PARSEME Corpus Release 1.3

Agata Savary
Paris-Saclay Univ,
CNRS, LISN, France
agata.savary@universite-paris-saclay.fr

Cherifa Ben Khelil
Univ of Tours, LIFAT, France
cherifa.bk@gmail.com

Carlos Ramisch
Aix Marseille U, CNRS,
LIS, France
carlos.ramisch@lis-lab.fr

Voula Giouli
ILSP
ATHENA Res Centre, Greece

Verginica Barbu Mititelu
RACAI
Romanian Academy

Najet Hadj Mohamed
U Tours, LIFAT, France
U Sfax, MIRACL, Tunisia

Cvetana Krstev
Univ of Belgrade, Serbia

Chaya Liebeskind
Jerusalem College of Technology
Israel

Hongzhi Xu
SISU, ICSA,
Shanghai, China

Menghan Jiang
MSU-BIT U Shenzhen China

Sara Stymne
Uppsala Univ, Sweden

Tunga Güngör
Boğaziçi Univ, Turkey

Thomas Pickard
University of Sheffield, UK

Bruno Guillaume
Univ de Lorraine, CNRS,
Inria, LORIA, France

Archna Bhatia
IHMC, USA

Alexandra Butler
Univ California, USA

Marie Candito
Univ Paris Diderot, France

Apolonija Gantar
Univ Ljubljana, Slovenia

Uxoa Iñurrieta
Univ Basque Country, Spain

Albert Gatt
Utrecht Univ, Netherlands

Jolanta Kovalevskaite
VMU, Lithuania

Simon Krek
Jožef Stefan Inst, Slovenia

Timm Lichte
Univ Tübingen, Germany

Nikola Ljubešić
Jožef Stefan Inst, Slovenia

Johanna Monti
UNIOR NLP Res Group
U Naples L'Orientale, Italy

Carla Parra Escartín
Dublin City U, Ireland

Mehrnoush Shamsfard
Shahid Beheshti Univ, Iran

Ivelina Stoyanova
IBL, BAS, Bulgaria

Veronika Vincze
ELKH-SZTE
Research Group on AI, Hungary

Abigail Walsh
ADAPT Centre, Dublin City
Univ, Ireland

Abstract

We present version 1.3 of the PARSEME multilingual corpus annotated with verbal multiword expressions. Since the previous version, new languages have joined the undertaking of creating such a resource, some of the already existing corpora have been enriched with new annotated texts, while others have been enhanced in various ways. The PARSEME multilingual corpus represents 26 languages now. All monolingual corpora therein use Universal Dependencies v.2 tagset. They are (re-)split observing the PARSEME v.1.2 standard, which puts impact on unseen VMWEs. With the current iteration, the corpus release process has been detached from shared tasks; instead, a process for continuous improvement and systematic releases has been introduced.

1 Introduction

The difficulty in automatically identifying multiword expressions (MWEs) in texts has been acknowledged for a while (Sag et al., 2002; Baldwin and Kim, 2010), and confirmed through results of experiments, many of which conducted as part of shared tasks (Schneider et al., 2016; Savary et al., 2017; Ramisch et al., 2018, 2020). MWEs, especially verbal ones (VMWEs), have been the focus of the PARSEME community since the homony-

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 24–35

mous COST Action took place[1] and are now paid further attention, in correlation with syntactic annotation and language typology, within the UniDive COST Action[2].

Training, tuning, and testing the systems that are able to identify VMWEs in texts need corpora annotated with such expressions. Within PARSEME, guidelines for annotating VMWEs were created and then improved with feedback provided during annotation. When we compare the differences between v. 1.0 of the guidelines[3] and their v. 1.1[4], we notice that the latter came with a refined VMWEs typology and an enhanced decision tree ensuring the consistent treatment of the phenomenon in a multilingual environment.

The guidelines contain the following types[5] of VMWEs, established with respect to their pervasiveness in the languages under study.
Universal types include: (i) VID (verbal idiom) e.g. (de) ***schwarz fahren*** (lit. 'black drive') 'take a ride without a ticket', (ii) LVC (light verb construction), which has two subtypes: LVC.full, e.g. (hr, sr) ***držati govor*** (lit. 'hold a speech') 'give a talk' and LVC.cause, e.g. (ro) ***da bătăi de cap*** (lit. 'give strikes of head') 'give a hard time'.
Quasi-universal types contain: (i) IRV (inherently reflexive verbs), e.g. (pt) ***se queixar*** 'complain', (ii) VPC (verb-particle construction), with two subtypes: VPC.full, e.g. (en) ***do in*** and VPC.semi, e.g. (en) ***eat up***, (iii) MVC (multi-verb construction), e.g. (fr) ***laisser tomber*** (lit. 'let fall') 'give up'.
Language-specific types - so far, only Italian has defined such a type: ICV (inherently clitic verb): (it) ***smetterla*** (lit. 'quit it') 'knock it off'.
Experimental category – IAV (inherently adpositional verbs), e.g. (es) ***entender de*** *algo* (lit. 'understand of something') 'know about something' – is annotated optionally. Whenever language-specific characteristics demand it, the decision trees are adjusted to reflect those characteristics, as in the case of Italian or Hindi.

The initiative of collecting and annotating corpora following common guidelines was initially joined by 18 language teams. With each new edition of the corpus, some teams remained active, some others were on standby and some new teams joined. In total, until edition 1.2, corpora for 26 were created but not unified within one single edition.

With this new release (v.1.3) which is the topic of this paper, our objectives are: (i) to release all past 26 languages[6] in a unified format, i.e. morphosyntactic annotation in Universal Dependencies[7] (UD) (Nivre et al., 2020) format, (ii) to detach the corpus releases from shared tasks, and (iii) to define a process of continuous improvement and systematic releasing (following the UD model).

This describes the novelties concerning the annotated data (Sec. 2–4), their underlying morphosyntactic annotation layers (Sec. 5), and their split (Sec. 6). Then, the statistics of the resulting corpus are provided (Sec. 7). We also describe recent developments of the technical infrastructure at the service of the corpus development (Sec. 5–9). We provide results of two VMWE identifiers trained on the new release, which establishes new state of the art for many languages (Sec. 10). We finally conclude and evoke perspectives for future work (Sec. 11). The corpus is available for download at `http://hdl.handle.net/11372/LRT-5124`.

2 New languages

We have two new languages on board: Arabic and Serbian.

The previous dataset for **Arabic** was created by Hawwari in PARSEME 1.1 (Ramisch et al., 2018). However, this corpus has never been published under an open license, being restricted to the Shared Task participants. The Arabic corpus in PARSEME 1.3 is a new corpus created from scratch. More than 4,700 VMWEs have been annotated in about 7,500 sentences taken from the UD corpus Prague Arabic Dependency Treebank (PADT) (Hajic et al., 2004), containing newspaper articles. This new annotated corpus is already available in the PARSEME repository under the CC-BY v4 license.

The **Serbian** language was not represented in the previous versions of the PARSEME corpus. The

[1] https://typo.uni-konstanz.de/parseme/
[2] https://unidive.lisn.upsaclay.fr/
[3] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.0/?page=home
[4] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/?page=home
[5] For their definition and examples in various languages, please see the guidelines: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/?page=home
[6] The 26 languages and their corresponding language codes are: Arabic (ar), Bulgarian (bg), Czech (cs), German (de), Greek (el), English (en), Spanish (es), Basque (eu), Farsi (fa), French (fr), Irish (ga), Hebrew (he), Croatian (hr), Hungarian (hu), Hindi (hi), Italian (it), Lithuanian (lt), Maltese (mt), Polish (pl), Portuguese (pt), Romanian (ro), Slovene (sl), Swedish (sv), Serbian (sr), Turkish (tr), Chinese (zh).
[7] universaldependencies.org

first step in preparing the Serbian PARSEME corpus consisted of the preparation of the large set of examples required for the guidelines.[8] Through this work, it became clear that the types of VMWEs to be encoded in Serbian texts were: LVC (full and cause), VID, and IRV. The Serbian corpus in PARSEME 1.3 consists of 3,586 sentences of newspaper texts covering mostly daily politics, and a small part dealing with fashion. The morphosyntactic annotation of texts was done using UDPipe (Straka, 2018). More than 1,300 VMWEs (approx. 640 different types) were annotated in it by one annotator. For the next edition of the corpus, we will try to recruit at least one more annotator for the same text.

3 Enlarged corpora

Three of the languages already present in previous editions were further enhanced with new annotated data: Greek, Swedish, and Chinese.

In the first edition of the PARSEME corpus, the **Greek** (EL) dataset was rather small and we have been committed since to adding new data in view of ultimately providing a corpus of adequate size. The new dataset comprises newswire texts (c. 26K sentences) also from sources that are characterized as bearing an informal register, lifestyle magazines, and newspapers, in order to account for new types of VMWEs. Only a fraction of the Greek dataset bears manual annotations at the lemma, POS, and dependency levels, namely the one originating from the UD initiative; the rest was completed automatically using UDpipe. VMWEs annotation was performed by two annotators; during the annotation process, extensive discussions were aimed at manually correcting common errors and avoiding inconsistencies.

The **Swedish** data set is expanded in comparison to PARSEME release 1.2. The Swedish annotations now cover the complete UD Swedish-Talbanken treebank, increasing the total size from 4,304 to 6,026 sentences. The Swedish corpus includes the manual morphosyntactic annotations from UD, now updated from version 2.5 to version 2.11. The new annotations were done in connection to the PARSEME 1.2 annotation campaign, by two trained annotators. As an extra decision support, the annotators were given access to the report from the consistency check for Swedish PARSEME 1.2, which both annotators reported as being very useful.

In this edition, the **Chinese** data includes 9,000 newly annotated sentences from the CoNLL 2017 Shared Task (Zeman et al., 2017). The columns were updated with the new UDPipe model to make the data consistent with the standard of UD 2.11. All the sentences were double annotated and the decisions were made by a trained linguistics student for the disagreed ones.

4 Enhancements of the existing data

The **Croatian** PARSEME annotations were, long overdue, transferred to the source hr500k dataset (Ljubešić et al., 2016)[9]. Sentences in hr500k that were annotated with PARSEME annotations are those that are annotated with gold UD linguistic annotation. With the PARSEME annotation transfer into hr500k, we enabled the gold UD annotations, which are being continuously improved, to be transferred back to the Croatian PARSEME dataset. The percentage of sentences that went through some change is rather staggering: from 3828 sentences, only 374 (9.8%) stayed identical as in PARSEME version 1.1, while the remaining sentences went through some sort of improvement in the linguistic annotation, either UD error correction or UD standard enhancement.

The **Romanian** corpus contained annotation of the VID, LVC.full, LVC.cause, and IRV types of MWEs in its previous releases. The new version contains annotation of IAVs, a type that was experimental in the Shared Task 1.2. Working with this type raised a few challenges, given that the class of such verbs seems to be heterogeneous with respect to the presence of the preposition in various syntactic structures in which the verb occurs. On the other hand, the test for identifying this type has proven insufficient in the case of some verbs, which shows the need for revisiting it. Given the frequency of this type in the corpus (a third of all VMWEs in the Romanian corpus is represented by IAVs, see Table 2), we consider it important to decide upon a common way of treating it in various languages.

In some languages, manual revision of previous annotations was performed. Thus, in **English**, the 1.1 version of the corpus went through a thorough process of consistency checks (Savary et al., 2018). In **Polish**, a number of controversial or inconsistent annotations were spotted by a new team member.

[8] Andjela Antić and Isidora Jaknić, master students at the University of Belgrade, helped in this task.

[9] `http://hdl.handle.net/11356/1183`

Grew-Match was also used to identify potential errors. Revealed errors were manually fixed. In the **Irish** corpus, a controversial category was removed (IRV), with MWEs of this type re-categorised as IAV or VID. Morphosyntactic annotations were also updated to be consistent with UD v2.11.

The **Turkish** corpus was improved in its morphosyntactic annotations. It was manually reviewed by one annotator and the incorrect annotations from the previous release were corrected. This resulted in changes in the form, lemma, UPOS, and XPOS fields in, respectively, 15, 2480, 1250, and 1266 tokens. The number of morphological features changed in the features field is 6451.

For two languages, **Czech** and **Maltese**, PARSEME corpora were released in version 1.0 only. The 1.0-to-1.1 upgrade of the PARSEME annotation guidelines[10] involved a few major changes, including a redesigned set of VMWE categories. Thus, upgrading 1.0 corpora to version 1.1 requires some manual intervention. Further upgrades to versions 1.2 and 1.3 were minor and mostly automatically applicable. For the present release, we could achieve a partial upgrade from version 1.0 to 1.3 in Czech and Maltese. Future work includes manual annotation of the LVC.cause category, which emerged in v 1.1.

5 Compatibility with Universal Dependencies

Syntactic and semantic properties of MWEs are deeply intertwined.[11] Therefore, the PARSEME corpus has, since its beginnings, been released with annotations for both VMWEs and morphosyntax for most languages. The morphosyntactic annotations have not been produced by PARSEME annotators but rather extracted from existing treebanks or generated by parsers.

To this end, we have been increasingly relying on the UD framework (de Marneffe et al., 2021), treebank collection (Nivre et al., 2020) and UDPipe parser (Straka, 2018), as PARSEME largely shares UD's objectives and principles of universality and diversity. Since edition 1.1, the PARSEME corpus uses the `.cupt` format, which extends the UD' CoNLL-U format with a VMWE annotation layer.[12] Since edition 1.2, we have strongly advocated compatibility with UD version 2.

This objective has been finally achieved in the current 1.3 edition. In 11 languages, we have at least partly manual morphosyntactic annotations. When those stem from UD treebanks, we synchronised them with the most recent UD release (2.11 from November 2022)[13]. In 16 languages, at least part of the morphosyntactic data had been automatically generated and we updated them using the most recent UDPipe models (mostly v 2.10).[14] Whenever several models per language existed, tagging/parsing performances and the genre of the training corpus were used as choice criteria.

As a result, all 26 language corpora now use the UD-2 tagsets (most often in the 2.11 version) for POS, morphological features, and dependency relations.[15] The README files were updated with details of the above updates and a change log now documents the history of releases.

In the future, the procedure for synchronising morphosyntactic annotations with recent UD releases or updating them with UDPipe should be made fully automatic. In the long run, we plan gradual convergence with UD, so as to possibly integrate the PARSEME annotations into UD treebanks (Savary et al., 2023).

6 Corpus re-split

The PARSEME Shared Task edition 1.2 involved dividing the annotated corpora provided by the task organizers into three subsets: training, development, and test (train/dev/test). The training data is used to train the MWE identification systems, the development data is used to perform model selection and fine-tuning, and the test data is used to evaluate the performance of the final models. Since new languages were added and others updated, we decided to follow the 1.2 standard (Ramisch et al., 2020) to re-split the annotated corpus for each language participating in the 1.3 release. This splitting method is based on two key parameters: the number of unseen VMWEs in the test data compared to the combined train and dev data, and the number of unseen VMWEs in the dev data compared to the

[10] `https://parsemefr.lis-lab.fr/parseme-st-guidelines/`

[11] In particular, PARSEME approximates semantic non-compositionality of MWE by their lexical and morphosyntactic inflexibility.

[12] `.cupt` is an instantiation of the CoNLL-U Plus Format.

[13] Exceptions are: (i) Czech, English, Polish, and Basque, where tracing PARSEME sentences to UD treebanks should be simplified, (ii) Italian, where the source treebank is not part of UD and did not evolve.

[14] `https://ufal.mff.cuni.cz/udpipe/2/models`

[15] Maltese lacks annotations for morphological features.

train data. The latter ensures that the dev data is similar to the test data, thereby making it possible for systems that are tuned on the dev data to perform well on the test data. Just as in the Shared Task 1.2, we set the number of unseen VMWEs in the test to 300 and the number of unseen VMWEs in the dev to 100. This configuration has been established to ensure a balanced split that meets the input specifications while preserving the natural distribution of the data, particularly the ratio of unseen to all VMWEs. This particular attention paid to unseen VMWE is motivated by the observation from Shared Task 1.1 that the performances of the VMWE identification systems correlate weakly with the size of the training data but strongly with the proportion of unseen VMWE in the test data. The statistics for the train/dev/test splits across 26 languages can be found in Table 2.

7 Statistics of the corpus

Table 2 presents the corpus statistics, including the number of annotated VMWEs per category. In total, the corpus amounts to over 9 million tokens in over 455,000 sentences, with an average of about 20 tokens per sentence.

Over 127,000 VMWEs are annotated across all 26 languages. The most frequent categories are LVC.full, IRV and VID. The universality (understood as existence in all languages under study) is confirmed for VIDs and LVC.full. LVC.cause, deemed universal, is not annotated in Czech and Turkish. In Czech this is due to the fact that the corpus development was on standby since edition 1.0 in which the LVC.cause category was not defined (cf. Sec. 4). In Turkish we might face a language-specific understanding of the guidelines.

The (quasi-universal) IRV category is present in all Slavic and Romance languages of the collection. Among Germanic languages, IRVs are present in German and Swedish but not in English. VPC.full is a pervasive category in Hungarian and in all 3 Germanic languages. It also occurs in Arabic, Greek, Hebrew, Irish, and Italian. VPC.semi is the dominating category in Chinese and is observed in Germanic languages, Hungarian, Irish, and Italian. IAVs are present in some languages and not others – this is not due to the nature of the language but rather to the fact that this category is considered experimental and has been annotated optionally. MVCs are pervasive in Chinese and in Hindi. Their high frequency in Spanish is probably due to a language-specific understanding of the guidelines.[16] Finally, LS.ICV is an Italian-specific category and obviously occurs in this language only.

All corpora are currently being released under various flavors of the Creative Commons license. Their publication via the LINDAT/CLARIN platform is upcoming.

8 Annotation guidelines

One important aspect of the PARSEME guidelines is the database of examples in multiple languages. Currently, the guidelines feature 232 example identifiers, each covering up to 28 languages. However, not all languages have examples for all example identifiers: we have a total of 1,980 examples, whereas, in theory, we could include up to $232 \times 28 = 6,496$ examples. In edition 1.2, the guidelines contained 1,801 examples; the newly added examples concern mostly Serbian and Arabic, i.e. the languages for which new corpora have been created for this release. Figure 1 shows a histogram with the number of examples per language, ranging from 188 for Spanish to only 1 example for Turkish, Hebrew, and Lithuanian.[17]

The examples in the guidelines are complex, including their form in the original language, lexicalised components in bold, literal, and idiomatic translations, as well as explanations, comments, negative counter-examples, etc. Their addition by language experts is a time-consuming and error-prone process that required much energy. One of the latest improvements on the PARSEME guidelines is a system for online example editing. The original XML language used to edit the examples on a shared online spreadsheet was replaced by an online editing system illustrated in Figure 2. We expect that this system will allow for a much quicker and more autonomous editing of examples by language teams.

9 Versioning, documenting and querrying

In order to help the maintenance of the different corpora, a new infrastructure was set up. All existing corpora, gathered from different previous releases were put in the same GitLab group[18], with each language having its own repository. Now, all

[16] The MVC category in Spanish seems to be used to signal compositional modal verb constructions.

[17] Statistics based on a dump of the examples database on September 14, 2022.

[18] https://gitlab.com/parseme

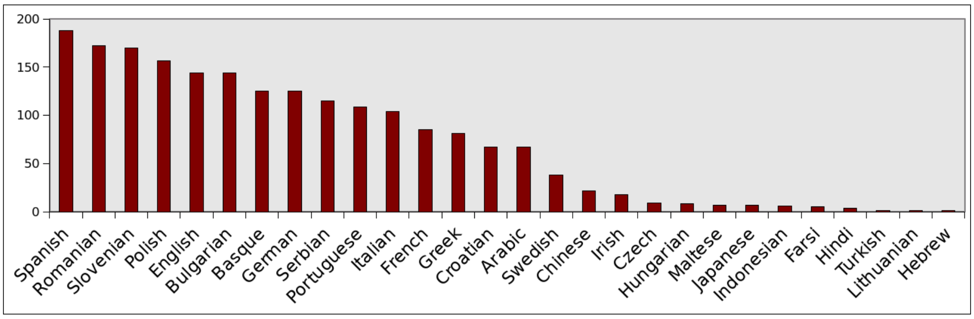

Figure 1: Number of examples per language in PARSEME 1.3 guidelines.

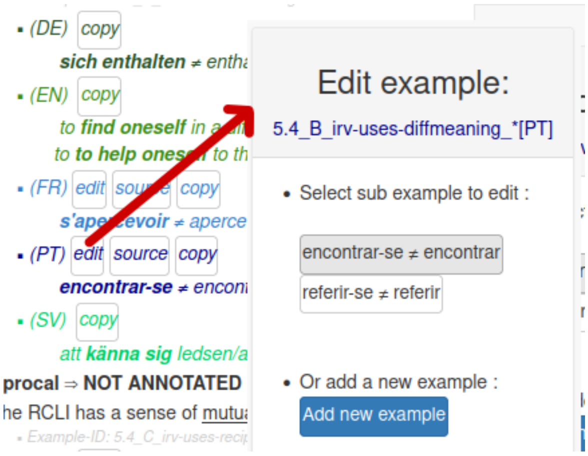

Figure 2: Screenshot of example editing GUI.

new updates on treebanks and new data are stored in this unique place. A rich collection of Wiki pages, available from the same GitLab space, gathers rich documentation of the PARSEME corpora and shared task initiatives, the corresponding tools and procedures, etc.

The Grew-match (Guillaume, 2021) tool has a new instance[19] which gives access to the PARSEME corpora. With this tool, it is possible to make graph-based queries to observe the annotated data; both PARSEME annotations and the underlying UD annotations can be used in queries. Data from each release is available. Moreover, thanks to a continuous integration system, data synchronized with the current development state of each of the 26 corpora (i.e. the data available on the `master` branch of each GitLab repository) can be accessed in Grew-match and the corresponding consistency checks web page is updated automatically when data changes[20].

As an example of Grew-match usage, a simple request[21] can be used to observe what verb lemmas are used in LVC.full annotation in a given corpus (the English one in the example).

10 System results

We began training two state-of-the-art systems, namely Seen2Seen (Pasquer et al., 2020) and MTLB-STRUCT (Taslimipoor et al., 2020), on each corpus of release 1.3. Ranked first in the PARSEME Shared Task edition 1.2 closed track (as far as the global MWE-based F-measure is concerned), Seen2Seen reads all annotated VMWEs in the train and then extracts from the test all candidate occurrences of the same multi-sets of lemmas. The system subsequently runs these candidates through a sequence of morpho-syntactic filters. In total, 8 filters are defined, and Seen2Seen chooses which filter to activate for each language during the training phase based on its performance on the dev corpus. MTLB-STRUCT is a semi-supervised system based on pre-trained BERT models that offers two learning approaches, single-task (where only VMWE annotations are used) or multi-task (where VMWE tags and dependency parse trees are learned jointly), to achieve semi-supervised training. This system has the best global MWE-based F-measure in the PARSEME Shared Task edition 1.2 open track and demonstrated the best performance for detecting unseen VMWEs. The training and evaluation process for MTLB-STRUCT has been completed only for the multi-task version of MTLB-STRUCT, and we report on this version only. The training of the single-task version is still ongoing.

Table 1 provides a comparison of the performance of Seen2Seen and of the multi-task version fo MTLB-STRUCT in identifying VMWEs, including their precision, recall, and F-measure

[19] http://parseme.grew.fr

[20] All the links to these services are available in the page: https://gitlab.com/parseme/corpora/-/wikis/home

[21] http://parseme.grew.fr/?custom=63edd82034bea

scores across 14 languages of the Shared Task edition 1.2 and 26 languages of the new release 1.3. For Seen2Seen, the F-score significantly increased in edition 1.3 for Basque, Hebrew, Hindi, and Swedish. In the case of Basque and Hindi, where no new VMWE annotations were added, this enhancement is certainly due to re-annotating the corpora with a recent version of UDPipe, which must have enhanced the quality of lemmas, used by Seen2Seen to extract VMWE candidates. In Swedish, the corpus size significantly grew, while in Hebrew its quality improved with consistency checks.

For MTLB-STRUCT, the evaluation of the release 1.3 models for Irish, Croatian, Hungarian and Romanian could not be performed for technical reasons. Among the other 10 languages covered both in release 1.2 and 1.3, the increase of the global F-measure is the most significant in Swedish. Also Basque, French, Portuguese and Turkish benefit from the data enhancements. For other languages, the F-measure is lower than in version 1.2, likely due to switching to the multi-task version of the model.

The primary focus of Figure 3 is to showcase how the F-score changes as the number of VMWE tokens in the training corpus varies between releases 1.2 and 1.3. By analyzing the F-scores of different languages, we can observe the effect of the number of VMWE tokens in the training corpus on the performance of the Seen2Seen and MTLB-STRUCT systems. For instance, the increase of the Swedish (SV) and Basque (EU) datasets brought about a higher F-score. Conversely, the F-score for Chinese (ZH) significantly decreased despite the increase in the number of VMWE annotations. This might be attributed to the increased number of unseen VMWEs in the larger corpus. Interestingly, the Turkish dataset decreased in edition 1.3 but the global F-score for both systems increase, which might stem from the higher quality of the 1.3 release data. For Seen2Seen, a large increase of the dataset brings a significant decrease of the F-score, which might indicate a biased nature of the 1.2 release, balanced in version 1.3.

Note that we restrict our comparison to edition 1.2. It would be less meaningful to compare the scores of editions 1.0 and 1.1 with the current version since the splitting methods used in those editions did not prioritize unseen VMWEs. But, even restricted to releases 1.2 and 1.3, the comparison may not be fully reliable, since: (i) each corpus was re-split into train, dev and test sets, i.e. the systems are not trained and evaluated with the same data partitions, (ii) only teh multi-task version of MTLB-STRUCT is examined for release 1.3.

11 Future work

This paper summarises the first release of the PARSEME corpora out of the context of a shared task. This fourth release (v.1.3) is the first one to cover the union of all the languages included in the previous three releases. Moreover, 2 new languages were included, a significant amount of additional data was added for 3 languages, and annotations for many languages were enhanced in various ways.

The future of the PARSEME corpus collection relies on the interests and availability of its volunteer contributors for each language. From the infrastructure perspective, we would like to consolidate the release methodology so that future yearly releases can smoothly integrate and make available the upgrades performed throughout the year by language teams. This includes further automation of procedures, in the spirit of CI/CD[22], including updates of the UD morpho-syntactic annotations, validating the file formats and MWE annotations, and checking the `README.md` documentation.

Another important goal of PARSEME is the extension of its guidelines to (a) non-verbal MWEs, (b) verbal MWEs not covered in the current guidelines, and (c) improved cross-lingual account of phenomena that are currently biased by the set of languages covered in the corpora.

Finally, we envisage synergies with the UD community so that the MWE layer and the morpho-syntactic annotations become gradually even more compatible. The challenges to achieving this goal include reaching compatible tokenisation decisions, unified terminology, reduction of redundancy (e.g. MWEs annotated as subrelations of syntactic dependencies), and syntactic connectiveness of annotated MWEs.

Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01). We thank Quentin Barrouyer and Baptiste Souche for their contribution to the editable annotation guidelines.

[22] Continuous integration and continuous deliver are concepts stemming from the domain of software engineering.

Lang	Seen2Seen						MTLB-STRUCT					
	Shared Task 1.2			*Release 1.3*			*Shared Task 1.2*			*Release 1.3*		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AR				58.33	45.29	50.99				59.54	61.47	60.49
BG				61.69	70.4	65.76				72.53	75.31	73.89
CS				71.54	77.02	74.18				84.99	83.56	84.27
DE	86.21	57.65	69.09	82.87	62.74	71.41	77.11	75.24	76.17	72.58	73.35	72.96
EL	73.55	61.4	66.93	65.81	**66.83**	66.31	72.54	72.69	72.62	71.83	71.48	71.66
EN				78.96	48.33	59.96				66.61	64.72	65.65
ES				57	54.27	55.6				55.45	56.27	55.86
EU	83.15	71.58	76.94	**85.15**	**79.42**	**82.18**	80.72	79.36	80.03	80.49	**80.9**	**80.69**
FA				86.56	61.49	71.9				87.3	85.46	86.37
FR	84.52	73.51	78.63	84.02	**74.17**	**78.79**	80.04	78.81	79.42	**81.57**	**79.18**	**80.36**
GA	77.17	16.28	26.89	36.21	**21.11**	26.67	37.72	25	30.07	*	*	*
HE	65.84	31.81	42.9	57.43	**39.64**	**46.91**	56.2	42.35	48.3	**58.1**	37.48	45.56
HI	86.56	39.23	53.99	**89.9**	**43.58**	**58.7**	72.25	75.04	73.62	**72.51**	72.64	72.57
HR				83.27	68.87	75.39				*	*	*
HU				95.6	19.23	32.02				*	*	*
IT	67.76	62.31	64.92	**67.82**	**62.5**	**65.05**	67.68	60.27	63.76	66.63	**60.37**	63.35
LT				78.03	35.66	48.95				62.47	47.75	54.12
MT				17.92	15.36	16.54				19.29	10.61	13.69
PL	91.15	74.28	81.85	**93.16**	74.07	**82.53**	82.94	79.18	81.02	82.2	78.88	80.51
PT	75.81	69.99	72.79	**79.71**	69.16	**74.06**	73.93	72.76	73.34	73.85	**74.04**	**73.95**
RO	82.69	81.81	82.25	65.74	**86.93**	74.87	89.88	91.05	90.46	*	*	*
SL				33.87	54.73	41.84				41.29	31.66	35.84
SR				87.46	48.11	62.08				69.09	62.4	65.57
SV	86.07	59.96	70.68	**93.27**	**73.56**	**82.25**	69.59	73.68	71.58	**73.94**	**80.44**	**77.06**
TR	61.69	65.33	63.46	60.24	**70.74**	65.07	68.41	70.55	69.46	66.48	**75.54**	**70.72**
ZH	44.84	54.71	49.28	25.47	56.3	35.07	68.56	70.74	69.63	64.5	61.92	63.18

Table 1: Comparing Seen2Seen and MTLB-STRUCT performance across 14 languages (Shared Task 1.2) and 26 languages (Release 1.3): Global MWE-based Precision (P), Recall (R), and F-measure (F1).

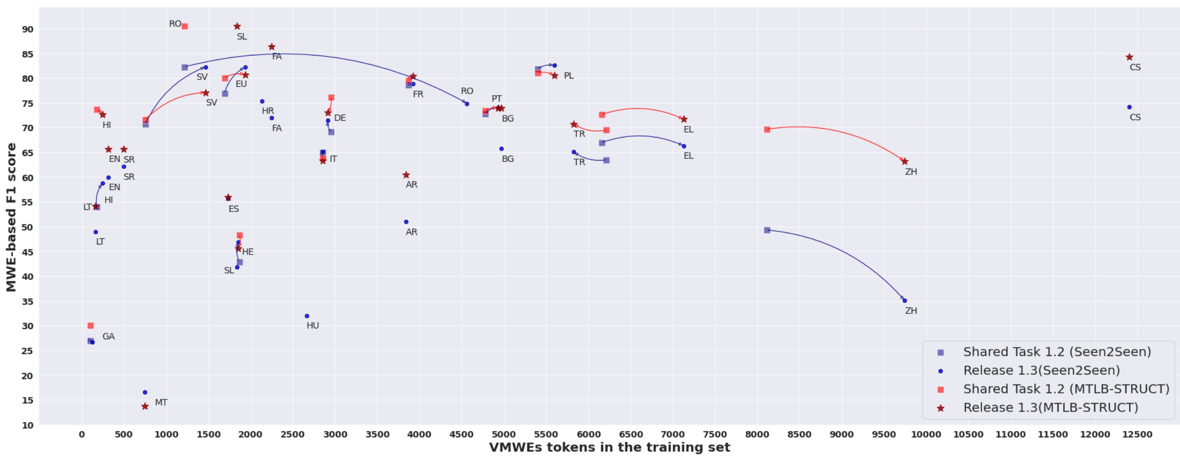

Figure 3: Seen2Seen and (multi-task) MTLB-SRUCT performance: A Comparison of MWE-based F1-Scores and VMWEs tokens in the training set between Shared Task 1.2 and release 1.3

Two other initiatives which contributed to the outcomes presented here are the CA21167 COST action UniDive (Universality, diversity and idiosyncrasy in language technology) and the Dagstuhl Seminar 21351 (Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics).

References

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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Jan Hajic, Otakar Smrz, Petr Zemánek, Jan Šnaidauf, and Emanuel Beška. 2004. Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, volume 1.

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020. Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3333–3345.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

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, 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, 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. Association for Computational Linguistics.

Agata Savary, Sara Stymne, Verginica Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, to appear.

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. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. Mtlb-struct@ parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. *arXiv preprint arXiv:2011.02541*.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Full corpus statistics and system results

Lang-split	Sentences	Tokens	Avg. length	VMWE	VID	IRV	LVC.full	LVC.cause	VPC.full	VPC.semi	IAV	MVC	LS.ICV
AR-train	6091	252456	41.4	3841	955	0	2178	236	0	0	468	4	0
AR-dev	342	14746	43.1	228	54	0	121	15	0	0	38	0	0
AR-test	1050	44541	42.4	680	173	0	379	52	0	0	75	1	0
AR-Total	7483	311743	41.6	4749	1182	0	2678	303	0	0	581	5	0
BG-train	15950	353748	22.1	4969	922	2421	1401	157	0	0	68	0	0
BG-dev	1380	30980	22.4	431	88	179	138	22	0	0	4	0	0
BG-test	4269	95685	22.4	1304	250	623	370	43	0	0	18	0	0
BG-Total	21599	480413	22.2	6704	1260	3223	1909	222	0	0	90	0	0
CS-train	42288	711213	16.8	12405	1353	8576	2476	0	0	0	0	0	0
CS-dev	1725	28697	16.6	523	68	357	98	0	0	0	0	0	0
CS-test	5418	93283	17.2	1608	192	1067	349	0	0	0	0	0	0
CS-Total	49431	833193	16.8	14536	1613	10000	2923	0	0	0	0	0	0
DE-train	6475	125081	19.3	2912	1015	230	222	23	1277	145	0	0	0
DE-dev	628	12046	19.1	281	103	25	22	6	119	6	0	0	0
DE-test	1893	36434	19.2	848	319	67	67	4	348	43	0	0	0
DE-Total	8996	173561	19.2	4041	1437	322	311	33	1744	194	0	0	0
EL-train	21983	587001	26.7	7128	2368	1	4430	154	127	0	0	48	0
EL-dev	1077	28833	26.7	348	107	0	228	9	4	0	0	0	0
EL-test	3115	82590	26.5	1032	366	0	635	16	12	0	0	3	0
EL-Total	26175	698424	26.6	8508	2841	1	5293	179	143	0	0	51	0
EN-train	2150	35534	16.5	317	44	0	98	12	112	16	22	13	0
EN-dev	1302	21660	16.6	199	35	0	63	10	62	7	13	9	0
EN-test	3984	67009	16.8	598	108	0	172	29	194	30	36	29	0
EN-Total	7436	124203	16.7	1114	187	0	333	51	368	53	71	51	0
ES-train	3424	112906	32.9	1732	200	433	259	54	0	0	328	458	0
ES-dev	521	17333	33.2	256	31	73	36	2	0	0	47	67	0
ES-test	1570	52125	33.2	751	96	208	97	25	1	0	136	188	0
ES-Total	5515	182364	33	2739	327	714	392	81	1	0	511	713	0
EU-train	5033	70017	13.9	1932	392	0	1444	96	0	0	0	0	0
EU-dev	1441	20957	14.5	560	130	0	404	26	0	0	0	0	0
EU-test	4684	66833	14.2	1754	358	0	1304	92	0	0	0	0	0
EU-Total	11158	157807	14.1	4246	880	0	3152	214	0	0	0	0	0
FA-train	2364	40110	16.9	2249	11	1	2237	0	0	0	0	0	0
FA-dev	321	5430	16.9	303	1	0	302	0	0	0	0	0	0
FA-test	932	16028	17.1	901	5	0	896	0	0	0	0	0	0
FA-Total	3617	61568	17	3453	17	1	3435	0	0	0	0	0	0
FR-train	14540	364414	25	3921	1529	1024	1286	63	0	0	0	19	0
FR-dev	1580	40107	25.3	437	157	123	146	11	0	0	0	0	0
FR-test	4841	121321	25	1297	471	354	446	23	0	0	0	3	0
FR-Total	20961	525842	25	5655	2157	1501	1878	97	0	0	0	22	0
GA-train	330	7104	21.5	127	25	0	43	19	3	6	31	0	0
GA-dev	318	7680	24.1	134	24	0	42	21	4	2	41	0	0
GA-test	1057	24123	22.8	398	57	0	115	78	21	12	115	0	0
GA-Total	1705	38907	22.8	659	106	0	200	118	28	20	187	0	0
HE-train	14035	283984	20.2	1855	848	0	740	158	109	0	0	0	0
HE-dev	1296	26766	20.6	171	59	0	90	10	12	0	0	0	0
HE-test	3869	77731	20	507	201	0	219	55	32	0	0	0	0
HE-Total	19200	388481	20.2	2533	1108	0	1049	223	153	0	0	0	0
HI-train	399	8641	21.6	242	13	0	155	7	0	0	0	67	0
HI-dev	322	6786	21	200	15	0	123	4	0	0	0	58	0
HI-test	963	20003	20.7	592	33	0	363	15	0	0	0	181	0
HI-Total	1684	35430	21	1034	61	0	641	26	0	0	0	306	0
HR-train	3357	77599	23.1	2131	161	657	476	81	0	0	756	0	0
HR-dev	672	15329	22.8	439	35	132	90	20	0	0	162	0	0
HR-test	2104	50018	23.7	1332	97	404	314	46	1	0	470	0	0
HR-Total	6133	142946	23.3	3902	293	1193	880	147	1	0	1388	0	0
HU-train	2139	54658	25.5	2664	39	0	400	130	1755	340	0	0	0
HU-dev	1000	25205	25.2	1259	19	0	173	69	843	155	0	0	0
HU-test	3020	76473	25.3	3837	46	0	570	202	2558	461	0	0	0
HU-Total	6159	156336	25.3	7760	104	0	1143	401	5156	956	0	0	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
LT-train	2281	42782	18.7	163	53	0	102	8	0	0	0	0	0
LT-dev	2181	41421	18.9	161	66	0	91	4	0	0	0	0	0
LT-test	6642	124309	18.7	488	189	0	286	13	0	0	0	0	0
LT-Total	11104	208512	18.7	812	308	0	479	25	0	0	0	0	0
MT-train	6460	154979	23.9	749	311	0	434	1	3	0	0	0	0
MT-dev	975	22924	23.5	119	53	0	65	0	0	0	0	1	0
MT-test	3165	74382	23.5	358	154	1	201	0	1	0	0	1	0
MT-Total	10600	252285	23.8	1226	518	1	700	1	4	0	0	2	0
PL-train	18037	303628	16.8	5595	637	2832	1881	245	0	0	0	0	0
PL-dev	1421	23865	16.7	430	54	199	163	14	0	0	0	0	0
PL-test	4089	68647	16.7	1288	142	657	434	55	0	0	0	0	0
PL-Total	23547	396140	16.8	7313	833	3688	2478	314	0	0	0	0	0

Lang-split	Sentences	Tokens	Avg. length	VMWE	VID	IRV	LVC.full	LVC.cause	VPC.full	VPC.semi	IAV	MVC	LS.ICV
PT-train	24594	557486	22.6	4926	999	782	3031	99	0	0	0	15	0
PT-dev	1867	42855	22.9	375	72	64	229	10	0	0	0	0	0
PT-test	5601	127728	22.8	1125	235	175	694	18	0	0	0	3	0
PT-Total	32062	728069	22.7	6426	1306	1021	3954	127	0	0	0	18	0
RO-train	26889	479681	17.8	4562	806	1799	246	87	0	0	1624	0	0
RO-dev	7668	139314	18.1	1257	222	516	64	22	0	0	433	0	0
RO-test	22107	395913	17.9	3689	616	1511	206	73	0	0	1283	0	0
RO-Total	56664	1014908	17.9	9508	1644	3826	516	182	0	0	3340	0	0
SL-train	15220	321377	21.1	1834	390	885	135	37	0	0	387	0	0
SL-dev	3054	64429	21	376	79	189	27	8	0	0	73	0	0
SL-test	9551	200381	20.9	1153	255	552	77	19	0	0	250	0	0
SL-Total	27825	586187	21	3363	724	1626	239	64	0	0	710	0	0
SR-train	1382	33839	24.4	492	100	212	158	22	0	0	0	0	0
SR-dev	544	13558	24.9	203	49	91	53	10	0	0	0	0	0
SR-test	1660	39970	24	609	120	261	191	37	0	0	0	0	0
SR-Total	3586	87367	24.3	1304	269	564	402	69	0	0	0	0	0
SV-train	2795	44904	16	1466	189	106	197	3	681	290	0	0	0
SV-dev	765	12328	16.1	421	66	29	54	2	199	71	0	0	0
SV-test	2466	39588	16	1268	186	102	166	5	581	228	0	0	0
SV-Total	6026	96820	16	3155	441	237	417	10	1461	589	0	0	0
TR-train	16730	248697	14.8	5824	3140	0	2679	0	0	0	0	5	0
TR-dev	1396	20679	14.8	466	250	0	216	0	0	0	0	0	0
TR-test	4180	62793	15	1439	751	0	688	0	0	0	0	0	0
TR-Total	22306	332169	14.8	7729	4141	0	3583	0	0	0	0	5	0
ZH-train	44103	738713	16.7	9744	877	0	1101	158	0	4177	0	3431	0
ZH-dev	1215	19936	16.4	274	23	0	26	7	0	117	0	101	0
ZH-test	3611	61698	17	801	73	0	87	12	0	335	0	294	0
ZH-Total	48929	820347	16.7	10819	973	0	1214	177	0	4629	0	3826	0
Total	455629	9264811	20.3	127498	26214	29062	40933	3238	9164	6443	7375	5032	37

Table 2: Statistics of the 1.3 release of the PARSEME corpus