Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions

Carlos Ramisch
Aix Marseille Univ, CNRS, LIS, France

Agata Savary
University of Tours, France

Bruno Guillaume
LORIA/Inria Nancy Grand-Est, France

Jakub Waszczuk
University of Duesseldorf, Germany

Marie Candito
Paris Diderot University, France

Ashwini Vaidya
IIT Delhi, India

Verginica Barbu Mititelu
Romanian Academy, Romania

Archna Bhatia
Florida IHMC, USA

Uxoa Iñurrieta
Univ. of the Basque Country, Spain

Voula Giouli
Athena Research Center, Greece

Tunga Güngör
Boğaziçi University Turkey

Menghan Jiang
PolyU, Hong Kong China

Timm Lichte
University of Tübingen, Germany

Chaya Liebeskind
Jerusalem College of Technology, Israel

Johanna Monti
"L'Orientale" University of Naples, Italy

Renata Ramisch
NILC, UFSCar, Brazil

Sara Stymne
Uppsala University, Sweden

Abigail Walsh
Dublin City University, Ireland

Hongzhi Xu
Shanghai International Studies Univ., China

Abstract

We present edition 1.2 of the PARSEME shared task on identification of verbal multiword expressions (VMWEs). Lessons learned from previous editions indicate that VMWEs have low ambiguity, and that the major challenge lies in identifying test instances never seen in the training data. Therefore, this edition focuses on unseen VMWEs. We have split annotated corpora so that the test corpora contain around 300 unseen VMWEs, and we provide non-annotated raw corpora to be used by complementary discovery methods. We released annotated and raw corpora in 14 languages, and this semi-supervised challenge attracted 7 teams who submitted 9 system results. This paper describes the effort of corpus creation, the task design, and the results obtained by the participating systems, especially their performance on unseen expressions.

1 Introduction

Multiword expressions (MWEs) such as *to **throw** someone **under the bus*** 'to cause one's suffering to gain personal advantage' are idiosyncratic word combinations which need to be identified prior to further semantic processing (Baldwin and Kim, 2010; Calzolari et al., 2002). The task of MWE identification, that is, automatically locating instances of MWEs in running text (Constant et al., 2017) has received growing attention in the last 4 years. Progress on this task was especially motivated by shared tasks such as DiMSUM (Schneider et al., 2016), and two editions of the PARSEME shared tasks, edition 1.0 in 2017 (Savary et al., 2017), and edition 1.1 in 2018 (Ramisch et al., 2018).

Previous editions of the PARSEME shared task focused on the identification of verbal MWEs (VMWEs), because of their challenging traits: complex structure, discontinuities, variability, ambiguity, etc. (Savary et al., 2017). The problem is addressed from a multilingual perspective: editions 1.0 and 1.1 covered 18 and 20 languages, respectively. The annotation guidelines and methodology are unified across languages, offering a rich playground for system developers.

The framework proposed by the (closed track of) previous shared tasks was tailored for supervised learning. An annotated training corpus for each language was made available for system developers. The

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 107–118
Barcelona, Spain (Online), December 13, 2020.

systems, building mostly on statistical and deep learning techniques, were then able to identify MWEs in the test data based on regularities learned from the training corpora. The strength of supervised machine learning approaches lies in (a) contextual disambiguation and (b) generalisation power. In other words, the identification of ambiguous expressions should be conditioned on their contexts, and new expressions or variants should be identified even if they were not observed in the training corpus.

However, corpus studies show that supervised methods can take limited advantage of these strengths for VMWE identification. Firstly, even if a number of studies have been dedicated to contextual disambiguation (between idiomatic and literal occurrences of MWEs), recent work shows that this task is quantitatively of minor importance, because literal readings occur surprisingly rarely in corpora. Namely, based on manual annotation in German, Greek, Basque, Polish, and Brazilian Portuguese, Savary et al. (2019b) discovered that most expressions are potentially ambiguous, but the vast majority of them never occur literally nor accidentally.

Secondly, MWE idiosyncrasies manifest at the level of types (sets of occurrences of the same expression) and not at the level of tokens (single occurrences). This fact, in addition to MWE's Zipfian distribution and low proliferation rate, makes it unlikely to detect new MWEs based on a few instances of known ones (Savary et al., 2019a). Thus, the generalisation power of supervised learning only applies to variants of expressions already observed in the training data.

These two findings motivated the current edition of the PARSEME shared task focusing on the identification of *unseen VMWEs*. A VMWE annotated in the test set is considered unseen if the multi-set of lemmas of its lexicalised components was never annotated in the training data.[1] Differently from edition 1.1, by training data we understand all the gold data released before the training stage, i.e. both the subset meant for training proper (train) and the one meant for development/fine-tuning (dev). Therefore, the main novelties in this edition are:

1. Evaluation is not only based on overall F1, but emphasises performance on unseen VMWEs;
2. Corpora are split so that test sets contain at least 300 VMWEs unseen in training sets;
3. Raw corpora are provided to foster the development of semi-supervised VMWE discovery;
4. Unseen VMWEs are now defined with respect to train and dev sets, rather than train alone.

Moreover, we extended and enhanced the corpus annotation effort, both in terms of languages covered and of methods to increase the quality of existing corpora. This included a stronger integration with the Universal Dependencies (UD) framework.[2] The remainder of this paper describes the design of edition 1.2 of the PARSEME shared task, and summarises its outcomes.[3]

2 Manually Annotated Corpora

The corpus used in the shared task and the underlying cross-lingually unified and validated annotation guidelines result from continuous efforts of a multilingual community since 2015.[4] The 1.2 guidelines mostly follow those from edition 1.1, with decision flowcharts based on linguistic tests, allowing annotators to identify and categorise candidates into the following categories:[5]

- inherently reflexive verbs (IRVs), e.g. FR ***se rendre*** (lit. 'return oneself') 'go'
- light verb constructions (LVCs), with 2 subcategories:
  - LVC.full, e.g. HE **לתת הסכמה** (lit. 'give consent') 'approve'
  - LVC.cause, e.g. RO ***pune la dispoziție*** (lit. 'put at disposal') 'make available'
- verbal idioms (VIDs), e.g. TR ***ileri sürmek*** (lit. 'lead forward') 'assert'
- verb-particle constructions (VPCs), with 2 subcategories:

[1] Instances whose lemmas match, but with different *forms* in training and test data, are considered seen VMWEs. We also distinguish seen-variant from seen-identical occurrences, to account for form mismatches.

[2] http://universaledependencies.org

[3] Although this paper was submitted anonymously and peer reviewed, the process may have been biased by public information about the shared task published online, including the names of organizers and language leaders who author this paper.

[4] https://gitlab.com/parseme/corpora/-/wikis/home

[5] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/

  - VPC.full, e.g. DE ***stellt her*** (lit. 'puts here') 'produces'
  - VPC.semi, e.g. ZH 获取到 (lit. 'capture arrive/to') 'capture'
- multi-verb constructions (MVCs), e.g. HI बैठ गया (lit. 'sit went') 'sat down'
- inherently adpositional verbs (IAVs), annotated non-systematically on an experimental basis, e.g. IT ***intendersi di*** (lit. 'understand of') 'to know about'

The only changes to these guidelines are language-specific additions: (i) a Chinese-specific decision tree for MVCs, (ii) two Swedish-specific sections about identifying multiword tokens and distinguishing particles from prepositions and prefixes.

The manually annotated corpus for edition 1.2 covers 14 languages: German (DE), Basque (EU), Greek (EL), French (FR), Irish (GA), Hebrew (HE), Hindi (HI), Italian (IT), Polish (PL), Brazilian Portuguese (PT), Romanian (RO), Swedish (SV), Turkish (TR) and Chinese (ZH).[6]

New Languages The underlined languages in the list above are those whose corpora are new or substantially increased with respect to editions 1.0 and 1.1.[7]

Chinese is the first language in the PARSEME collection in which word boundaries are not spelled out in running text. Thus, tokenisation constitutes a major challenge. We used previously tokenised texts from the Chinese UD treebank and some raw texts from the CoNLL 2017 parsing shared task corpus.[8] The latter was tokenised automatically and manually corrected when segmentation errors affected the right scope of a VMWE. About 48% of the annotated VMWEs consist in a single (multiword) token.

Irish is our first language of the Celtic genus, with new VMWE-related challenges. Firstly, frequent contractions of prepositions with personal pronouns make it hard to annotate IAVs. The preposition is usually lexicalised while the pronoun is not, as in GA ***chuir*** *sé* ***orm*** (lit. 'put he on-me') 'he bothered me'. However, since these contractions are seen in UD as inflected prepositions, they are represented as single words and lemmatised into the preposition alone.[9] Therefore, the only possible VMWE annotation is to consider the pronoun as an inflectional ending, i.e. part of the lexicalised preposition (***chuir*** *sé* ***orm***). Secondly, some copula constructions, like GA *X is ainm dom* (lit. 'X is name to-me') 'my name is X', are idiomatic and would normally find their place among the VIDs. This is, however, currently not possible because, according to our guidelines, a VMWE (in its syntactically least marked form) has to be headed by a verb. However, following the UD lexicalist morphosyntactic annotation principles, the head of a copula construction is the predicative noun (*ainm* 'name') rather than the copula (*is* 'is').

Swedish had a small annotated corpus in edition 1.0, but the new corpus was annotated from scratch. The main challenge was related to particle-verb combinations occurring as single tokens. Some of them can be seen either as unique words, i.e. no VMWE candidates, or as multiword tokens (MWTs), i.e. potential VPCs. This depends on whether they can occur both in the joint (one-token) and in the split (two-token) configuration, with the same or a different meaning. For instance, SV ***pågå*** (lit. 'on-go') 'be in progress' can be split but only with a changed meaning SV ***gå på*** (lit. 'go on') 'keep bringing the same issue up'. In SV ***överleva*** (lit. 'over-live') 'survive' the particle (*över*) is easily distinguished from the verb but the split configuration never occurs. Other compound verbs, like SV ***sysselsätta*** (lit. 'activity-put') 'put into work', cannot be split either. Currently, all such cases are considered MWTs and annotated as VPCs or VIDs. About 49% of the annotated VMWEs contain a single (multiword) token.

Enhancements in Previous Languages For all other 11 languages, the current corpus builds upon edition 1.1, with some extensions and enhancements. In Greek, Hebrew, Polish and Brazilian Portuguese, new texts were annotated (mostly in the centralised FLAT platform)[10], which increased the pre-existing

[6]The annotated corpus for the 1.2 edition is available at `http://hdl.handle.net/11234/1-3367`

[7]Some languages present in editions 1.0 and 1.1 are not covered because the corpora were not upgraded: Arabic, Bulgarian, Croatian, Czech, English, Farsi, Hungarian, Lithuanian, Maltese, Slovene and Spanish.

[8]`http://hdl.handle.net/11234/1-2184`

[9]Note that other languages also have inflected (reflexive) pronouns, e.g. in IRVs: FR *je* ***me rends*** (lit. 'I return myself') 'I go', *il* ***se rend*** (lit. 'he returns himself') 'he goes', etc. The difference is that, in the Irish examples, the pronoun is not lexicalized and should normally not be annotated as a VMWE component.

[10]`https://proycon.anaproy.nl/software/flat/`

	S	A_1	A_2	F_{span}	κ_{span}	κ_{cat}
Greek (EL)	$874_{(1617)}$	$293_{(428)}$	$394_{(462)}$	$0.652_{(0.694)}$	$0.608_{(0.665)}$	$\mathbf{0.715}_{(0.673)}$
Irish (GA)	800	312	270	0.715	0.663	0.835
Polish (PL)	$900_{(2079)}$	$252_{(759)}$	$296_{(707)}$	$\mathbf{0.774}_{(0.619)}$	$\mathbf{0.732}_{(0.568)}$	$\mathbf{0.907}_{(0.882)}$
Br. Portuguese (PT)	$1251_{(1000)}$	$253_{(275)}$	$232_{(241)}$	$0.672_{(0.713)}$	$0.640_{(0.684)}$	$\mathbf{0.928}_{(0.837)}$
Swedish (SV)	700	364	257	0.734	0.671	0.847
Chinese (ZH)	3953	883	840	0.584	0.544	0.833

Table 1: Inter-annotator agreement on S sentences with A_1 and A_2 VMWEs per annotator. F_{span} shows inter-annotator F-measure, κ_{span} shows chance-corrected agreement on annotation span, and κ_{cat} on category. Subscripts indicate agreement in edition 1.1 (on different samples).

corpora by 13%-209% in terms of the annotated VMWEs. In other languages, previous annotations were corrected in the layers of tokenisation, lemmatisation, morphosyntax or VMWEs.

Quality All 14 languages now benefit from morphosyntactic tagsets compatible with UD version 2. The tokenisation, lemmatisation, and morphosyntactic layers contain manual annotations for some languages (Chinese, French, Irish, Italian, Swedish, partly German, Greek, Polish and Portuguese) and automatic ones for the others (mostly with UDPipe[11] trained on UD version 2.5). The homogenisation of the morphosyntactic layer via a widely adopted framework such as UD facilitates the development of tools for corpus processing as well as for MWE identification by shared task participants.

In each language, most of the VMWE annotations were performed by a single annotator per file, except for Chinese and Turkish, where double annotation and adjudication was systematic. In most languages the post-annotation use of a custom consistency checking tool helped to reduce silence and noise (Savary et al., 2018, section 5.4). For the data annotated from scratch in edition 1.2 (Chinese, Greek, Irish, Polish and Portuguese)[12] we performed double annotation of a sample to estimate inter-annotator agreement (Savary et al., 2017; Ramisch et al., 2018). Compared to edition 1.1 (where roughly the same guidelines and methodology were used), the scores presented in Tab. 1 for Greek, Polish and Portuguese are clearly higher for categorisation.[13] For span, they are slightly lower in Greek and Portuguese but significantly higher in Polish. For all 6 languages, the contrast between the last two columns confirms the observation of previous editions that, once a VMWE has been correctly identified by an annotator, assigning it to the correct category is significantly easier.

Finally, we applied a set of validation scripts to ensure that all files respect the CUPT format (see below); each VMWE has a single category label among those specified in the guidelines; all dependency trees are acyclic; the mandatory metadata `text` and `source_sent_id` are present and the latter is well formatted; and that the same set of tokens is never annotated twice.

Corpus Release The annotated corpora were split into training, development and test set (see Section 5). They were released to participants in an instance of the CoNLL-U Plus format[14] called CUPT.[15] As described in more detail by Ramisch et al. (2018), it is a TAB-separated textual format with one token per line and 11 columns: the first 10 correspond to morpho-syntactic information identical to CoNLL-U such as the token's LEMMA and UPOS tags, and the 11th column contains the VMWE annotations in the form of numerical indices and a category label. Appendix B presents some corpus statistics, including the number of annotated VMWEs per category. Virtually all corpora are released

[11] `http://ufal.mff.cuni.cz/udpipe`

[12] Hebrew was excluded due to insufficient quantity of newly annotated data.

[13] Chinese had 17 annotators. They were numbered and assigned corpus sentences so that annotator n shared sentences with annotators n-1 and n+1. The outcomes of all annotators with even numbers were grouped into one cluster, and of those with odd numbers into another cluster, as if they were produced by two pseudo-annotators. For Irish, with only one active annotator, self-agreement was measured between the beginning and the end of the annotation process. For Greek, Polish and Portuguese, a subcorpus was annotated by 2 independent annotators.

[14] `http://universaldependencies.org/ext-format.html`

[15] `http://multiword.sourceforge.net/cupt-format`

Language	DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
tokens ($\times 10^6$)	185	25.6	21.3	915	34.2	12.9	78	281	1,902	307	12.7	2,474	19.8	67.2
sentences ($\times 10^6$)	10	1.04	1.33	34	1.38	0.45	3.6	12.3	159	26	0.48	164	1.39	4.11
tokens/sentence	18.5	24.5	16.0	26.9	24.8	38.5	21.7	22.9	12.0	11.8	26.6	15.1	14.5	16.3

Table 2: Number of tokens, sentences and average tokens/sentence ratio in the raw corpora

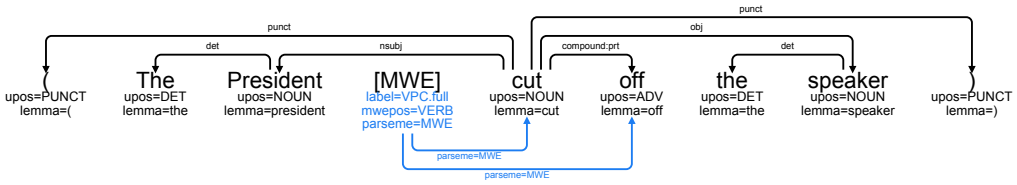

Figure 1: Example of Grew-match visualisation of a MWE annotation.

under various flavours of Creative Commons.[16]

3 Raw Corpora

In addition to the VMWE-annotated data, each language team prepared a large "raw" corpus, i.e., a corpus annotated for morphosyntax but not for VMWEs.[17] Raw corpora, uniformly released in the UD format, were meant for discovering unseen VMWEs. They have very different sizes (cf. Tab. 2) ranging from 12.7 to 2,474 millions of tokens. The genre of the data depends on the language, but efforts were put into making it consistent with the annotated data. The most frequent sources are CoNLL 2017 shared-task data, Wikipedia and newspapers.

For all languages except Italian, the raw corpus was parsed with UDPipe (Straka and Straková, 2017) using models trained on UD treebanks (2.0, 2.4 or 2.5). The Italian corpus was converted into UD from the existing annotated PAISÀ Corpus.[18] To ease their use by participants, each raw corpus was split into smaller files. We checked with a UD tool[19] that in the first 1,000 sentences of each file: (1) each sentence contains the required metadata, (2) the POS and dependency tags comply with the UD 2 tagsets, (3) the syntactic annotation forms a tree.

4 New Tools and Resources

Documentation Up to now, the release of data was coordinated with the organisation of shared tasks. This time, effort has been put into dissociating corpus annotation from shared tasks. Each language team was given a git repository containing development versions of the corpora. We have created a wiki containing instructions for language leaders to prepare data, recruit and train annotators, use common tools to create and manipulate data (e.g. the centralised annotation platform FLAT), etc. This documentation should evolve as the initiative moves towards more frequent releases of the data. We hope that this will allow more flexible resource creation, in accordance with each team's needs and resources. Moreover, extensions and enhancements in the corpora will be integrated into MWE identification tools faster.

Grew-match All along the annotation phase, the latest version of the annotated corpora (on a git repository) was searchable online via the Grew-match querying tool.[20] Grew-match is a generic graph-matching tool which was adapted to take into account the MWE annotations, by adding MWE-specific graph nodes and arcs, as shown in Figure 1: each MWE gives rise to a fake "token" node, heading arcs to all the components of the MWE. Language teams thus used Grew-match to identify potential errors

[16] Except parts of the CoNLL-U data, under other open (French, Polish, Irish) or unknown (Irish) licenses.

[17] The raw corpus for edition 1.2 is available at `http://hdl.handle.net/11234/1-3416` and described at `http://gitlab.com/parseme/corpora/wikis/Raw-corpora-for-the-PARSEME-1.2-shared-task`

[18] `http://www.corpusitaliano.it`

[19] `https://github.com/universalDependencies/tools`

[20] `http://match.grew.fr/` – tab "PARSEME".

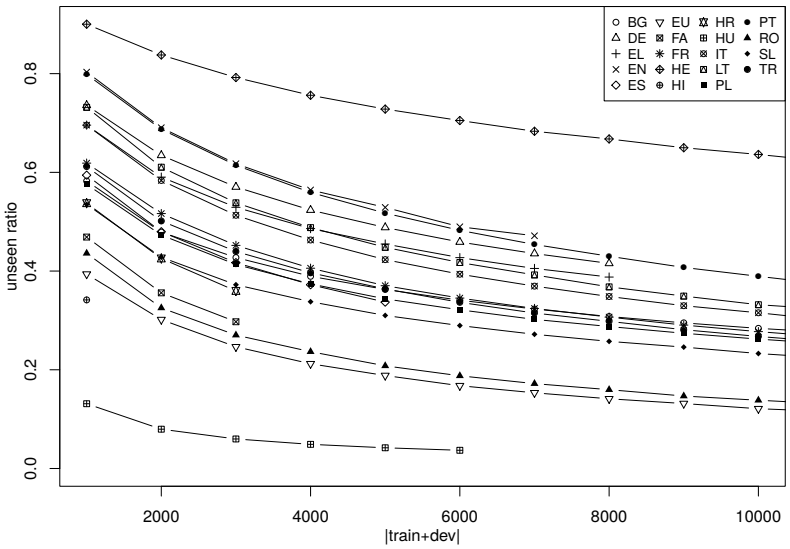

Figure 2: Per-language unseen ratios as a function of train+dev size (data from edition 1.1).

and inconsistencies, e.g., the VMWE in Figure 1 would be retrieved by searching for VMWEs lacking a verbal component (in this case, the MWE annotation is correct whereas the POS of *cut* is incorrect).

Evaluation Tools We adopt the script and metrics developed in edition 1.1 and described in detail by Ramisch et al. (2018). In addition to global and token-based precision (P), recall (R) and F-measure (F1), per language and macro-averaged, we evaluate participating systems on specific VMWE phenomena (e.g. continuous vs. discontinuous) and categories (e.g. VID, IRV, LVC.full). Especially relevant for this edition are the scores on unseen VMWEs, that is, those whose multi-set of lemmas never occur in the training data. In edition 1.1, by training data we meant the train subset only. Recently, we found that this introduced bias from those VMWEs which occurred in dev but not in train: they were still known in the gold data during the system development and tuning. Therefore, in edition 1.2, we redefined an unseen VMWE as a multiset of lemmas annotated in test but not in train+dev. Also differently from edition 1.1, the final macro-averaged and language-specific rankings emphasise results on unseen VMWEs.

5 Corpus Splits

Some datasets in edition 1.1 contained very few unseen VMWEs.[21] Using them as is would lead to statistically unreliable assessment of systems' performance on unseen VMWEs. Thus, we had to design a strategy to re-split the corpora controlling for the distribution of unseen VMWEs. Our two prerequisites were to: (i) ensure a sufficient absolute number of unseen VMWEs for each language (ii) adapt the strategy to the (7 out of 14) languages with no new annotated data compared to previous editions. Hence we could not use the strategy of the WNUT2017 shared task on novel and emerging entity recognition, which would consist in annotating new texts, pre-filtered so as not to contain the VMWEs already present in the existing data (Derczynski et al., 2017). Therefore, we decided to split the whole annotated data for each language by randomly placing sentences in the training (train), development (dev) or test sets.

We considered several splitting methods differing in the parameters that were controlled. Apart from the absolute *number* of unseen VMWEs, the unseen/all VMWE *ratio*, as well as the test/whole corpus size ratio, seemed like desirable parameters of the splitting method. However, these three parameters interact. Figure 2, which plots the average unseen ratio as a function of the train+dev size (in terms of the number of sentences), shows that unseen ratios greatly vary across languages, *even when controlling for train+dev size*. Furthermore, we can see that this ratio depends on the relative size of the train+dev/test sets. So while the unseen ratio may well depend on some traits intrinsic to the language, it clearly depends on other, external, factors (e.g. the chosen text genres and the particular split).

[21] E.g. Romanian, Basque, and Hungarian contain 26, 57, and 62 unseen VMWEs in test w.r.t. train+dev.

Language		DE	EL	EU	FR	GA	HE	HI	IT	PL	PT	RO	SV	TR	ZH
Dev w.r.t.	Nb.	100	100	100	101	100	101	100	101	100	100	100	100	100	100
train	Rate	0.37	0.32	0.19	0.24	0.79	0.61	0.54	0.31	0.23	0.25	0.12	0.37	0.27	0.38
Test w.r.t.	Nb.	301	300	300	300	301	302	300	300	301	300	299	300	300	300
train+dev	Rate	0.37	0.31	0.15	0.22	0.69	0.60	0.45	0.29	0.22	0.24	0.07	0.31	0.26	0.38

Table 3: Number and rate of unseen VMWEs in dev w.r.t. train and in test w.r.t. train+dev.

On the other hand, the unseen VMWE ratio was proved to better (inversely) correlate with MWE identification performance than with the training set size alone (Al Saied et al., 2018). The analysis above dissuaded us from controlling for a "natural" (i.e. close to the average across random splits) unseen ratio. Therefore two options were considered: (1) perform random splits using predetermined proportions for train/dev/test sets and pick a split that best approaches the "natural" unseen ratio for that language, while reaching a sufficient absolute number of unseen VMWEs in the test set; (2) target roughly the same absolute number of unseen VMWEs per language, while the test size and unseen ratio follow from it naturally. Both options restrict the unseen ratio (which still varies depending on the specific split). We preferred the second one because it gives equal weights to each language in system evaluation.

Implemented Splitting Method The splitting method relies on two parameters: the number of unseen VMWEs in test with respect to train+dev, and the number of unseen VMWEs in dev with respect to train. The latter ensures that dev is similar to test, so that systems tuned on dev have similar performances on test. The method strives to find a three-way train/dev/test split satisfying the input specification while preserving the "natural" data distribution (in particular, the unseen/all VMWEs ratios).

The same procedure is applied to split the full data into test and train+dev, and then to split train+dev into train and dev, so only the former is detailed below. The procedure takes as input a set of sentences, a target number of unseen VMWEs u_t, and a number N of random splits:

- We estimate s_t, the size (number of sentences) of the target test set leading to the desired value of u_t. As the average number of unseen VMWEs grows with the size of the test set,[22] we can use binary search to determine s_t.[23] In the course of the search, for a given test size, the average number of unseen VMWEs is estimated based on N random splits.
- For the resulting test size s_t, we compute the average unseen ratio r_t over the same N splits.
- N random splits with test size s_t are performed, and the one that best fits u_t and r_t is selected. More precisely, best fit means here the split, with u unseen and unseen ratio r, that minimises the cost function $c(u, r, u_t, r_t) = \frac{|u_t - u|}{u_t} + |r_t - r|$.

Splitting Results Table 3 shows the statistics of the splits obtained for all languages of the shared task using the above method, with N=100, u_t=300 (in test) and then u_t=100 (in dev). Due to different sizes and characteristics of the individual datasets and languages, the obtained test/train+dev and dev/train unseen ratios vary considerably, the former varying from 0.07 for Romanian to 0.69 for Irish.[24]

6 Systems and Results

Seven teams submitted 9 results to edition 1.2, summarised in Table 4. They use a variety of techniques including recurrent neural networks (ERMI, MultiVitamin, MTLB-STRUCT and TRAVIS), syntax-based candidate extraction and filtering including association measures (HMSid, Seen2Seen), and rule-based joint parsing and MWE identification (FipsCo). The VMWE-annotated corpora are used for model training or fine-tuning, as well as for tuning patterns and filters. Surprisingly, the provided raw corpora

[22] The input dataset is fixed, hence a larger test set means a smaller train set, therefore more unseen VMWEs.

[23] If the input set has T sentences, we iterate using a binary search for the test set size in the $[1, T-1]$ interval. For instance, the first iteration picks $s = \lfloor T/2 \rfloor$, the interval considered next ($[1, s-1]$ or $[s+1, T-1]$) depends on $U(s)$, the average number of unseen VMWEs in N random splits with test set of size s: if the current value is higher than $U(s)$, then the next binary search will operate on $[1, s-1]$, and so on. The final value of s is assigned to s_t.

[24] Romanian's unseen ratio results from sentence pre-selection and leads to outstanding identification results.

System	Architecture	Use of corpora/resources		
		Annotated	**Raw**	**External resources**
ERMI	bidirectional LSTM + CRF	train model	train embed.	–
FipsCo	rule-based joint parsing+identification			VMWE lexicon
HMSid	syntactic patterns, association measures (AMs)	tune patterns and AMs		idiom dataset, FrWac corpus
MTLB-STRUCT	neural language model, fine-tuned for joint parsing+identification	tune BERT		multilingual BERT
MultiVitamin	neural binary ensemble classifier	train classifier		XLM-RoBERTa
Seen2Seen	rule-based extraction + filtering	tune filters		–
Seen2Unseen	+ lexical replacement, translation, AMs			Google Trans., Wiktionary, CoNLL 2017 corpus/embed.
TRAVIS-mono	neural language model, fine-tuned for MWE identification	tune BERT		monolingual BERT
TRAVIS-multi				multilingual BERT

Table 4: Architecture of the systems, and their use of provided and external resources.

System	#Lang	Unseen MWE-based				Global MWE-based				Global Token-based			
		P	R	F1	#	P	R	F1	#	P	R	F1	#
ERMI	14/14	25.3	27.2	26.2	1	64.8	52.9	58.2	2	73.7	54.5	62.6	2
Seen2Seen	14/14	36.5	00.6	01.1	2	76.2	58.6	66.2	1	78.6	57.0	66.1	1
MTLB-STRUCT	14/14	36.2	41.1	38.5	1	71.3	69.1	70.1	1	77.7	70.9	74.1	1
TRAVIS-multi	13/14	28.1	33.3	30.5	2	60.7	57.6	59.1	3	70.4	60.1	64.8	2
TRAVIS-mono	10/14	24.3	28.0	26.0	3	49.5	43.5	46.3	4	55.9	45.0	49.9	4
Seen2Unseen	14/14	16.1	12.0	13.7	4	63.4	62.7	63.0	2	66.3	61.6	63.9	3
FipsCo	3/14	04.3	05.2	05.7	5	11.7	8.8	10.0	5	13.3	8.5	10.4	5
HMSid	1/14	02.0	03.8	02.6	6	04.6	04.9	04.7	6	04.7	04.8	04.8	6
MultiVitamin	7/14	00.1	00.1	00.1	7	00.2	00.1	00.1	7	03.5	01.3	01.9	7

Table 5: Unseen MWE-based (w.r.t. train+dev), global MWE-based, and global token-based Precision (P), Recall (R), F-measure (F1) and F1 ranking (#). Closed track above separator, open track below.

seem to have been used by one system only, for training word embeddings (ERMI). We expected that the teams would use the raw corpus to apply MWE discovery methods such as those described in Constant et al. (2017, Sec. 2), but they may have lacked time to do so. The external resources used include morphological and VMWE lexicons, external raw corpora, translation software, pre-trained non-contextual and contextual word embeddings, notably including pre-trained mono- and multi-lingual BERT.

Table 5 shows the participation of the systems in the two tracks, the number of languages they covered, and their macro-average F1 score ranking across 14 languages.[25] Two system results were submitted to the closed track and 7 to the open track. Four results covered all 14 languages.[26] As this edition focuses on performances on unseen VMWEs, these scores are presented first. [27] In the open track, the best F1 obtained by MTLB-STRUCT (38.53) is by over 10 points higher the corresponding best score in the edition 1.1 (28.46, by SHOMA). These figures are, however, not directly comparable, due to differences in the languages covered in the two editions, the size and quality of the corpora. The closed-track system ERMI achieves promising results, likely thanks to word embeddings trained on the raw corpus.

The global MWE-based F1 scores for all, both seen and unseen, VMWEs exceed 66 and 70 for the closed and open track, respectively, against 54 and 58 in edition 1.1. Like for the unseen score, it remains to be seen how much this significant difference owes to new/enhanced resources, different language sets, and novel system architectures. The second best score across the two tracks is achieved by a closed-track system (Seen2Seen) using non-neural rule-based candidate extraction and filtering. Global token-based

[25]Full results: `http://multiword.sourceforge.net/sharedtaskresults2020/`

[26]Macro-averages are meaningless for systems not covering some languages, for which P=R=F1=0.

[27]When we first published the results, we wrongly considered the unseen in test with respect to train only. Here we provide the results with unseen with respect to train+dev, as explained in Section 4. Results will be updated on the website and in the final versions of system description papers.

System	Unseen MWE-based F1 score										
	DE	**EL**	**EU**	**FR**	**HE**	**HI**	**IT**	**PL**	**PT**	**RO**	**TR**
ERMI	21.98	29.81	26.99	24.40	08.40	39.25	12.71	25.92	28.33	21.28	36.46
MTLB-STRUCT	**49.34**	**42.47**	**34.41**	42.33	**19.59**	**53.11**	20.81	39.94	**35.13**	34.02	43.66
TRAVIS-mono	46.89	7.25	–	**48.01**	–	0.64	**26.16**	**43.44**	–	**40.26**	48.40
TRAVIS-multi	37.25	37.86	30.38	37.27	15.51	34.90	21.48	38.95	–	28.34	41.74
SHOMA (1.1)	18.40	29.67	18.57	44.66	14.42	47.74	11.83	17.67	29.36	17.95	**50.27**
Nb. VMWE (1.2)	3,217	6,470	2,226	4,295	2,030	361	3,178	5,841	5,174	2,036	6,579
Nb. VMWE (1.1)	3,323	1,904	3,323	5,179	1,737	534	3,754	4,637	4,983	5,302	6,635
Nb. unseen (1.2)	301	300	300	300	302	300	300	301	300	299	300
Nb. unseen (1.1)	232	192	57	240	307	214	179	137	141	26	378

Table 6: F1 scores on unseen VMWEs (in train+dev) of the 4 best systems in ed. 1.2, of the best open system in ed. 1.1 (SHOMA), nb. of VMWEs (train+dev), and nb. of unseen VMWEs (train+dev).

F1 scores are often slightly higher than corresponding MWE-based scores. An interesting opposition appears when comparing the global scores with those for unseen VMWEs. In the former, precision is usually higher than recall, whereas in the latter, recall exceeds precision, except for 2 systems.

As macro-averages hide inter-language variability, Table 6 shows unseen F1 scores for 11 languages present in editions 1.1 and 1.2. Results are not comparable across editions due to different corpora, but for languages with similar number of annotated total and unseen VMWEs, some systems reach higher unseen F1 scores than the best 1.1 system SHOMA (e.g. in German, French, and Hindi). However, this is not systematic (see Turkish) and the best scores are not always obtained by the same systems, preventing us from drawing strong conclusions. Performances for Chinese (not shown in Table 6) are surprisingly high, reaching unseen F1=60.19 (TRAVIS-mono). In Chinese, a many VMWEs are syntactically and lexically regular. A simple system with two rules would reach unseen MWE-based F1=27.33.[28]

One finding from the previous shared task editions (Section 5), is that performance for a given language is better explained by the unseen ratio for this language than by the size of the training set. This is even truer for the 1.2 edition, as we could measure a very high negative linear correlation between the highest MWE-based F1 score for a given language and the unseen ratio for that language (Pearson coefficient = -0.90). In contrast, the correlation between the best F1 and the size of the number of annotated VMWEs in the training set is quite poor (Pearson coefficient = 0.23). Appendix C plots these correlations graphically.

7 Conclusions and Future Work

The contributions of the PARSEME shared task 1.2 can be summarised as: (1) the creation and enhancement of VMWE-annotated corpora including three new languages, (2) an evaluation methodology to split the corpora ensuring the representativity of the target phenomenon, and (3) encouraging results hinting at improvements on the identification of unseen VMWEs. In the future, we would like to implement continuous corpus development, with frequent releases independent of shared tasks, so that new languages can join at any time and system developers benefit from latest corpus versions. Additionally, our long-term aim is to increase the coverage of MWE categories, including nominal expressions, adverbials, etc. Finally, we would like to pursue our efforts to design innovative setups for combining (unsupervised) MWE discovery, automatic and manual lexicon creation, and supervised MWE identification.

Acknowledgments

This work was supported by the IC1207 PARSEME COST action and project PARSEME-FR (ANR-14-CERA-0001). Thanks to Maarten van Gompel for his help with FLAT and to the University of Düsseldorf for hosting the server. Thanks to language leaders and annotators (Appendix A) for their hard work.

[28] R1: verbs ending with 入 are single-token VMWEs; R2: pairs of consecutive verbs linked with `mark` and such that the dependant's lemma belongs to a list of 7 lemmas: 到, 为, 出, 在, 成, 至 and 出 are VMWEs.

References

Hazem Al Saied, Marie Candito, and Mathieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth*. Language Science Press.

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.

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1934–1940, Las Palmas.

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.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, September. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240. ACL.

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 Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. 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*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019a. Without lexicons, multiword expression identification will never fly: A position statement. In *MWE-WN 2019*, pages 79–91, Florence, Italy. ACL.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019b. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54, apr.

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.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

A Composition of the Corpus Annotation Teams

DE: Timm Lichte (LL[29]), Rafael Ehren; **EL**: Voula Giouli (LL), Vassiliki Foufi, Aggeliki Fotopoulou, Stella Markantonatou, Stella Papadelli, Sevasti Louizou **EU**: Uxoa Iñurrieta (LL), Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez, Antton Gurrutxaga, Larraitz Uria, Ruben Urizar; **FR**: Marie Candito (LL), Matthieu Constant, Bruno Guillaume, Carlos Ramisch, Caroline Pasquer, Yannick Parmentier, Jean-Yves Antoine, Agata Savary; **GA**: Abigail Walsh (LL), Jennifer Foster, Teresa Lynn; **HE**: Chaya Liebeskind (LL), Hevi Elyovich, Yaakov Ha-Cohen Kerner, Ruth Malka; **HI**: Archna Bhatia (LL), Ashwini Vaidya (LL), Kanishka Jain, Vandana Puri, Shraddha Ratori, Vishakha Shukla, Shubham Srivastava; **IT**: Johanna Monti (LL), Carola Carlino, Valeria Caruso, Maria Pia di Buono, Antonio Pascucci, Annalisa Raffone, Anna Riccio, Federico Sangati, Giulia Speranza; **PL**: Agata Savary (LL), Jakub Waszczuk (LL), Emilia Palka-Binkiewicz; **PT**: Carlos Ramisch (LL), Renata Ramisch (LL), Silvio Ricardo Cordeiro, Helena de Medeiros Caseli, Isaac Miranda, Alexandre Rademaker, Oto Vale, Aline Villavicencio, Gabriela Wick Pedro, Rodrigo Wilkens, Leonardo Zilio; **RO**: Verginica Barbu Mititelu (LL), Mihaela Ionescu, Mihaela Onofrei, Monica-Mihaela Rizea; **SV**: Sara Stymne (LL), Elsa Erenmalm, Gustav Finnveden, Bernadeta Griciūtė, Ellinor Lindqvist, Eva Pettersson; **TR**: Tunga Güngör (LL), Zeynep Yirmibeşoğlu, Gozde Berk, Berna Erden; **ZH**: Menghan Jiang (LL), Hongzhi Xu (LL), Jia Chen, Xiaomin Ge, Fangyuan Hu, Sha Hu, Minli Li, Siyuan Liu, Zhenzhen Qin, Ruilong Sun, Chengwen Wang, Huangyang Xiao, Peiyi Yan, Tsy Yih, Ke Yu, Songping Yu, Si Zeng, Yongchen Zhang, Yun Zhao.

B Statistics of the Corpora

Lang-part	Sent.	Tokens	Avg. length	VMWE	VID	IRV	LVC full	LVC cause	VPC full	VPC semi	IAV	MVC	LS ICV
DE-train	6568	126830	19.3	2950	1039	249	212	24	1286	140	0	0	0
DE-dev	602	11756	19.5	267	95	14	26	2	122	8	0	0	0
DE-test	1826	34976	19.1	824	303	59	73	7	336	46	0	0	0
DE-Total	8996	173562	19.2	4041	1437	322	311	33	1744	194	0	0	0
EL-train	17733	479679	27	6155	1933	0	3982	101	96	0	0	43	0
EL-dev	909	23911	26.3	315	98	0	203	5	4	0	0	5	0
EL-test	2805	75442	26.8	974	323	0	612	19	17	0	0	3	0
EL-Total	21447	579032	26.9	7444	2354	0	4797	125	117	0	0	51	0
EU-train	4440	61867	13.9	1690	347	0	1261	82	0	0	0	0	0
EU-dev	1418	20509	14.4	536	127	0	383	26	0	0	0	0	0
EU-test	5300	75431	14.2	2020	406	0	1508	106	0	0	0	0	0
EU-Total	11158	157807	14.1	4246	880	0	3152	214	0	0	0	0	0
FR-train	14377	360070	25	3870	1494	1037	1253	70	0	0	0	16	0
FR-dev	1573	39502	25.1	425	157	117	144	5	0	0	0	2	0
FR-test	5011	126420	25.2	1359	505	347	481	22	0	0	0	4	0
FR-Total	20961	525992	25	5654	2156	1501	1878	97	0	0	0	22	0
GA-train	257	6242	24.2	100	14	0	35	23	2	2	24	0	0
GA-dev	322	7020	21.8	126	22	0	29	22	6	5	42	0	0
GA-test	1121	25954	23.1	436	69	6	137	74	20	13	117	0	0
GA-Total	1700	39216	23	662	105	6	201	119	28	20	183	0	0
HE-train	14152	286262	20.2	1864	825	0	765	166	108	0	0	0	0
HE-dev	1254	25392	20.2	166	64	0	80	13	9	0	0	0	0
HE-test	3794	76827	20.2	503	219	0	204	44	36	0	0	0	0
HE-Total	19200	388481	20.2	2533	1108	0	1049	223	153	0	0	0	0
HI-train	282	5764	20.4	175	11	0	109	3	0	0	0	52	0
HI-dev	289	6272	21.7	186	11	0	126	0	0	0	0	49	0
HI-test	1113	23394	21	673	39	0	406	23	0	0	0	205	0
HI-Total	1684	35430	21	1034	61	0	641	26	0	0	0	306	0
IT-train	10641	292065	27.4	2854	999	783	502	112	74	2	343	19	20
IT-dev	1202	32652	27.1	324	109	81	52	18	11	0	44	4	5
IT-test	3885	106072	27.3	1032	376	280	180	44	20	0	110	10	12
IT-Total	15728	430789	27.3	4210	1484	1144	734	174	105	2	497	33	37
PL-train	17731	298437	16.8	5398	629	2723	1807	239	0	0	0	0	0
PL-dev	1425	23950	16.8	443	49	219	162	13	0	0	0	0	0

[29]LL stands for language leader.

Lang-part	Sent.	Tokens	Avg. length	VMWE	VID	IRV	LVC full	LVC cause	VPC full	VPC semi	IAV	MVC	LS ICV
PL-test	4391	73753	16.7	1345	148	687	451	59	0	0	0	0	0
PL-Total	23547	396140	16.8	7186	826	3629	2420	311	0	0	0	0	0
PT-train	23905	542497	22.6	4777	945	763	2960	98	0	0	0	11	0
PT-dev	1976	43676	22.1	397	80	73	236	6	0	0	0	2	0
PT-test	6236	142377	22.8	1263	281	191	763	23	0	0	0	5	0
PT-Total	32117	728550	22.6	6437	1306	1027	3959	127	0	0	0	18	0
RO-train	10920	195718	17.9	1218	304	771	108	35	0	0	0	0	0
RO-dev	7714	134340	17.4	818	228	504	56	30	0	0	0	0	0
RO-test	38069	685566	18	4135	1114	2552	352	117	0	0	0	0	0
RO-Total	56703	1015624	17.9	6171	1646	3827	516	182	0	0	0	0	0
SV-train	1605	24970	15.5	752	105	41	95	6	345	160	0	0	0
SV-dev	596	8889	14.9	270	40	24	42	0	108	56	0	0	0
SV-test	2103	31623	15	969	146	50	142	5	418	208	0	0	0
SV-Total	4304	65482	15.2	1991	291	115	279	11	871	424	0	0	0
TR-train	17945	267503	14.9	6212	3351	0	2858	0	0	0	0	3	0
TR-dev	1062	15935	15	367	187	0	179	0	0	0	0	1	0
TR-test	3304	48791	14.7	1151	604	0	546	0	0	0	0	1	0
TR-Total	22311	332229	14.8	7730	4142	0	3583	0	0	0	0	5	0
ZH-train	35326	575590	16.2	8113	676	0	927	148	0	3156	0	3206	0
ZH-dev	1141	18258	16	265	18	0	33	6	0	108	0	100	0
ZH-test	3462	55728	16	786	63	0	94	13	0	300	0	316	0
ZH-Total	39929	649576	16.2	9164	757	0	1054	167	0	3564	0	3622	0
Total	279785	5517910	19.7	68503	18553	11571	24574	1809	3018	4204	680	4057	37

C Correlation of Performance and Unseen Ratio/Training Set Size

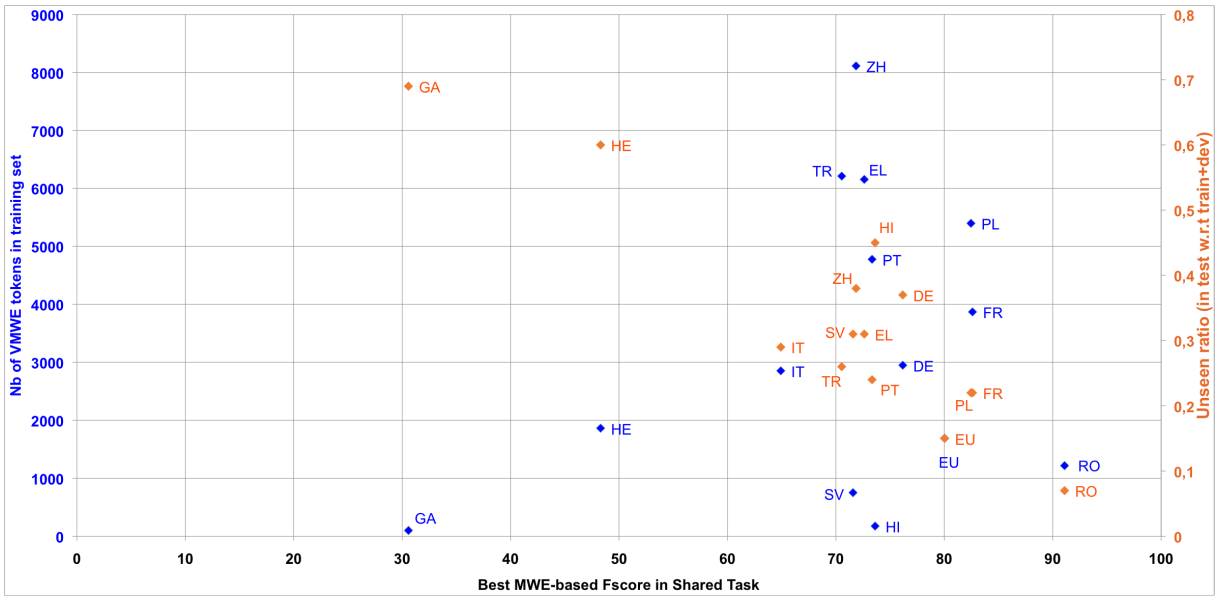

Figure 3: Relation between the performance of each language and its unseen ratio (red) and number of VMWEs tokens in the training set (blue). X axis: best MWE-based F1 score. Blue Y axis: Number of VMWEs in training set. Red Y axis: Unseen ratio.