Cross-type French Multiword Expression Identification with Pre-trained Masked Language Models

Van-Tuan Bui, Agata Savary
CNRS, Paris-Saclay University, LISN, France
van-tuan.bui@lisn.upsaclay.fr, agata.savary@universite-paris-saclay.fr

Abstract

Multiword expressions (MWEs) pose difficulties for natural language processing (NLP) due to their linguistic features, such as syntactic and semantic properties, which distinguish them from regular word groupings. This paper describes a combination of two systems: one that learns verbal multiword expressions (VMWEs) and another that learns non-verbal MWEs (nVMWEs). Together, these systems leverage training data from both types of MWEs to enhance performance on a cross-type dataset containing both VMWEs and nVMWEs. Such scenarios emerge when datasets are developed using differing annotation schemes. We explore the fine-tuning of several state-of-the-art neural transformers for each MWE type. Our experiments demonstrate the advantages of the combined system over multi-task approaches or single-task models, addressing the challenges posed by diverse tagsets within the training data. Specifically, we evaluated the combined system on a French treebank named Sequoia, which features an annotation layer encompassing all syntactic types of French MWEs. With this combined approach, we improved the F1-score by approximately 3% on the Sequoia dataset.

Keywords: Multiword Expressions, Transformers, Deep Learning

1. Introduction

A multiword expression (MWE) is a combination of words which exhibits lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies (Baldwin and Kim, 2010). MWE identification, a subtask of MWE processing, takes a corpus as input and adds an annotation layer to indicate where MWE instances are located. This identification process shares some characteristics with the named entity recognition (NER), but it is worth noting that MWEs consist of at least two words which do not have to be adjacent. Recognizing MWEs is useful for various NLP tasks, such as parsing, machine translation, semantic processing and information retrieval (Constant et al., 2017). A recent paper by Baziotis et al., 2023, uses MWE-specific machine translation evaluation measures and shows that MWEs are more difficult to translate than regular words and that pretraining a translation model on monolingual data (of the source language) containing MWEs increases the quality of the final translation. Thus, if MWE identification is reliable, we can easily select such monolingual pre-training corpora. Also, Haviv et al., 2023 show one can increase interpretability of language models (LMs) by focusing on MWEs, many of which are items that transformer LMs necessarily memorize (as opposed to generalising).

Recent studies in MWE identification have explored fine-tuning of generic pre-trained masked transformer-based models (Taslimipoor et al., 2020; Premasiri and Ranasinghe, 2022; Avram et al., 2023; Kurfalı, 2020). In such a setting, high-quality features are generated autonomously, which eliminates the need of manually crafted features. Moreover, this approach integrates context into the learning process more effectively than methods such as Recurrent Neural Networks (RNNs), syntax-based candidate extraction and filtering with association measures, and rule-based joint parsing (Ramisch et al., 2020). A number of pre-trained language models exist, both monolingual and multilingual. They differ in which neural transformer architectures and training corpora they use.

Fine-tuning these pre-trained models calls for annotated data fitting the task at hand. In our case, the task is to annotate MWEs of all syntactic types (verbal, nominal, adjectival, adverbial, functional, etc.), henceforth called, allMWEs, in French. We have access to two manually annotated datasets: (i) Sequoia, a small French corpus annotated for allMWEs, including verbal ones (VMWEs) (Candito et al., 2020), and (ii) PARSEME, a much larger multilingual corpus, including French, but annotated for VMWEs only (Savary et al., 2023).

The problem is how to combine these heterogeneous datasets, and which pre-trained models to fine-tune on them, for an effective allMWE identifier in French. To this aim, we carried out three experiments:

1. Identification of allMWEs is treated as a *single task*. We fine-tune seven distinct transformer models, using Sequoia only.
2. Identification of non-verbal MWEs (nVMWEs) and VMWEs is treated as *two independent tasks*. We use two versions of Sequoia: one with nVMWE and one with VMWE annotations. The latter is augmented with PARSEME. The

LREC-COLING 2024, pages 4198–4204

predictions of the two fine-tuned systems are merged using a union approach.

3. Identification of nVMWEs and VMWEs is treated as a *multi-task* problem. We fine-tune a model on VMWEs and nVMWEs jointly.

Our findings indicate that the union of two independent models outperforms both the single-task and multi-task models. To the best of our knowledge, Sequoia is the first corpus wherein the PARSEME annotation schema has been extended to include nVMWEs. However, to date, no system has been trained on this corpus for the identification of all syntactic types of MWEs. Therefore, our findings constitute the first reported results for this specific problem.

More broadly, we contribute to the problem of training models on heterogenous datasets. More precisely, we offer a protocol for the scenario of *selective annotation* as defined by Beryozkin et al. (2019).

2. Datasets

2.1. PARSEME Corpora

The PARSEME multilingual corpora have been annotated with VMWEs to serve, notably, as both training and testing resources for shared tasks (Savary et al., 2017; Ramisch et al., 2018, 2020), with the aim of enhancing the identification of VMWEs in written content. Presently in version 1.3, the PARSEME corpora encompass 26 languages, including French. Collectively, the corpora comprise $455,629$ sentences, equivalent to 9 million tokens, and $127,498$ VMWEs. Specifically, the French corpus contains $20,961$ sentences, equivalent to $525,842$ tokens, with $5,655$ annotated VMWEs .

The VMWE types are detailed in the annotation guidelines.[1] The VMWE types annotated specifically for French are as follows:

- IRV (inherently reflexive verbs): e.g. ***s'évanouir*** (lit. 'to faint oneself') 'to faint'
- LVC.full (light verb constructions in which the verb is semantically totally bleached): e.g. ***faire*** *une* ***présentation*** 'to make a presentation'
- LVC.cause (light verb constructions in which the verb adds a causative meaning to the noun): e.g. ***donner*** *le* ***droit*** 'to grant the right'
- VID (verbal idioms): e.g. ***se faire des idées*** (lit. 'make oneself ideas') 'to imagine something false'
- MVC (multi-verb constructions): e.g. *ce mot* ***veut dire*** *autre chose* (lit. 'this word wants to mean something else') 'this word means something else'

2.2. Sequoia Corpus

Sequoia (Candito et al., 2020) is a French treebank that encompasses various written genres (news, parliamentary debates, wikipedia narratives, and medical reports). The published corpus,[2] currently in version 9.2, includes $3,451$ allMWEs, of which 981 are VMWEs and $2,470$ are nVMWEs. VMWE annotations in this corpus were adopted from the multilingual PARSEME corpora, specifically the Sequoia corpus, which shares the same name. Additionally, the corpus contains annotations for nVMWEs, such as ***à la suite*** (lit. 'in the follow-up') 'following', ***dur à la tâche*** (lit. 'hard in the task') 'hardworking' and ***sécurité routière*** 'road safety'. Despite its relatively modest size, this corpus stands as the sole open-source treebank with annotations for allMWEs in French.[3]

2.3. Heterogeneity in the Two Datasets

The texts in Sequoia bear annotations for both VMWEs and nVMWEs. The same texts are included in the French PARSEME corpus but with VMWE annotations only. Consequently, merging the two datasets and training a single model across all MWE types could lead to numerous false negatives due to the absence of the nVMWE annotations in PARSEME. Such issues have previously been observed in NER tasks (Beryozkin et al., 2019; Greenberg et al., 2018). This scenario emerges when varying annotation schemes are applied to multiple datasets for the same task (Beryozkin et al., 2019). These authors propose a taxonomy of such scenarios and our case falls under *selective annotation*, since a pre-existing tagset for VMWEs (IRV, LVC.full, LVC.cause, VID, MVC) is extended with a new tag (nVMWE) but annotations are completed with this new tag only for a small part of the initial corpus.

A straightforward strategy is to rely solely on the Sequoia corpus to cover the entire tagset (Sec. 3.1). However, the number of VMWE examples in this corpus is relatively limited (981 annotated VMWEs) when compared to the French corpus in PARSEME, which contains $5,655$ annotated VMWEs. The solution is to employ two versions of the same training corpus: one with VMWE and another with nVMWE annotations only (Sec. 3.2–3.3).

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

[2] https://deep-sequoia.inria.fr/

[3] Laporte et al., 2008 present another openly available French corpus annotated for MWEs of various syntactic types, however their criteria for defining MWEhood are incompatible with ours, since semantic noncompositionality seems not to be a required property.

2.4. Corpus Splits

Tables 1 and 2 provide a summary of the overall statistics of the Sequoia and PARSEME datasets used in our experiments. We divide Sequoia into training, development, and test sets. Given that PARSEME is employed exclusively for training, we select only its training subset, excluding the Sequoia part.

Dataset	Sentences	VMWEs	nVMWEs
Sequoia	3,099	981	2,470
-train	1,921	610	1,536
-dev	283	80	240
-test	895	291	694

Table 1: Sequoia statistics.

Lang	Sentences	VMWEs
FR-train[4]	15,253	4,005
-GSD-UD-train	14,450	3,661
-ParTUT-UD-train	803	344
Multi-languages-train[5]	303,943	85,970

Table 2: Statistics of the PARSEME training corpus (with Sequoia excluded).

From these two datasets, we extract five distinct subsets for our experiments:

- allSequoia: Contains annotations for both VMWEs and nVMWEs, representing the entire Sequoia corpus, subdivided into train/dev/test sets, as shown in Table 1.
- V-Sequoia: Includes only VMWE annotations. Both the split and the number of sentences are identical to Sequoia, but nVMWE annotations are removed.
- N-Sequoia: Contains only nVMWE annotations (keeping the same sentences and split).
- F-PARSEME: Corresponds to the “FR-train" section in Table 2.
- M-PARSEME: Represents the “Multi-languages-train" section in Table 2.

Various combinations of these subsets are used in the experiments described in the following section.

3. System Description

To obtain an effective allMWE identifier for French, we use the general architecture of the MTLB-STRUCT system (Taslimipoor et al., 2020). In the PARSEME shared task edition 1.2 (Ramisch et al., 2020), this system achieved the highest F1-score across all 14 languages. It scored second for French, however, in the post-shared task experiments, it took the lead by leveraging a French-specific rather than a multilingual pre-trained model.

3.1. Single-task Models

We conducted the first experiment using the single-task version of MTLB-STRUCT with seven BERT-based case-sensitive models: the multilingual BERT (mBERT) (Devlin et al., 2019), CamemBERT (both base and large variants) (Martin et al., 2020), FlauBERT (both base and large variants) (Le et al., 2020), and XLM-RoBERTa (both base and large variants) (Conneau et al., 2020) for allMWE identification on allSequoia. An illustration of the model can be seen in Figure 1. Note that the model outputs different labels for their different MWE categories, but these differences are omitted in evaluation, which only considers the identification but not the categorisation task.

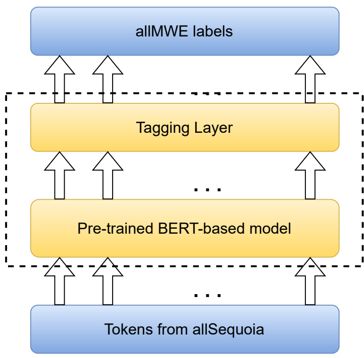

Figure 1: The overall architecture of the single-task model of MTLB-STRUCT. We use a BERT-based model to extract the encoded representations from the input tokens. All these representations are fed into a linear tagging layer to predict allMWE labels.

3.2. Union of Independent Models

In this second experiment, we selected three transformer-based models, CamemBERT-large, FlauBERT-large and XLM-RoBERTa-large, identified as the best single-task models in the first experiment (Sec. 3.1), as presented in the Table 3. Each of these three models keeps the same architecture from Fig. 1 but with a different corpus on input. More precisely, each model is fine-tuned into two separate instances: one focused on VMWEs, trained on V-Sequoia and either F-PARSEME (for monolingual training) or M-PARSEME (for multilingual training); and the other focused on nVMWEs

[4]The two training corpora were sourced from the official PARSEME GitLab repository, version 1.3, available at: https://gitlab.com/parseme/parseme_corpus_fr/-/tree/rp1.3

[5]Comprises 25 languages, with French included but Maltese omitted for the reasons detailed in Sec. 3.2.

trained on N-Sequoia. Since both the CamemBERT and FlauBERT models are monolingual, they are fine-tuned on F-PARSEME. XLM-RoBERTa, conversely, is multilingual, so we fine-tune it on M-PARSEME, which intentionally excludes the Maltese language since it is not part of the XLM-RoBERTa's training data.

We then combine the outputs of both instances through a union strategy. For instance, in *ils* [***ont***[***rendez-vous***]$_2$]$_1$ 'they have a meeting' , the VMWE and the nVMWE with indices 1 and 2 are identified by two independent models and both annotations are added to the final result.

3.3. Multi-task Learning

In our third experiment, we trained a single model to address two distinct tasks: VMWE and nVMWE identification. Both tasks utilize a shared text representation layer, which corresponds to the final layer of the pre-trained masked language model, and each task has its dedicated tagging layer. For training, we select CamemBERT-large, FlauBERT-large, XLM-RoBERTa-large, using pairs of the same datasets specific to each model and the union method introduced in the second experiment (Sec. 3.2). For instance, CamemBERT-large is trained on V-Sequoia, F-PARSEME and N-Sequoia, while XLM-RoBERTa-large uses V-Sequoia, M-PARSEME and N-Sequoia.

Figure 2 depicts the architecture of this multi-task system.

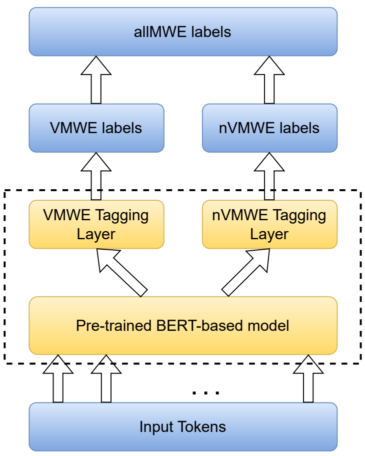

Figure 2: The architecture of the multi-task learning model with two branches built on top of BERT: one with a linear classifier layer for VMWE tagging and the other for nVMWE tagging. These branches are integrated into a unified allMWE tagging through a union approach.

In all the above experiments, we used batches of 10 sentences each and the AdamW optimizer with a learning rate of $3e-5$. All models were trained

Single-task models on allMWEs	Global MWE-based P	R	F1
mBERT	75.54	68.63	71.92
CamemBERT-base	73.20	73.20	73.20
CamemBERT-large	**79.87**	**79.08**	**79.47**
FlauBERT-base	74.42	73.20	73.81
FlauBERT-large	79.19	77.12	78.15
XLM-RoBERTa-base	75.96	71.24	73.52
XLM-RoBERTa-large	79.44	74.51	76.90

Table 3: Results of the single-task models on the allSequoia dev set (Sec. 3.1).

over 20 epochs, incorporating a linear schedule where 10% of the training steps were allocated to the warmup phase. For each model, we saved its iteration which achieved the best F1-score on the development set.

4. Results

We present the performance of all three methods on the allSequoia dev set in Tab. 3, 4 and 5. We use the global MWE-based precision, recall, and F1 measures (Savary et al., 2017), as used in the PARSEME shared tasks. This means that we consider each predicted MWE as correct only if all of its tokens match the ground-truth. Categorisation (into IRV, LVC.full, LVC.cause, VID, MVC or nVMWE) is disregarded because MWE identification is defined as highlighting occurrences of MWEs in the text, regardless of their category.

The union of FlauBERT-large model for VMWEs and the CamemBERT-large for nVMWEs (Tab. 4) surpasses the best single-task model (Tab. 3) by 3.24% F1 and the best multi-task model (Tab. 5) by 2.17% F1. Both of these are monolingual models and they exceed the performance of the two multilingual models (mBERT and XLM-RoBERTa) in VMWE and nVMWE identification. This trend has also been observed by Kurfalı, 2020; Nozza et al., 2020 in VMWE identification as well as in multiple other NLP tasks. Furthermore, Le et al., 2020 reported that FlauBERT demonstrates competitive performance in comparison to CamemBERT, with both pre-trained language models exhibiting complementary strengths and weaknesses. Regarding multilingual training, the union of the two XLM-RoBERTa-large models fined-tuned on VMWEs and nVMWEs (bottom right-hand cell in Tab. 4) performs better than XLM-RoBERTa-large fine-tuned on monolingual data by 1.41% (Tab. 3).

Table 6 displays the results of the best models for each of the three methods on the allSequoia test set. These findings further validate that the

Global MWE-based scores				Single-task VMWE models								
				CamemBERT-large			FlauBERT-large			XLM-RoBERTa-large		
				P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
				89.47	76.40	82.42	89.61	77.53	83.13	89.47	76.40	82.42
Single-task nVMWE models	CamemBERT-large	**P**	84.54	85.87			**85.92**			85.87		
		R	80.65		79.41			**79.74**			79.41	
		F1	82.55			82.51			**82.71**			82.51
	FlauBERT-large	**P**	78.83	81.54			81.88			81.54		
		R	80.65		79.41			79.74			79.41	
		F1	79.73			80.46			80.79			80.46
	XLM-RoBERTa-large	**P**	78.37	81.34			81.69			81.34		
		R	75.12		75.49			75.82			75.49	
		F1	76.71			78.31			78.64			78.31

Table 4: Results of combining independent models (Sec. 3.2) on allSequoia dev. In line 1 and column 1: scores on VMWE and nVMWE identification separately. In the remaining cells: results of the union.

Multi-task models	Global MWE-based		
	P	**R**	**F1**
CamemBERT-large	**83.51**	**77.78**	**80.54**
FlauBERT-large	78.48	77.45	77.96
XLM-RoBERTa-large	48.10	33.01	39.15

Table 5: Results of the multi-task models (Sec. 3.3) on the allSequoia dev set.

Best models	Global MWE-based		
	P	**R**	**F1**
Single-task: CamemBERT-large	79.41	81.42	80.40
Union: FlauBERT-large + CamemBERT-large	**85.27**	**82.84**	**84.04**
Multi-task: CamemBERT-large	79.70	81.73	80.70

Table 6: Results of the best models for each of the three methods on the allSequoia test set.

union method outperforms both the single-task and multi-task approaches, exhibiting an improvement of 3.64% F1 and 3.34% F1, respectively.

The results also reveal a complementary nature of the opportunities and challenges behind VMWEs vs. nVMWEs. Note that annotated VMWEs are 3 times more numerous than nVMWEs in the train sets $((4,005 + 610)/1,536$ in Tab. 1 and 2). Still CamemBERT-large gets slightly better results (Tab. 4) for nVMWEs (F1=82.55) than for VMWEs (F1=82.42). This should be due to the fact that nVMWEs are mostly continuous, have a fixed word order and their components rarely inflect (***dur/dure à la tâche*** (lit. 'hard.MASC/FEM in the task') 'hard-working'). Conversely, VMWEs are know to exhibit relatively large flexibility in terms of inflection, word order and discontinuity (*le* ***droit*** *qui a été* ***donné*** 'the right which was granted') and are therefore harder to identify (Constant et al., 2017).

5. Conclusions

This paper leveraged two datasets, one focused on verbal MWEs and the other on non-verbal MWEs. We employed a union approach, integrating two independent identification tasks for these two distinct types of MWEs. In our experiments, the proposed approach outperformed both the single-task and the multi-task methods on both the development and test sets in a cross-type dataset.

6. Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01). We extend our gratitude to Carlos Ramisch and Marie Candito for their insightful discussions and feedback throughout this project. We are also thankful to the anonymous reviewers for their valuable comments and to Saclay-IA for providing the GPU grant.

7. Bibliographical References

Andrei Avram, Verginica Barbu Mititelu, and Dumitru-Clementin Cercel. 2023. Romanian multiword expression detection using multilingual adversarial training and lateral inhibition. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 7–13, Dubrovnik, Croatia. Association for Computational Linguistics.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2nd edition, pages 267–292. CRC Press.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700, Dubrovnik, Croatia. Association for Computational Linguistics.

Genady Beryozkin, Yoel Drori, Oren Gilon, Tzvika Hartman, and Idan Szpektor. 2019. A joint named-entity recognizer for heterogeneous tagsets using a tag hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 140–150, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

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. Association for Computational Linguistics.

Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. 2018. Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2824–2829, Brussels, Belgium. Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

Murathan Kurfalı. 2020. TRAVIS at PARSEME shared task 2020: How good is (m)BERT at seeing the unseen? In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 136–141, online. Association for Computational Linguistics.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A french corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Damith Premasiri and Tharindu Ranasinghe. 2022. Bert (s) to detect multiword expressions. *arXiv preprint arXiv:2208.07832*.

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.

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.

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. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

8. Language Resource References

Candito, Marie and Constant, Mathieu and Ramisch, Carlos and Savary, Agata and Guillaume, Bruno and Parmentier, Yannick and Cordeiro, Silvio Ricardo. 2020. *A French corpus annotated for multiword expressions and named entities*. Polska Akademia Nauk. Instytut Podstaw Informatyki PAN. PID https://deep-sequoia.inria.fr.

Savary, Agata and Khelil, Chérifa Ben and Ramisch, Carlos and Giouli, Voula and Mititelu, Verginica Barbu and Mohamed, Najet Hadj and Krstev, Cvetana and Liebeskind, Chaya and Xu, Hongzhi and Stymne, Sara and others. 2023. *PARSEME corpus release 1.3*. PID http://hdl.handle.net/11372/LRT-5124.