MultiVitaminBooster at PARSEME Shared Task 2020: Combining Window- and Dependency-Based Features with Multilingual Contextualised Word Embeddings for VMWE Detection

Sebastian Gombert and Sabine Bartsch
Corpus- and Computational Linguistics, English Philology
Institute of Linguistics and Literary Studies, Technische Universität Darmstadt
sebastiang@outlook.de
sabine.bartsch@tu-darmstadt.de

Abstract

In this paper, we present *MultiVitaminBooster*, a system implemented for the *PARSEME shared task on semi-supervised identification of verbal multiword expressions - edition 1.2*. For our approach, we interpret detecting verbal multiword expressions as a token classification task aiming to decide whether a token is part of a verbal multiword expression or not. For this purpose, we train gradient boosting-based models. We encode tokens as feature vectors combining multilingual contextualized word embeddings provided by the *XLM-RoBERTa* language model (Conneau et al., 2019) with a more traditional linguistic feature set relying on context windows and dependency relations. Our system was ranked 7th in the official open track ranking of the shared task evaluations with an encoding-related bug distorting the results. For this reason we carry out further unofficial evaluations. Unofficial versions of our systems would have achieved higher ranks.

1 Introduction

Multiword expressions (MWEs) are an object of research in various areas of linguistics and NLP. On the one hand, areas such as lexico-semantics and construction grammar have a distinct research interest in the form and semantics of different classes of multiword expressions (Masini, 2005); their automatic detection makes them accessible to large-scale corpus analyses. On the other hand, various NLP-systems, especially in the area of machine translation where the detection of MWEs prevents spurious literal translations, can benefit from detecting MWEs (Zaninello and Birch, 2020), as well.

In this paper, we present an approach to the automatic detection of verbal multiword expressions (VMWE), MWEs which form around a head verb, which was our contribution to the *PARSEME shared task 2020 on verbal multiword expressions* (Ramisch et al., 2020a). The data set provided for the shared task (Ramisch et al., 2020b) distinguishes 7 general categories of VMWEs, some with additional subcategories:

VMWE category	Tag	Example
Verbal idioms	VID	to let the cat out of the bag.
Light-verb constructions	LVC.full, LVC.cause	to make a decision
Verb-particle constructions	VPC.full, VPC.cause	to go on
Multi-verb constructions	MVC	to make do
Inherently reflexive verbs	IRV	sich beschäftigen (to deal with; to be concerned)
Inherently adpositional verbs	IAV	to stand for s. th.
Inherently clitic verbs	LS.ICV	se ne frega (he does not care)

Table 1: The different categories of VMWEs dealt with during the shared task.

Past approaches in the area of MWE detection rely on the usage of statistical association measures (Evert et al., 2017; Pecina, 2005; Ramisch et al., 2008; Tsvetkov and Wintner, 2010) and machine learning (Klyueva et al., 2017; Moreau et al., 2018; Stodden et al., 2018; Waszczuk, 2018), sometimes combining

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 149–155
Barcelona, Spain (Online), December 13, 2020.

both (Mandravickaitė and Krilavičius, 2017). Our approach follows this tradition while paying tribute to the latest developments in the area of multilingual transformer-based neural language modeling.

Transformer-based language models such as *BERT* (Devlin et al., 2019) or *RoBERTa* (Liu et al., 2019) are pre-trained on large-scale corpora and are able to achieve state-of-the-art results for various standard tasks in NLP. They can either be fine-tuned to solve a specific task or used to provide contextualised word embeddings. The difference between such contextualised embeddings and the static ones based on traditional methods such as *GloVe* (Pennington et al., 2014) or *word2vec* (Mikolov et al., 2013) is that the former can account for different local word contexts and encode a given word individually with regard to the observed local context as well as global distributional information.

By intuition, this aspect should make contextualised word embeddings a feasible candidate for the detection of MWEs, as they should encode a word differently when observed as part of an MWE compared to an occurrence in open distribution, given both cases were reflected during pre-training. As this aspect can, however, not be guaranteed in all cases, we complement the embeddings with regular window- and dependency-based features to make the results of our systems less dependent on the pre-training of the language model used, and, thus, more robust. To be able to account for the multiple languages represented in the shared task without providing a distinct transformer language model for each language, we use the multilingual *XLM-RoBERTa* (Conneau et al., 2019) which was trained on *Common Crawl* data in 100 languages (including all languages represented in the shared task) and is, as a consequence, able to generate contextualised word embeddings for all of them.

2 System Description

For our system, *MultiVitaminBooster*, we interpret detecting MWEs as a binary token classification task whose goal is to decide whether a token is part of a given MWE or not. We train respective binary classifiers per language and MWE category to account for the phenomenon of overlapping MWEs from different categories using the *train* and *dev* sets provided for the shared task. For re-assembling single tokens into coherent MWEs, we calculate sub-graphs of the dependency tree of a given sentence where all nodes not marked as being part of a given MWE category are filtered out. We interpret the remaining connected components within them as coherent MWEs and tag the tokens accordingly.

2.1 Feature Encoding

XLM-RoBERTa-based contextualised word embeddings: we represent each token by the contextualised word embedding generated by XLM-RoBERTa by itself as well as for its parent within the dependency tree of a sentence and the root of the respective sentence. To this end, we add the averages of the contextualised embeddings of all children and siblings of a given token. For acquiring the embeddings, we rely on the version of *XLM-RoBERTa* provided by *huggingface.co* (Wolf et al., 2019) and use the *base model*. As this language model requires a more fine-grained segmentation of tokens than present in the training data (e. g. highly productive morphemes are regarded as independent tokens to save input dimensions) and because, as a consequence, a token within the training set might correspond to multiple sub-tokens and, thus, to multiple contextualised embeddings, we average these embeddings in such cases.

XLM-RoBERTa attention values: as transformer-based language models use attention during the calculation of representations (Vaswani et al., 2017), they provide numerical values directly indicating the importance of tokens for the semantics of each other. Our intuition is that the question whether two tokens are part of a given MWE, or not, could be reflected in the attention they show for each other. We encode each token with the attention it pays its parent and the root of a given sentence as well as the attention both of these pay to the token itself. Analogous to the embeddings themselves, we average the attention values in cases where multiple sub-tokens correspond to a token from the training data.

Linguistic features (window- and dependency-based): the training- and evaluation corpora of the shared task (Ramisch et al., 2020b) comply to the format of the *universal dependencies* project (McDonald et al., 2013). The majority of them were either automatically annotated with lemmata, *UD* POS tags,

language-specific POS tags, *universal features*[1] and *UD* dependency relations using *UDPipe* (Straka and Straková, 2017) or taken directly from official *UD treebanks*.

We encode each token with its corresponding lemma (we filter out all lemmata not observed as being part of an VMWE within the training corpus), its language-specific POS tag, its *universal features* and its dependency tag. To this, we add the corresponding annotations of the parent, siblings and children of a token within the dependency tree of a given sentence, the respective annotations of the root token of this sentence within this tree, and the corresponding annotations of neighbouring tokens given a size of two for the left and the right context. We encode these features as one-hot respectively *n*-hot-vectors.

2.2 Classification

Gradient Boosting: for *MultiVitaminBooster*, we use gradient boosting (Mason et al., 1999) relying on the implementation provided by *CatBoost* (Dorogush et al., 2018; Prokhorenkova et al., 2018). Gradient boosting creates an ensemble of weak learners in the form of regression trees in order to create a strong one. The logit parameters predicted by these trees are combined into final prediction scores using a variant of logistic regression. We chose gradient boosting as an algorithm as it is able to create complex and powerful classification models for heterogeneous feature sets. We use the default parameters and train for 1000 epochs. Per language and VMWE category present for this language, we train a binary token classifier whose goal is to decide whether a token is part of a VMWE of the respective category.

2.3 VMWE re-assembly

This leaves us with tagged tokens. However, the task requires VMWEs to form connected units indicating relations between the different corresponding tokens within the output data. To reconnect the single tokens tagged as VMWE within a given sentence into such complete units, we implemented the following heuristics which is executed per language and VMWE category:

- We instantiate the dependency tree of a sentence as a graph.
- Within this graph, we delete all nodes corresponding to tokens without a respective VMWE tag.
- We interpret the remaining connected components (= remaining sub-graphs consisting of one or more connected nodes) as coherent VMWEs.

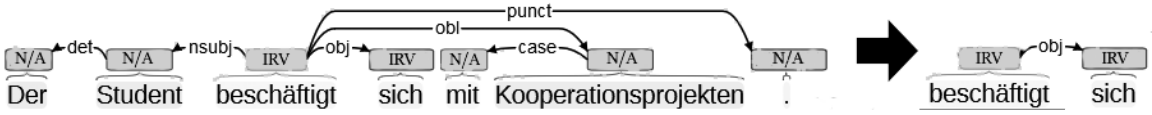

Figure 1: An example illustrating this heuristic. Translation of this utterance: *the student is concerned with cooperation projects.*

3 Results and Analysis

As already stated, our official system (*MVB*) was ranked last in the official shared task evaluations with an output encoding-related bug distorting results (under a common condition, it was likely that multiple sentences were assigned the same VMWE tags). For this reason, we evaluated a bug-fixed system (*MVB (bug free)*) for seven out of the 14 languages represented in the shared task (*DE*, *EU*, *GA*, *HI*, *IT*, *SV* and *TR*; due to time-related reasons, we only managed to evaluate our system for these languages for the official shared task which is why we decided to focus on them throughout all other evaluations).

In addition to our bug-fixed system, we trained a system exclusively on the contextualised embeddings and attention values (*Emb. Att. B.*), another system exclusively on the window- and dependency-based linguistic features (*Ling. Feats. B.*) and a third system which relies on the same feature set as *MVB* but

[1] https://universaldependencies.org/u/feat/index.html

uses *logistic regression* as classification algorithm (we rely on the implementation provided by *scikit-learn* (Pedregosa et al., 2011) for this and train for 1000 iterations) as additional baselines.[2]

The official evaluations of the shared task were separated into two tracks. Systems participating in the closed track were obliged to only rely on the data sets provided directly by the organisers (Ramisch et al., 2020b), while systems participating in the open track were allowed to use external resources such as external corpora or lexical resources, as well. All our systems except for the one relying solely on window- and dependency-based linguistic features would have participated in the open track due to the usage of the external contextualised embeddings if submitted for the official shared task. The systems were evaluated with regard to different categories with the three most important ones being *unseen MWE-based*, a category that evaluates the performance of systems in respect to VMWEs not observed within the training data, *global MWE-based*, a category that evaluates the general detection of VMWEs as connected units, and *global token-based*, a category evaluating the detection of VMWEs on a token level.

System	Unseen MWE-based				Global MWE-based				Global Token-based			
	P	**R**	**F1**	**Rank**	**P**	**R**	**F1**	**Rank**	**P**	**R**	**F1**	**Rank**
MVB	0.05	0.07	0.06	7	0.19	0.09	0.12	7	3.49	1.26	1.85	7
MVB (bug free)	3.74	3.48	3.61	*6**	52.21	30.02	38.12	*5**	**84.10**	35.16	49.59	*5**
Ling. Feats. B.	14.88	10.03	11.98	*2***	56.62	35.72	43.80	*3***	**82.52**	41.49	55.22	*3***
Emb. Att. B.	13.44	0.42	0.81	*7**	25.81	0.89	1.72	*7**	67.11	1.73	3.37	*7**
MVLR	3.65	21.12	6.22	*5**	21.60	46.89	29.58	*5**	37.87	70.76	49.34	*5**
MTLB-STRUCT	*36.24*	*41.12*	*38.53*	*1*	*71.26*	*69.05*	*70.14*	*1*	*77.69*	*70.9*	*74.14*	*1*
SEEN2SEEN	*36.47*	*0.57*	*1.12*	*2*	*76.21*	*58.56*	*66.23*	*1*	*78.64*	*57.02*	*66.11*	*1*

Table 2: The overall results of our evaluations for the seven languages. ** = unofficial rank in the open track. ** = unofficial rank in the closed track. MTLB-STRUCT* and *SEEN2SEEN* are the winning systems of the two tracks of the shared tasks. We provide their results for reasons of comparability. **Bold** marks the highest score reached within a category throughout all shared task results within a given track. Underline marks the best score reached among our systems.

The bug-free version of *MultiVitaminBooster* would have been ranked fifth within the *global MWE-based* and *global token-based* evaluation categories and sixth within the *unseen MWE-based* category if it had participated in the official shared task. While this is a huge improvement over the bugged version, these results can be considered subpar, especially in comparison to the winning systems *MTLB-STRUCT* and *SEEN2SEEN*.

Two further observations which speak against our form of usage of multilingual contextualised word embeddings for the given task can be made here, as well: on the one hand, the system which was trained solely on them (*Emb. Att. B.*) performed by far worst out of all our unofficial systems, and, on the other hand, the system which was trained solely on the window- and dependency-based linguistic feature set (*Ling. Feats. B.*) performed best out of all our systems and even manages to put *MultiVitaminBooster* into place. If submitted to the shared track, it would have been ranked second in the *unseeen MWE-based* category and third in the *global MWE-based* and *global token-based* categories for the closed track. Here, the question whether these results would have turned out more successful if another language model had been used instead of *XLM-RoBERTa* (Conneau et al., 2019) or if our results reflect a general inadequacy of the approach to use transformer-based word embeddings for detecting VMWEs remains.

Our model based on logistic regression (*MVLR*) achieved lower precision but higher recall scores than all our systems based on gradient boosting and, by average, lower F1-scores than the best gradient boosting-based models. This raises the question to what extent the results would differ when applying other classification algorithms.

One other important finding is that for all our systems, there is a discrepancy between the precision scores observed for *global MWE-based* and *global token-based*. While the precision achieved for the *global MWE-based* category turned out subpar, *MultivitaminBooster* and the system relying solely on the linguistic feature set achieve the best precision scores out of all our unoffical and all official systems within the *global token-based* category. We attribute this discrepancy to our heuristics used for re-assembling VMWEs.

[2]Our code and our full evaluation results can be found under `https://github.com/SGombert/MultiVitaminBoosterResults`

A closer look onto an example can explain this: *"'Es **tut weh**, die **Sprache** derer **benützen** zu müssen, die **dich schinden**', **heißt es** da beispielsweise schon am **Anfang** [...]." ("'It hurts having to use the language of those who maltreat you', it says, for example, at the beginning.")* In this German sentence, all tokens marked as bold were recognised as one large VMWE instead of multiple ones, illustrating a problematic pattern which is observable for our systems throughout all languages evaluated. Tokens of multiple VMWEs of the same category can form connected components within the dependency tree of a given sentence which our heutristics is not able to resolve in a correct way. A solution to this would be to further inspect the dependency relations for which this phenomenon is observable and to try to identify criteria to filter them out under given circumstances.

System	Unseen MWE-based				Global MWE-based				Global Token-based			
	P	R	F1	Rank	P	R	F1	Rank	P	R	F1	Rank
MVB (bug free)	0.65	2.27	1.02	*4**	65.14	50.82	57.10	*3**	87.07	59.50	70.69	*2**
Ling. Feats. B.	1.63	4.55	2.40	*2***	70.50	56.46	62.71	*1***	86.09	64.16	73.52	*1***
MTLB-STRUCT	*48.75*	*58.33*	*53.11*	*1*	*72.25*	*75.04*	*73.62*	*1*	*81.20*	*77.24*	*79.17*	*1*
ERMI	*37.09*	*41.67*	*39.25*	*1*	*63.48*	*56.32*	*59.69*	*1*	*79.48*	*62*	*69.66*	*1*

Table 3: The results of two of our systems for the langugage *Hindi*. ** = unofficial rank in the open track. ** = unofficial rank in the closed track. MTLB-STRUCT* and *ERMI* are the winning systems of the two tracks of the shared task for this language. We provide their results for reasons of comparability.

A further observation is that in the case of the language *Hindi*, the results achieved by our systems show positive outliers. Here, *MultiVitaminBooster* would have ranked third in the *global MWE-based* category and second in the *global token-based* category. The unofficial system trained solely on window- and dependey-based features even manages to achieve unofficial first ranks in the closed track.

4 Conclusion and Future Work

We presented *MultiVitaminBooster* and three unofficial systems implemented for the *PARSEME shared task 2020 on verbal multiword expressions*. We evaluated our systems for seven languages. The best of our systems would have ranked fifth in the official shared task. A positive outlier can be observed for the language *Hindi*, where our systems achieved more competitive results. The usage of multilingual contextualized word embeddings for our systems can be considered a failure, as the same deteriorated results and our system relying solely on the linguistic feature set achieved superior results. It is, however, to explore if this would have turned out differently with another language model.

To summarise, there remains room for improvement. Using statistical association measures induced from large scale corpora as additional features may be a route to further explore this. An improved redesign of the heuristics used for assembling tokens into VMWEs built on a more complex rule set could lead to improvements in *MWE-based* precision scores and close the gap to the *token-based* ones. Different classification algorithms, such as CRFs or SVMs, could be explored as alternatives to gradient boosting and logistic regression, as well as different variations and combinations of training hyperparameters to aim for better regularisation.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *CoRR*, abs/1810.11363.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. 2017. E-VIEW-Alation – a Large-Scale Evaluation Study of Association Measures for Collocation Identification. In Miloš J Jelena K Simon K Iztok K, Carole T

and Vít B, editors, *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 Conference*, page 531–549, Brno. Lexical Computing.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Justina Mandravickaitė and Tomas Krilavičius. 2017. Identification of multiword expressions for Latvian and Lithuanian: Hybrid approach. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 97–101, Valencia, Spain, April. Association for Computational Linguistics.

Francesca Masini. 2005. Multi-word expressions between syntax and the lexicon : The case of italian verb-particle constructions. *Sky Journal of Linguistics*, 18:145–173.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. 1999. Boosting algorithms as gradient descent. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 512–518, Cambridge, MA, USA. MIT Press.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, and Carl Vogel. 2018. CRF-seq and CRF-DepTree at PARSEME shared task 2018: Detecting verbal MWEs using sequential and dependency-based approaches. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 241–247, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Pavel Pecina. 2005. An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop*, pages 13–18, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. Catboost: unbiased boosting with categorical features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 6639–6649.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *In Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008*, pages 50–53.

Carlos Ramisch, Bruno Guillaume, Agata Savary, Jakub Waszczuk, Marie Candito, and Ashwini Vaidya. 2020a. Shared task on semi-supervised identification of verbal multiword expressions - edition 1.2. `http://multiword.sourceforge.net/PHITE.php?sitesig=CONF&page=CONF_02_MWE-LEX_2020___lb__COLING__rb__&subpage=CONF_40_Shared_Task`. Accessed: 2020-09-09.

Carlos Ramisch, Bruno Guillaume, Agata Savary, 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 Stymme, Abigail Walsh, Hongzhi Xu, Emilia Palka-Binkiewicz, Rafael Ehren, Sara Stymne, Matthieu Constant, Caroline Pasquer, Yannick Parmentier, Jean-Yves Antoine, Carola Carlino, Valeria Caruso, Maria Pia Di Buono, Antonio Pascucci, Annalisa Raffone, Anna Riccio, Federico Sangati, Giulia Speranza, Renata Ramisch, Silvio Ricardo Cordeiro, Helena de Medeiros Caseli, Isaac Miranda, Alexandre Rademaker, Oto Vale, Aline Villavicencio, Gabriela Wick Pedro, Rodrigo Wilkens, Leonardo Zilio, Monica-Mihaela Rizea, Mihaela Ionescu, Mihaela Onofrei, Jia Chen, Xiaomin Ge, Fangyuan Hu, Sha Hu, Minli Li, Siyuan Liu, Zhenzhen Qin, Ruilong Sun, Chenweng Wang, Huangyang Xiao, Peiyi Yan, Tsy Yih, Ke Yu, Songping Yu, Si Zeng, Yongchen Zhang, Yun Zhao, Vassiliki Foufi, Aggeliki Fotopoulou, Stella Markantonatou, Stella Papadelli, Sevasti Louizou, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez, Antton Gurrutxaga, Larraitz Uria, Ruben Urizar, Jennifer Foster, Teresa Lynn, Hevi Elyovitch, Yaakov Ha-Cohen Kerner, Ruth Malka, Kanishka Jain, Vandana Puri, Shraddha Ratori, Vishakha Shukla, Shubham Srivastava, Gozde Berk, Berna Erden, and Zeynep Yirmibeşoğlu. 2020b. Annotated corpora and tools of the PARSEME shared task on semi-supervised identification of verbal multiword expressions (edition 1.2). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Regina Stodden, Behrang QasemiZadeh, and Laura Kallmeyer. 2018. TRAPACC and TRAPACCS at PARSEME shared task 2018: Neural transition tagging of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 268–274, Santa Fe, New Mexico, USA, August. 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.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In Chu-Ren Huang and Dan Jurafsky, editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 1256–1264. Chinese Information Processing Society of China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jakub Waszczuk. 2018. TRAVERSAL at PARSEME shared task 2018: Identification of verbal multiword expressions using a discriminative tree-structured model. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 275–282, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3816–3825, Marseille, France, May. European Language Resources Association.