every annotator annotated at least one VMWE, as attested by A_1 and A_2.

[19]When no POS information was available (i.e. for AR), we approximated V as the number of sentences S, i.e. $V \approx S$.

[20]Notice that in 2017, the $V \approx S$ approximation was used for all languages, so both scores are not directly comparable.

5 Shared Task Organization

Each language in the shared task was handled by a team that was responsible for the choice of sub-corpora and for the annotation of VMWEs, in a similar setting as in the previous edition. For each language, we then split its corpus into training, test and development sets (train/test/dev), as follows:

- If the corpus has less than 550 VMWEs: Take sentences containing 90% of the VMWEs as test, and the other 10% as a small training corpus.
- If the corpus has between 550 and 1500 VMWEs: Take sentences containing 500 VMWEs as test, and take the rest for training.
- If the corpus has between 1,500 and 5,000 VMWEs: Take sentences containing 500 VMWEs as test, take sentences containing 500 VMWEs as dev, and take the rest for training.
- If the corpus has more than 5,000 VMWEs: Take sentences containing 10% of the VMWEs as test, take sentences containing 10% of the VMWEs as dev, and take the remaining 80% for training.

As in edition 1.0, participants could submit their systems to two tracks: open and closed. Systems in the closed track were only allowed to train their models on the train and dev files provided.

In this edition, we distinguished sentences based on their origin, so as to make sure that the fraction of each sub-corpus is the same in all splits for each language. For example, around 59% of all Basque sentences came from UD, while the other 41% came from the sub-corpus Elhuyar. We have made sure that similar percentages also applied to test/train/dev when taken in isolation. Due to this balancing act, for most languages, we could not keep the VMWEs in the same split as in edition 1.0.

6 Evaluation Measures

The goal of the evaluation measures is to represent the quality of system predictions when compared to the human-annotated gold standard for a given language. As in edition 1.0, we define two types of evaluation measures: a strict *per-VMWE* score (in which each VMWE in gold is either deemed predicted or not, in a binary fashion); and a fuzzy *per-token* score (which takes partial matches into account). For each of these two, we can calculate precision (P), recall (R) and F_1-scores (F).

Orthogonally to the type of measure, there is the choice of what subset of VMWEs to take into account from gold and system predictions. As in the previous edition, we calculate a general category-agnostic measure (both per-VMWE and per-token) based on the totality of VMWEs in both gold and system predictions — this measure only considers whether each VMWE has been properly predicted, regardless of category. We also calculate category-specific measures (both per-VMWE and per-token), where we consider only the subset of VMWEs associated with a given category.

We additionally consider the following phenomenon-specific measures, which focus on some of the challenging phenomena specifically relevant to MWEs (Constant et al., 2017):

- *MWE continuity*: We calculate per-VMWE scores for two different subsets: continuous e.g. TR ***istifa edecek*** 'resignation will-do'⇒'he/she will resign', and discontinuous VMWEs e.g. SL ***imajo*** *investicijske* ***načrte*** 'they-have investment plans'⇒'they have investment plans'.
- *MWE length*: We calculate per-VMWE scores for two different subsets: single-token, e.g. DE ***anfangen*** 'at-catch'⇒'begin', ES ***abstenerse*** 'abstain-REFL'⇒'abstain', and multi-token VMWEs e.g. FA **چشم انداختن** 'eye throw'⇒'to look at'.
- *MWE novelty*: We calculate per-VMWE scores for two subsets: seen and unseen VMWEs. We consider a VMWE in the (gold or prediction) test corpus as seen if a VMWE with the same multiset of lemmas is annotated at least once in the training corpus. Other VMWEs are deemed unseen. For instance, given the occurrence of EN ***has*** *a new* ***look*** in the training corpus, the occurrence of EN ***had*** *a* ***look*** *of innocence* and of EN ***having a look*** *at this report* in the test corpus would be considered seen and unseen, respectively.
- *MWE variability*: We calculate per-VMWE scores for the subset of VMWEs that are variants of VMWEs from the training corpus. A VMWE is considered a variant if: (i) it is deemed as a seen VMWE, as defined above, and (2) it is not identical to another VMWE, i.e. the training corpus does not contain the sequence of surface-form tokens as seen in this VMWE (including non-lexicalized

components in between, in the case of discontinuous VMWEs). E.g., BG накриво ли беше стъпил is a variant of стъпя накриво 'to step to the side'⇒'to lose (one's) footing'.

Systems may predict VMWEs for all languages in the shared task, and the aforementioned measures are independently calculated for each language. Additionally, we calculate a macro-average score based on all of the predictions. In this case, the precision P for a given measure (e.g. for continuous VMWEs) is the average of the precisions for all 19 languages. Arabic is not considered due to delays in the corpus release. Missing system predictions are assumed to have P = R = 0. The recall R is averaged in the same manner, and the average F score is calculated from these averaged P and R scores.

7 System Results

For the 2018 edition of the PARSEME Shared Task, 12 teams submitted 17 system results: 13 to the closed track and 4 to the open track. No team submitted system results for all 20 languages of the shared task, but 11 teams covered 19 languages (all except Arabic). Detailed result tables are reported on the shared task website.[21] In the tables, systems are referred to by anonymous nicknames. System authors and their affiliations are available in the system description papers published in these proceeings.

Most of the systems (Deep-BGT, GBD-NER-standard, GBD-NER-resplit, mumpitz, mumpitz-preinit, SHOMA, TRAPACC, TRAPACC-S and Veyn) exploited neural networks. Syntactic trees and parsing methods were employed in other systems (Milos, MWETreeC and TRAVERSAL) while CRF-DepTree-categ and CRF-Seq-noncateg are based on a tree-structured CRF. Polirem-basic and Polirem-rich use statistical methods and association measures whereas varIDE relies on a Naive Bayes classifier.

As for the best performing systems, TRAPACC and TRAVERSAL were ranked first for 8 languages and 7 language, respectively. TRAVERSAL is more effective in Slavic and Romance languages, whereas TRAPACC works well for German and English. In the "Other" language group, GDB-NER achieved the best results for Farsi and Turkish, and CRF approaches proved to be the best for Hindi. The best results for Bulgarian were obtained by varIDE, based on a Naive Bayes classifier.

Results per language show that, Hungarian and Romanian were the "easiest" languages for the systems, with best MWE-based F-scores of 90.31 and 85.28, respectively. Hebrew, English and Lithuanian show the lowest MWE-based F-scores, not exceeding 23.28, 32.88 and 32.17, respectively. This is likely due to the amount of annotated training data: Hungarian had the highest, whikle English and Lithuanian the lowest, number of VMWEs in the training data. A notable exception to this tendency is Hindi, where good results (an F-score of 72.98) could be achieved building on a small amount of training data. This is probably due to the high number of multi-verb constructions (MVCs) in Hindi, which are usually formed by a sequence of two verbs, hence relatively easily identified by relying on POS tags.

Table 12 shows the effectiveness of MWE identification with regard to MWE categories. The highest F-scores were achieved for IRVs (especially for Balto-Slavic languages). This might be due to the fact that the IRVs tend to be continuous and must contain a reflexive pronoun/clitic, therefore the presence of such a pronoun in the immediate neighborhood of a verb is a strong predictor for IRVs. The LVC.full category is present in all languages. Interestingly, they are most effectively identified in the "Other" language group. Idioms occur in the test corpora of almost all languages (except Farsi), and they can be identified to the greatest extent in Romance languages. VPCs seem to be the easiest to find in Hungarian.

In regards to phenomenon-specific macro-average results (Tables 4 to 11), let us have a closer look at the F_1-MWE measure of the 11 systems which submitted results to all 19 languages, except MWE-TreeC (whose results are hard to interpret). The differences are: (i) from 13 to 28 points (17 points on average) for continuous vs. discontinuous VMWEs, (ii) from 14 to 43 points (27 points on average) for multitoken vs. single-token VMWEs, (iii) from 45 to 56 points (50 points on average) for seen-in-train vs. unseen-in-train VMWEs, and (iv) from 13 to 27 points (20 points on average) for identical-to-train vs. variant-of-train VMWEs. These results confirm that the phenomena they focus on are major challenges in the VMWE identification task, and we suggest that the corresponding measures should be systematically used for future evaluation. The hardest challenge is the one of identifying unseen-in-train VMWEs. This result is not a suprise since MWE-hood is, by nature, a lexical phenomenon, that is, a

[21] http://multiword.sourceforge.net/sharedtaskresults2018

particular idiomatic reading is available only in presence of a combination of particular lexical units. Replacing one of them by a semantically close lexeme usually leads to the loss of idiomatic reading, e.g. ***force*** *one's* ***hand*** 'compel someone to act against her will' is an idiom, while *force one's arm* can only be understood literally. Few other, non-lexical, hints are given to distinguish a particular VMWE occurrence from a literal expression, because a VMWE usually takes syntactically regular forms. Morphosyntactic idiosyncrasy (e.g. the fact that a given VMWE allows some and blocks some other regular syntactic transformations) is a property of types rather than tokens. We expect, therefore, satisfactory unseen-in-train VMWE identification results mostly from systems using large-scale VMWE lexicons or semi/unsupervised methods and very large corpora.

8 Conclusions and Future Work

We reported on edition 1.1 of the PARSEME Shared Task aiming at identifying verbal MWEs in texts in 20 languages. We described our corpus annotation methodology, the data provided to the participants, the shared task modalities and evaluation measures. The official results of the shared task were also presented and briefly discussed. The outputs of individual systems[22] should be compared more thoroughly in the future, so as to see how systems with different architectures cope with different phenomena. For instance, it would be interesting to check if, as expected, discontinuous VMWEs are handled better by parsing-based methods vs. sequential taggers, or by LSTMs vs. other neural network architectures.

Compared to the first edition in 2017, we attracted a larger number of participants (17 vs. 7), with 11 of the submissions covering 19 languages. We expect that this growing interest in modeling and computational treatment of verbal MWEs will motivate teams working on corpus annotation, especially from new language families, to join the initiative. We expect to maintain and continuously increase the quality and the size of the existing annotated corpora. For instance, we have identified weaknesses in the guidelines for MVCs that will require enhancements. Furthermore, we need to collect feedback about the IAV experimental category, and decide whether we consolidate its annotation guidelines.

Our ambitious goal for a future shared task is to extend annotation to other MWE categories, not only verbal ones. We are aware of corpora and guidelines for individual languages (e.g. English or French) and/or MWE categories (e.g. noun-noun compounds). However, a considerable effort will be required to design and apply universal annotation guidelines for the annotation of new MWE categories. We strongly believe that the large community and collective expertise gathered in the PARSEME initiative will allow us to take on this challenge. We definitely hope that this initiative will continue in the next years, yielding available multilingual annotated corpora that can foster MWE research in computational linguistics, as well as in linguistics and translation studies.

Acknowledgments

This work was supported by the IC1207 PARSEME COST action[23], and national funded projects: LD-PARSEME[24] (LD14117) in the Czech Republic, and PARSEME-FR[25] (ANR-14-CERA-0001) in France. Carla Parra Escartín is funded by the European Union's Horizon 2020 programme under the Marie Skłodowska-Curie grant agreement No 713567, and Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) at Dublin City University. Behrang QasemiZadeh and Timm Lichte are funded by the Deutsche Forschungsgemeinschaft (DFG) within the CRC 991 "The Structure of Representations in Language, Cognition, and Science". Veronika Vincze was supported by the UNKP-17-4 New National Excellence Program of the Ministry of Human Capacities, Hungary. The Slovenian team was supported by the Slovenian Research Agency via New grammar of contemporary standard Slovene: sources and methods (J6-8256 project). The Turkish team was supported by Boğaziçi University Research Fund Grant Number 14420. We are grateful to Maarten van Gompel for his help with adapting the FLAT

[22] Available at https://gitlab.com/parseme/sharedtask-data/tree/master/1.1/system-results.

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

[24] https://ufal.mff.cuni.cz/grants/ld-parseme

[25] http://parsemefr.lif.univ-mrs.fr/

annotation platform to our needs. Our thanks go also to all language leaders (LLs) and annotators, listed in Appendix A, for their their feedback on the annotation guidelines and preparing the annotated corpora.

References

Hazem Al Saied, Matthieu Constant, and Marie Candito. 2017. The ATILF-LLF system for Parseme shared task: a transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 127–132, Valencia, Spain, April. Association for Computational Linguistics.

Maria Jesús Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uria. 2015. Automatic conversion of the Basque dependency treebank to Universal Dependencies. In Markus Dickinsons, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski, editors, *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14)*, pages 233–241. Warszawa, Poland. Institute of Computer Science of the Polish Academy of Sciences.

Špela Arhar Holdt, Vojko Gorjanc, and Simon Krek. 2007. FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools. In *Proceedings of the Corpus Linguistics Conference, CL2007*, Birmingham.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia, 2015. *The Hindi/Urdu Treebank Project*. Springer Press.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of the First Conference on Machine Translation (WMT16), Volume 2: Shared Task Papers*, pages 131–198.

Tiberiu Boroş, Sonia Pipa, Verginica Barbu Mititelu, and Dan Tufiş. 2017. A data-driven approach to verbal multiword expression detection. PARSEME Shared Task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 121–126, Valencia, Spain, April. Association for Computational Linguistics.

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.

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of TALN 2012 (in French)*, Grenoble, France, June.

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.

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

Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged TreeBank. In Václav Matousek, Pavel Mautner, and Tomás Pavelka, editors, *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005*, Lecture Notes in Computer Science, pages 123–132, Berlin / Heidelberg, September. Springer.

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.

Natalia Klyueva, Antoine Doucet, and Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 60–65, Valencia, Spain, April. Association for Computational Linguistics.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling*, 0(1):65–110.

Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Polona Gantar, and Taja Kuzman. 2017. Training corpus ssj500k 2.0. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1165.

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

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

Alfredo Maldonado, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel, and Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 114–120, Valencia, Spain, April. Association for Computational Linguistics.

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.

Luka Nerima, Vasiliki Foufi, and Eric Wehrli. 2017. Parsing and MWE detection: Fips at the PARSEME shared task. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 54–59, Valencia, Spain, April. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1659–1666, Portorož, Slovenia, May.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik. 2011. National Corpus of Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 259–263, Poznań, Poland.

Behrang QasemiZadeh and Saeed Rahimi. 2006. Persian in MULTEXT-East framework. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing*, pages 541–551, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain, April. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebes kind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. forthcoming. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*. Language Science Press, Berlin, Germany.

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.

Katalin Ilona Simkó, Viktória Kovács, and Veronika Vincze. 2017. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 48–53, Valencia, Spain, April. Association for Computational Linguistics.

Mariona Taulé, Aina Peris, and Horacio Rodríguez. 2016. Iarg-AnCora: Spanish corpus annotated with implicit arguments. *Language Resources and Evaluation*, 50(3):549–584, Sep.

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.

Appendix A: Composition of the corpus anotation teams

Balto-Slavic languages: (BG) Ivelina Stoyanova (LL), Tsvetana Dimitrova, Svetlozara Leseva, Valentina Stefanova, Maria Todorova; (HR) Maja Buljan (LL), Goranka Blagus, Ivo-Pavao Jazbec, Kristina Kocijan, Nikola Ljubešić, Ivana Matas, Jan Šnajder; (LT) Jolanta Kovalevskaitė (LL), Agnė Bielinskienė, Loic Boizou; (PL) Agata Savary (LL), Emilia Palka-Binkiewicz; (SL): Polona Gantar (LL), Simon Krek (LL), Špela Arhar Holdt, Jaka Čibej, Teja Kavčič, Taja Kuzman.
Germanic languages: (DE) Timm Lichte (LL), Rafael Ehren; (EN) Abigail Walsh (LL), Claire Bonial, Paul Cook, Kristina Geeraert, John McCrae, Nathan Schneider, Clarissa Somers.
Romance languages: (ES) Carla Parra Escartín (LL), Cristina Aceta, Héctor Martínez Alonso; (FR) Marie Candito (LL), Matthieu Constant, Carlos Ramisch, Caroline Pasquer, Yannick Parmentier, Jean-Yves Antoine, Agata Savary; (IT) Johanna Monti (LL), Valeria Caruso, Maria Pia di Buono, Antonio Pascucci, Annalisa Raffone, Anna Riccio; (RO) Verginica Barbu Mititelu (LL), Mihaela Onofrei, Mihaela Ionescu; (PT) Renata Ramisch (LL), Aline Villavicencio, Carlos Ramisch, Helena de Medeiros Caseli, Leonardo Zilio, Silvio Ricardo Cordeiro.
Other languages: (AR) Abdelati Hawwari (LL), Mona Diab, Mohamed Elbadrashiny, Rehab Ibrahim; (EU) Uxoa Inurrieta (LL), Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez, Antton Gurrutxaga, Ruben Urizar; (EL) Voula Giouli (LL), Vassiliki Foufi, Aggeliki Fotopoulou, Stella Markantonatou, Stella Papadelli, Natasa Theoxari; (FA) Behrang QasemiZadeh (LL), Shiva Taslimipoor; (HE) Chaya Liebeskind (LL), Yaakov Ha-Cohen Kerner (LL), Hevi Elyovich, Ruth Malka; (HI) Archna Bhatia (LL), Ashwini Vaidya (LL), Kanishka Jain, Vandana Puri, Shraddha Ratori, Vishakha Shukla, Shubham Srivastava; (HU) Veronika Vincze (LL), Katalin Simkó, Viktória Kovács; (TR) Tunga Güngör (LL), Gözde Berk, Berna Erden.

Appendix B: Shared task results

Lang-split	Sent.	Tok.	Sent. length	VMWE	VID	IRV full	LVC cause	LVC full	VPC semi	VPC	IAV	MVC ICV	LS
AR-train	2370	231030	97.4	3219	1272	17	940	0	957	0	0	33	0
AR-dev	387	16252	41.9	500	17	0	419	0	64	0	0	0	0
AR-test	380	17962	47.2	500	31	0	410	0	59	0	0	0	0
AR-Total	3137	265244	84.5	4219	1320	17	1769	0	1080	0	0	33	
BG-train	17813	399173	22.4	5364	1005	2729	1421	135	0	0	74	0	0
BG-dev	1954	42020	21.5	670	173	240	214	35	0	0	8	0	0
BG-test	1832	39220	21.4	670	82	254	274	52	0	0	8	0	0
BG-Total	21599	480413	22.2	6704	1260	3240	1909	222	0	0	90	0	0
DE-train	6734	130588	19.3	2820	977	220	218	28	1264	113	0	0	0
DE-dev	1184	22146	18.7	503	181	48	34	2	221	17	0	0	0
DE-test	1078	20559	19	500	183	40	42	2	210	23	0	0	0
DE-Total	8996	173293	19.2	3823	1341	3548	294	32	1695	153	0	0	0
EL-train	4427	122458	27.6	1404	395	0	938	44	19	0	0	8	0
EL-dev	2562	66431	25.9	500	81	0	376	34	8	0	0	1	0
EL-test	1261	35873	28.4	501	169	0	308	11	11	0	0	2	0
EL-Total	8250	224762	27.2	2405	645	3548	1622	89	38	0	0	11	0
EN-train	3471	53201	15.3	331	60	0	78	7	151	19	16	0	0
EN-test	3965	71002	17.9	501	79	0	166	36	146	26	44	4	0
EN-Total	7436	124203	16.7	832	139	3548	244	43	297	45	60	4	0
ES-train	2771	96521	34.8	1739	167	479	223	36	0	0	360	474	0
ES-dev	698	26220	37.5	500	65	114	84	17	0	0	87	133	0
ES-test	2046	59623	29.1	500	95	121	85	28	1	0	64	106	0
ES-Total	5515	182364	33	2739	327	4262	392	81	1	0	511	713	0
EU-train	8254	117165	14.1	2823	597	0	2074	152	0	0	0	0	0
EU-dev	1500	21604	14.4	500	104	0	382	14	0	0	0	0	0
EU-test	1404	19038	13.5	500	73	0	410	17	0	0	0	0	0
EU-Total	11158	157807	14.1	3823	774	4262	2866	183	0	0	0	0	0
FA-train	2784	45153	16.2	2451	17	1	2433	0	0	0	0	0	0
FA-dev	474	8923	18.8	501	0	0	501	0	0	0	0	0	0
FA-test	359	7492	20.8	501	0	0	501	0	0	0	0	0	0
FA-Total	3617	61568	17	3453	17	4263	3435	0	0	0	0	0	0

Continued on next page.

Continued from previous page.

Lang-split	Sent.	Tok.	Sent. length	VMWE	VID	IRV full	LVC cause	LVC full	VPC semi	VPC	IAV	MVC ICV	LS
FR-train	17225	432389	25.1	4550	1746	1247	1470	68	0	0	0	19	0
FR-dev	2236	56254	25.1	629	207	154	252	15	0	0	0	1	0
FR-test	1606	39489	24.5	498	212	108	160	14	0	0	0	4	0
FR-Total	21067	528132	25	5677	2165	5772	1882	97	0	0	0	24	0
HE-train	12106	237472	19.6	1236	519	0	545	113	59	0	0	0	0
HE-dev	3385	65843	19.4	501	258	0	148	61	34	0	0	0	0
HE-test	3209	65698	20.4	502	182	0	211	49	60	0	0	0	0
HE-Total	18700	369013	19.7	2239	959	5772	904	223	153	0	0	0	0
HI-train	856	17850	20.8	534	23	0	321	14	0	0	0	176	0
HI-test	828	17580	21.2	500	38	0	320	12	0	0	0	130	0
HI-Total	1684	35430	21	1034	61	5772	641	26	0	0	0	306	0
HR-train	2295	53486	23.3	1450	113	468	303	45	0	0	521	0	0
HR-dev	834	19621	23.5	500	34	139	143	26	1	0	157	0	0
HR-test	708	16429	23.2	501	33	118	131	31	0	0	188	0	0
HR-Total	3837	89536	23.3	2451	180	6497	577	102	1	0	866	0	0
HU-train	4803	120013	24.9	6205	84	0	892	363	4131	735	0	0	0
HU-dev	601	15564	25.8	779	10	0	85	10	539	135	0	0	0
HU-test	755	20759	27.4	776	10	0	166	28	486	86	0	0	0
HU-Total	6159	156336	25.3	7760	104	6497	1143	401	5156	956	0	0	0
IT-train	13555	360883	26.6	3254	1098	942	544	147	66	0	414	23	20
IT-dev	917	32613	35.5	500	197	106	100	19	17	2	44	6	9
IT-test	1256	37293	29.6	503	201	96	104	25	23	0	41	5	8
IT-Total	15728	430789	27.3	4257	1496	7641	748	191	106	2	499	34	37
LT-train	4895	90110	18.4	312	106	0	195	11	0	0	0	0	0
LT-test	6209	118402	19	500	202	0	284	14	0	0	0	0	0
LT-Total	11104	208512	18.7	812	308	7641	479	25	0	0	0	0	0
PL-train	13058	220465	16.8	4122	373	1785	1531	180	0	0	253	0	0
PL-dev	1763	26030	14.7	515	57	245	153	33	0	0	27	0	0
PL-test	1300	27823	21.4	515	73	249	149	15	0	0	29	0	0
PL-Total	16121	274318	17	5152	503	9920	1833	228	0	0	309	0	0
PT-train	22017	506773	23	4430	882	689	2775	84	0	0	0	0	0
PT-dev	3117	68581	22	553	130	83	337	3	0	0	0	0	0
PT-test	2770	62648	22.6	553	118	91	337	7	0	0	0	0	0
PT-Total	27904	638002	22.8	5536	1130	10783	3449	94	0	0	0	0	0
RO-train	42704	781968	18.3	4713	1269	3048	250	146	0	0	0	0	0
RO-dev	7065	118658	16.7	589	169	373	29	18	0	0	0	0	0
RO-test	6934	114997	16.5	589	173	363	34	19	0	0	0	0	0
RO-Total	56703	1015623	17.9	5891	1611	14567	313	183	0	0	0	0	0
SL-train	9567	201853	21	2378	500	1162	176	40	0	0	500	0	0
SL-dev	1950	38146	19.5	500	121	224	30	12	0	0	113	0	0
SL-test	1994	40523	20.3	500	106	245	35	13	0	0	101	0	0
SL-Total	13511	280522	20.7	3378	727	16198	241	65	0	0	714	0	0
TR-train	16715	334880	20	6125	3172	0	2952	0	0	0	0	1	0
TR-dev	1320	27196	20.6	510	285	0	225	0	0	0	0	0	0
TR-test	577	14388	24.9	506	233	0	272	0	0	0	0	1	0
TR-Total	18612	376464	20.2	7141	3690	16198	3449	0	0	0	0	2	0
Total	280838	6072331	21.6	79326	18757	16198	28190	2285	8527	1156	3049	1127	37

Table 2: Statistics on the training (train), development (dev), and test corpora. Number of sentences (Sent.), number of tokens (Tok.), average sentence length in number of tokens (Sent. length), total number of annotated VMWEs (VMWE), and number of annotated VMWEs broken down by category (VID, IRV, . . .)

System	Track	#Langs	P MWE	R MWE	F1 MWE	Rank MWE	P Tok	R Tok	F1 Tok	Rank Tok
TRAVERSAL	closed	19/19	67.58	44.97	54	1	77.41	48.55	59.67	1
TRAPACC_S	closed	19/19	62.28	41.4	49.74	2	68.54	42.06	52.13	4
TRAPACC	closed	19/19	55.68	44.67	49.57	3	62.1	46.37	53.09	3
CRF-Seq-nocategs	closed	19/19	56.13	39.12	46.11	4	73.44	43.49	54.63	2
varIDE	closed	19/19	61.49	36.71	45.97	5	64.13	37.63	47.43	6
CRF-DepTree-categs	closed	19/19	52.33	37.83	43.91	6	64.65	41.56	50.6	5
GBD-NER-standard	closed	19/19	36.56	48.3	41.62	7	41.11	52.21	46	7
GBD-NER-resplit	closed	19/19	30.26	52.95	38.51	8	33.83	58.03	42.74	9
Veyn	closed	19/19	42.76	32.51	36.94	9	58.13	36.57	44.9	8
mumpitz	closed	7/19	17.14	13.03	14.81	10	24.95	15.5	19.12	11
Polirem-rich	closed	3/19	10.9	2.87	4.54	11	13.07	3.89	6	12
Polirem-basic	closed	3/19	10.78	0.65	1.23	12	11.33	0.68	1.28	13
MWETreeC	closed	19/19	0.21	3.72	0.4	13	23.5	24.78	24.12	10
SHOMA	open	19/19	66.08	51.82	58.09	1	76.22	54.27	63.4	1
Deep-BGT	open	10/19	33.41	25.29	28.79	2	39.77	26.47	31.78	2
Milos	open	4/19	9.17	7.87	8.47	3	11.5	8.25	9.61	3
mumpitz-preinit	open	1/19	2.28	1.9	2.07	4	3.71	2.35	2.88	4

Table 3: General results.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
TRAVERSAL	closed	19/19	68.19	49.78	57.55	1
TRAPACC_S	closed	19/19	65.12	48.18	55.38	2
TRAPACC	closed	19/19	59.09	51.99	55.31	3
CRF-Seq-nocategs	closed	19/19	54.99	49.84	52.29	4
varIDE	closed	19/19	78.03	37.98	51.09	5
CRF-DepTree-categs	closed	19/19	52.8	42.44	47.06	6
GBD-NER-standard	closed	19/19	38.76	55.2	45.54	7
GBD-NER-resplit	closed	19/19	33.5	57.92	42.45	8
Veyn	closed	19/19	41.76	37.76	39.66	9
mumpitz	closed	7/19	16.83	15.32	16.04	10
Polirem-rich	closed	3/19	10.9	4.78	6.65	11
Polirem-basic	closed	3/19	10.78	1.09	1.98	12
MWETreeC	closed	19/19	0.21	4.21	0.4	13
SHOMA	open	19/19	66.07	59.73	62.74	1
Deep-BGT	open	10/19	36.05	27.54	31.23	2
Milos	open	4/19	9.42	9.49	9.45	3
mumpitz-preinit	open	1/19	1.97	2.31	2.13	4

Table 4: Results for continuous MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
TRAVERSAL	closed	19/19	61.14	34.81	44.36	1
varIDE	closed	19/19	44.53	32.24	37.4	2
CRF-DepTree-categs	closed	19/19	48.8	26.4	34.26	3
TRAPACC_S	closed	19/19	53.23	24.88	33.91	4
TRAPACC	closed	19/19	43.29	27.3	33.48	5
GBD-NER-standard	closed	19/19	29.32	33.69	31.35	6
GBD-NER-resplit	closed	19/19	23.22	41.41	29.76	7
Veyn	closed	19/19	40.53	19.07	25.94	8
CRF-Seq-nocategs	closed	19/19	54.2	15.48	24.08	9
mumpitz	closed	7/19	18.34	8.71	11.81	10
Polirem-rich	closed	3/19	3.51	0.06	0.12	11
Polirem-basic	closed	3/19	0	0	0	n/a
MWETreeC	closed	19/19	0	0	0	n/a
SHOMA	open	19/19	62.95	32.87	43.19	1
Deep-BGT	open	10/19	28.83	19.4	23.19	2
Milos	open	4/19	9.37	5.79	7.16	3
mumpitz-preinit	open	1/19	3.25	1.44	2	4

Table 5: Results for discontinuous MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
TRAVERSAL	closed	19/19	74.66	44.59	55.83	1
TRAPACC	closed	19/19	57.23	43.42	49.38	2
TRAPACC_S	closed	19/19	63.6	39.97	49.09	3
CRF-Seq-nocategs	closed	19/19	66.16	38.23	48.46	4
CRF-DepTree-categs	closed	19/19	60.52	37.21	46.09	5
varIDE	closed	19/19	61.49	36.18	45.56	6
GBD-NER-standard	closed	19/19	36.56	51.05	42.61	7
GBD-NER-resplit	closed	19/19	30.26	55.86	39.26	8
Veyn	closed	19/19	52.16	30.33	38.36	9
mumpitz	closed	7/19	22.23	12.47	15.98	10
Polirem-rich	closed	3/19	10.9	2.87	4.54	11
Polirem-basic	closed	3/19	10.78	0.65	1.23	12
MWETreeC	closed	19/19	0	0	0	n/a
SHOMA	open	19/19	73.37	50.65	59.93	1
Deep-BGT	open	10/19	35.04	25.09	29.24	2
Milos	open	4/19	10.37	6.89	8.28	3
mumpitz-preinit	open	1/19	3	1.61	2.1	4

Table 6: Results for multi-token MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
TRAPACC	closed	5/5	35.13	30.8	32.82	1
TRAPACC_S	closed	5/5	34.64	30.49	32.43	2
TRAVERSAL	closed	5/5	30.7	22.49	25.96	3
CRF-DepTree-categs	closed	5/5	28.49	21.81	24.71	4
Veyn	closed	5/5	22.91	25.76	24.25	5
CRF-Seq-nocategs	closed	5/5	24.95	22.69	23.77	6
varIDE	closed	5/5	36.47	6.43	10.93	7
mumpitz	closed	1/5	4.87	12.62	7.03	8
MWETreeC	closed	5/5	0.79	61.8	1.56	9
Polirem-rich	closed	0/5	0	0	0	n/a
Polirem-basic	closed	0/5	0	0	0	n/a
GBD-NER-standard	closed	5/5	0	0	0	n/a
GBD-NER-resplit	closed	5/5	0	0	0	n/a
SHOMA	open	5/5	27.77	28.9	28.32	1
Deep-BGT	open	3/5	27.61	24.33	25.87	2
Milos	open	2/5	12.23	15.84	13.8	3
mumpitz-preinit	open	1/5	6.43	9.8	7.77	4

Table 7: Results for single-token MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
TRAVERSAL	closed	19/19	86.54	63	72.92	1
GBD-NER-resplit	closed	19/19	82.76	63.82	72.07	2
TRAPACC	closed	19/19	82.72	61.41	70.49	3
GBD-NER-standard	closed	19/19	82.92	60.74	70.12	4
TRAPACC_S	closed	19/19	82.04	57.06	67.31	5
CRF-Seq-nocategs	closed	19/19	78.27	52.71	63	6
CRF-DepTree-categs	closed	19/19	83.23	50.03	62.49	7
varIDE	closed	19/19	62.8	56.2	59.32	8
Veyn	closed	19/19	76.6	43.7	55.65	9
mumpitz	closed	7/19	30.69	17.81	22.54	10
Polirem-rich	closed	3/19	14.98	4.44	6.85	11
MWETreeC	closed	19/19	13.16	3.99	6.12	12
Polirem-basic	closed	3/19	15.79	1.16	2.16	13
SHOMA	open	19/19	89	66.78	76.31	1
Deep-BGT	open	10/19	46.25	30.36	36.66	2
Milos	open	4/19	16.46	10.4	12.75	3
mumpitz-preinit	open	1/19	4.34	3.07	3.6	4

Table 8: Results for seen-in-train MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
GBD-NER-standard	closed	19/19	14.33	31.54	19.71	1
GBD-NER-resplit	closed	19/19	12.74	37.66	19.04	2
TRAVERSAL	closed	19/19	23.94	13.61	17.35	3
CRF-DepTree-categs	closed	19/19	18.71	15.58	17	4
TRAPACC	closed	19/19	19.19	14.52	16.53	5
TRAPACC_S	closed	19/19	24.07	12.47	16.43	6
CRF-Seq-nocategs	closed	19/19	20.49	13.63	16.37	7
Veyn	closed	19/19	11.57	10.58	11.05	8
mumpitz	closed	7/19	5.5	5.92	5.7	9
varIDE	closed	19/19	14.61	3.31	5.4	10
Polirem-rich	closed	3/19	1.76	0.36	0.6	11
MWETreeC	closed	19/19	0.02	1.99	0.04	12
Polirem-basic	closed	3/19	0	0	0	n/a
SHOMA	open	19/19	31.73	25.8	28.46	1
Deep-BGT	open	10/19	12.99	13	12.99	2
Milos	open	4/19	5.56	5.89	5.72	3
mumpitz-preinit	open	1/19	0.75	0.72	0.73	4

Table 9: Results for unseen-in-train MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
TRAPACC	closed	19/19	90.44	77.94	83.73	1
TRAVERSAL	closed	19/19	89.15	75.71	81.88	2
TRAPACC_S	closed	19/19	85.56	73.04	78.81	3
GBD-NER-resplit	closed	19/19	87.27	71.18	78.41	4
GBD-NER-standard	closed	19/19	87.39	69.44	77.39	5
CRF-Seq-nocategs	closed	19/19	80.32	70.54	75.11	6
CRF-DepTree-categs	closed	19/19	85.85	60.25	70.81	7
varIDE	closed	19/19	82.23	57.52	67.69	8
Veyn	closed	19/19	81.15	53.37	64.39	9
mumpitz	closed	7/19	31.57	22.25	26.1	10
Polirem-rich	closed	3/19	15.31	5.99	8.61	11
MWETreeC	closed	19/19	13.16	4.69	6.92	12
Polirem-basic	closed	3/19	15.79	2.03	3.6	13
SHOMA	open	19/19	90.26	85.15	87.63	1
Deep-BGT	open	10/19	46.45	36.71	41.01	2
Milos	open	4/19	17.2	11	13.42	3
mumpitz-preinit	open	1/19	4.25	3.66	3.93	4

Table 10: Results for identical-to-train MWEs.

System	Track	#Langs	P-MWE	R-MWE	F1-MWE	Rank-MWE
GBD-NER-resplit	closed	19/19	76.6	56.48	65.02	1
TRAVERSAL	closed	19/19	83.22	50.82	63.1	2
GBD-NER-standard	closed	19/19	76.63	52.16	62.07	3
TRAPACC	closed	19/19	74.73	47.22	57.87	4
TRAPACC_S	closed	19/19	76.22	43.11	55.07	5
varIDE	closed	19/19	52.7	53.97	53.33	6
CRF-DepTree-categs	closed	19/19	78.29	39.01	52.07	7
CRF-Seq-nocategs	closed	19/19	73.17	35.48	47.79	8
Veyn	closed	19/19	70.65	34.62	46.47	9
mumpitz	closed	7/19	29.96	13.77	18.87	10
Polirem-rich	closed	3/19	14.51	3.21	5.26	11
MWETreeC	closed	19/19	7.89	2.16	3.39	12
Polirem-basic	closed	3/19	10.53	0.48	0.92	13
SHOMA	open	19/19	85.95	50.03	63.25	1
Deep-BGT	open	10/19	45.04	22.42	29.94	2
Milos	open	4/19	15.96	9.97	12.27	3
mumpitz-preinit	open	1/19	4.44	2.67	3.33	4

Table 11: Results for variant-of-train MWEs.

	IAV		IRV		LVC.cause		LVC.full		MVC		VID		VPC.full		VPC.semi		LS.IVC	
	MF	TF	MF	TF	MF	TF	MF	TF	MF	TF	MF	TF	MF	TF	MF	TF	MF	TF
BG	0.00	0.92	65.56	66.20	13.82	14.19	36.67	37.57			22.68	25.70						
HR	31.61	45.03	42.34	48.00	9.17	10.32	19.16	21.11			4.83	6.40						
LT					18.31	19.74	17.05	21.61			5.20	5.77						
PL	34.89	44.17	58.36	64.16	9.13	10.80	34.01	38.12			13.34	19.64						
SL	33.38	37.60	48.95	50.97	7.78	13.42	15.09	17.38			12.03	16.10						
AV	24.97	31.93	53.80	57.33	11.64	13.69	24.40	27.16			11.61	14.72						
ES	17.23	23.24	27.62	30.35	3.29	4.24	12.89	16.54	19.72	26.29	10.16	13.53	0.00	0.00				
FR			46.73	50.27	2.91	3.37	30.55	35.50	7.28	7.28	34.31	42.88						
IT	20.69	26.41	31.08	33.30	17.58	20.21	24.16	26.05	9.52	9.66	16.58	20.66	19.93	21.02			4.93	7.31
PT			49.70	50.45	11.09	14.82	43.80	46.55			31.73	35.64						
RO			69.34	74.46	65.56	77.74	52.26	55.88			62.96	68.24						
AV	18.96	24.83	44.89	47.76	20.09	24.08	32.73	36.10	12.17	14.41	31.15	36.19	9.97	10.51			4.93	7.31
DE			20.75	32.05	5.56	4.76	5.69	7.94			18.99	25.15	45.47	49.11	10.77	11.64		
EN	7.92	8.75			0.00	0.00	10.15	11.25	0.00	0.00	3.68	3.70	28.16	30.22	2.79	5.06		
AV	7.92	8.75	20.75	32.05	2.78	2.38	7.92	9.60	0.00	0.00	11.33	14.43	36.82	39.66	6.78	8.35		
EL					2.03	3.05	38.36	46.37	12.12	14.14	21.36	27.88	7.68	12.66				
EU					19.19	20.03	57.24	59.82			28.37	30.03						
FA							62.07	70.10										
HE					15.77	20.03	19.13	23.29			13.54	18.40	0.00	0.00				
HI					7.57	8.69	57.01	63.02	68.95	73.68	7.21	6.30						
HU					49.91	52.33	45.68	50.75			47.36	56.09	61.47	64.04	58.74	58.54		
TR							25.13	27.83	0.00	0.00	19.33	22.21						
AV					23.11	25.27	44.38	49.13	34.48	36.84	23.16	26.61	30.73	32.02	58.74	58.54		
MA	19.76	25.16	44.61	49.03	15.06	17.29	31.11	34.72	16.42	18.23	20.50	24.37	24.02	25.92	27.57	28.43	4.93	7.31

Table 12: Average F1-scores per category for each language and language group. MF: MWE-based F1-score, TF: token-based F1-score, AV: average, MA: macro-average.