A Survey of MWE Identification Experiments: The Devil is in the Details

Carlos Ramisch
Aix Marseille Univ, CNRS,
LIS, Marseille, France
first.last@lis-lab.fr

Abigail Walsh
ADAPT Centre
Dublin City University, Ireland
first.last@adaptcentre.ie

Thomas Blanchard
Centrale Marseille, Aix Marseille Univ,
CNRS, LIS, Marseille, France
first.last@centrale-marseille.fr

Shiva Taslimipoor
ALTA Institute
University of Cambridge, UK
first.last@cl.cam.ac.uk

Abstract

Multiword expression (MWE) identification has been the focus of numerous research papers, especially in the context of the DiMSUM and PARSEME Shared Tasks (STs). This survey analyses 40 MWE identification papers with experiments on data from these STs. We look at corpus selection, pre- and post-processing, MWE encoding, evaluation metrics, statistical significance, and error analyses. We find that these aspects are usually considered minor and/or omitted in the literature. However, they may considerably impact the results and the conclusions drawn from them. Therefore, we advocate for more systematic descriptions of experimental conditions to reduce the risk of misleading conclusions drawn from poorly designed experimental setup.

1 Introduction

The task of identifying Multiword Expressions (MWEs) in texts, as defined by Constant et al. (2017), can be modeled using several paradigms: syntactic parsing (Nagy T. and Vincze, 2014; Constant and Nivre, 2016), compositionality prediction of MWE candidates (Cook et al., 2008; Haagsma et al., 2020; Garcia et al., 2021), or sequence annotation (Constant et al., 2012; Schneider et al., 2014). The *sequence annotation* paradigm has been recently popularised by the DiMSUM shared task (Schneider et al., 2016), and by three editions of the PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2018a, 2020). Automatic methods designed to solve MWE identification (MWEI) seen as sequence annotation range from more traditional structured sequence tagging (Al Saied et al., 2017) to more free-form recent transformer-based token classification (Taslimipoor et al., 2020).

While the sequence annotation paradigm makes it possible to analyse various idiosyncratic aspects of MWEI in full text, empirical model evaluation is still a challenge. Our survey focuses on experimental design choices that are not always clearly described and discussed in the literature (§ 2).

The *data* used to learn, tune and evaluate MWEI models can influence a study's conclusions. For instance, the PARSEME corpora contain only verbal MWEs; evaluations based on it favour systems that can manage discontinuities (§ 3). Moreover, annotation schemes have different approaches to deal with discontinuity, variability, nesting, and overlaps, which are particular to MWEs. Traditionally, variations of BIO labelling were used to represent some of these aspects (Ramshaw and Marcus, 1995). PARSEME proposes a generic corpus format, taking these above-mentioned phenomena into account. However, the lack of standardisation with the selection and application of labelling schemes leaves the door open for system developers to decide how they want to model MWEs (§ 4).

Another important aspect of evaluation is the choice of the *evaluation metrics* used to assess system performance. While global exact and fuzzy metrics based on precision, recall and F-score are traditionally employed (Green et al., 2013; Constant and Nivre, 2016), they ignore a model's capability to deal with challenging traits like MWE discontinuity, seen/unseen MWEs, and their variability. From edition 1.1, PARSEME designed focused measures to evaluate for these aspects (Ramisch et al., 2018a). We discuss and compare these metrics, and the way systems report and discuss them in papers (§ 5). Furthermore, most related work does not assess whether a superior performance is likely due to chance, that is, whether observed performance differences are statistically significant. Thus, we propose a framework, a free implementation, and report significance analyses on the PARSEME 1.2 shared task results (§ 6). Finally,

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 106–120
May 6, 2023 ©2023 Association for Computational Linguistics

we look at whether and how MWEI papers report error analysis (§ 7).

In short, we shed some light on these apparently minor aspects which actually can have a great impact on results and conclusions. We look at corpus constitution and split, pre- and post-processing, MWE tagging, evaluation metrics, statistical significance of system comparison, and error analyses. We compare the experiments of 40 MWEI papers and discuss best practices in designing experimental setup and evaluation.

2 Survey scope

Our survey covers a total of 40 papers selected according to the following criteria:

- Available on the ACL Anthology, and
- Focus on MWEI as per Constant et al. (2017), report experimental results, and:
  - are shared task (ST) or system description papers submitted to DiMSUM (2016) or to one of the 3 editions of the PARSEME STs (2017, 2018, 2020), or
  - are published after the first ST (2016) and report experiments on the DiMSUM or PARSEME corpora.

Our selection is not exhaustive, disregarding influential MWEI articles with experiments on other corpora, e.g. Green et al. (2013); Constant and Nivre (2016), and recent papers on in-context compositionality prediction, e.g. Zeng and Bhat (2021); Tayyar Madabushi et al. (2022). To keep the number of papers manageable, we arbitrarily disregard papers published in venues absent from the ACL Anthology, e.g. Maldonado and QasemiZadeh (2018).[1] Moreover, our sample is certainly biased towards over-represented languages (e.g. English for DiMSUM) and MWE categories (e.g. verbal MWEs for PARSEME). Nonetheless, we believe that it represents a large fraction of work in the MWE annotation paradigm, and could be complemented by a larger survey in the future.

The goal of our survey is to base our discussion on quantitative data extracted from the papers. Thus, intuitions can be confirmed and concrete proposals can be made for clearly identified gray zones. Thus, for each of the surveyed papers, we systematically answered the following questions:

- Languages of the corpora,
- Corpus splits used (train/dev/test),
- MWE categories identified by the models,
- Corpus pre-processing and post-processing,
- MWE encoding and decoding, especially for classification and tagging models,
- Evaluation metrics reported,
- Statistical significance of model comparison,
- Aspects looked at in error analyses.

Hereafter, we distinguish the 27 papers submitted to one of the four recent shared tasks (ST papers) from the 9 standalone papers, not submitted to a shared task (non-ST papers). Moreover, 4 of the papers are overall shared task description papers. For the others, we will use the terms *systems* and *models* interchangeably, as these papers describe experiments using a system that relies on a proposed model or family of models.

3 Corpus constitution and selection

The first aspect that we look at is the corpora used in the MWEI experiments.

Languages The languages of the corpora used mostly depend on the data available for STs. The SEMEVAL DiMSUM ST provided corpora in English (Schneider et al., 2016), whereas PARSEME STs provided corpora for 18 languages in edition 1.0 (Savary et al., 2017), 19 languages in edition 1.1 (Ramisch et al., 2018a), and 14 languages in edition 1.2 (Ramisch et al., 2020). The DiMSUM corpus is based on Streusle (Schneider et al., 2014) and is annotated for most major MWE categories (nominal, verbal, adverbial, functional), but does not include category labels. The PARSEME corpora, on the other hand, contain fine-grained MWE category annotations, but only cover verbal MWEs.

Figure 1 shows the distribution of papers across the 24 languages considered by our paper sample. The reasons that lead to choosing a given corpus and/or set of languages in non-ST works are various: language diversity (Zampieri et al., 2019), corpus domain (Liu et al., 2021), and corpus quality and size (Pasquer et al., 2020b).

Conversely to the number of papers per language, we can also look at the number of languages addressed by each paper. Most papers (26 out of 40) address more than one language, with the following distribution: 1-3 languages: 15 papers; 4-10

[1] One exception was made for the SHOMA system paper, available only on arXiv, but listed in the PARSEME ST 1.1 paper and website (Taslimipoor and Rohanian, 2018).

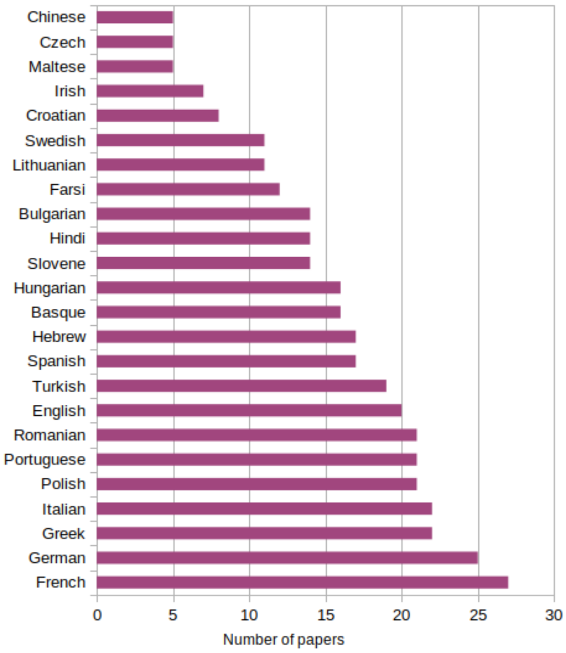

Figure 1: Number of papers per language.

languages: 6 papers, 11 languages or more: 19 papers. Among the 9 non-ST papers, 6 cover only one language, whereas 3 are multilingual.

Only 2 papers reported limiting their predictions to a subset of MWE categories (Foufi et al., 2017; Pasquer et al., 2018), otherwise the target MWE categories are by default all those present in the corpora. The prevalence of multilingual systems is probably due to the large amount of available corpora in the PARSEME collection, and to the use of largely language-independent methods based on these corpora. On the other hand, high cross-lingual variability is observed in most MWEI experiments. This can be due to the heterogeneity in the corpora and/or in the MWEs in each language (and how MWEI methods model them). Language-specific PARSEME corpus description papers not covered here can provide details, e.g. for Basque (Iñurrieta et al., 2018), Chinsese (Jiang et al., 2018), English (Walsh et al., 2018), Irish (Walsh et al., 2020), Italian (Monti and di Buono, 2019), Polish (Savary and Waszczuk, 2020), Portuguese (Ramisch et al., 2018b), Romanian (Barbu Mititelu et al., 2019), Turkish (Berk et al., 2018b; Ozturk et al., 2022), among others.

Domains Corpus domain may play an important role in MWEI. DiMSUM includes texts from 3 domains: web reviews, TED talk transcriptions, and tweets, and the ST paper analyses results per domain. One paper in our sample focuses on tweets, using this corpus (Zampieri et al., 2022b). PARSEME corpora contain mostly newspapers, with a few exceptions (e.g. French contains also Wikipedia, transcripts, and drug notices). One interesting case is that of the PARSEME Hungarian corpus, which contains barely any idioms, due to its highly specialised nature (law texts). Thus, systems using this corpus tend to report good performance, since this difficult category is under-represented (Savary et al., 2018). Liu et al. (2021) report cross-corpus (thus cross-domain) experiments using fine-tuned pre-trained language models with fine MWE+supersense labels.

Corpus and splits The four STs propose a corpus split: DiMSUM and PARSEME 1.0 randomly split the corpora into training and test sets. The PARSEME 1.1 and 1.2 STs add a third part: the development (dev) set (or validation set).[2] In the following discussion, we exclude the 4 general ST description papers, so our total is 36 system papers instead of 40.

External resources, rather than the training corpora, are used in 2 systems (Foufi et al., 2017; Colson, 2020), and 2 papers train models on the Streusle corpus and use PARSEME/DiMSUM only for test (Liu et al., 2021; Zampieri et al., 2022b), while the remaining 32 papers train their models on the PARSEME/DiMSUM training sets.

In DiMSUM, 4 papers mention a fixed train/dev split used to tune the systems, 1 paper mentions tuning on held-out data without further details (Kirilin et al., 2016) and two systems do not mention the issue (Björne and Salakoski, 2016; Scherbakov et al., 2016). For PARSEME 1.0, 3 papers use cross-validation to tune features (Al Saied et al., 2017; Maldonado et al., 2017; Boros et al., 2017), one system used a fixed train/dev split (Klyueva et al., 2017), and one system does not mention the issue (Simkó et al., 2017). For PARSEME 1.1, the languages with no dev set were usually tuned on the dev set of other languages, (Stodden et al., 2018; Taslimipoor and Rohanian, 2018, e.g.).

The use of standard corpus splits is a current practice in the NLP community. It ensures comparability across papers, e.g. to establish leaderboards and define state-of-the-art systems. However, standard splits have been criticised as their use may lead to unreplicable results (Gorman and Bedrick, 2019). Conversely, the use of multiple random splits also presents some disadvantages, leading to

[2] No dev in Hindi, English, and Lithuanian in edition 1.1.

over-estimated performances (Søgaard et al., 2021). As each splitting strategy has advantages and disadvantages, it is crucial to report how splits were obtained and why a given strategy was chosen.

Unseen MWEs The discussion in Ramisch et al. (2020) motivates the adoption of a less naturally distributed split in the PARSEME 1.2 ST corpora. The split is artificially biased to contain at least 100 unseen MWEs in the dev corpus, and 300 unseen MWEs in the test set.[3] While the results of this ST focus on generalisation, their definition of *unseen MWE* may require language-specific adaptations, e.g. Savary et al. (2019) argue that Basque canonical forms should include some morphological features. The use of automatically lemmatised corpora may also induce errors in the definition of unseen MWEs and thus influence the corpus splitting procedure.

The PARSEME 1.2 ST provided raw corpora not annotated for MWEs. However, there is no guarantee that MWEs in the dev and test corpora occur in the raw corpora. Moreover, pre-trained language models now popular in NLP are trained on corpora that are not always known or released, making it tricky to assess whether a given MWE is unseen, i.e. whether it has been observed in pre-training data. Future work on MWEI could propose strategies to address these challenges in assessing the generalisation of models.

Other corpora Finally, we mention corpora not included in our sample and not discussed here. Prior to DiMSUM and PARSEME, treebanks were often used to derive MWE annotations as a by-product. MWEI experiments were reported using the French Treebank (Constant et al., 2016), the Penn Treebank (Shigeto et al., 2013), the Arabic Treebank (Green et al., 2013), and the Szeged treebank (Vincze et al., 2013). For English, Wiki50 was one of the first full-text MWE-annotated corpora (Vincze et al., 2011), followed by the Streusle corpus (Schneider et al., 2014), of which the DiMSUM corpus is an extension.

Quite a few papers explore the task of distinguishing literal from idiomatic occurrences of pre-listed potentially idiomatic expressions. Corpora for this task include the English VNC-tokens corpus (Cook et al., 2008), the German preposition-noun-verb (Fritzinger et al., 2010) and infinitive-verb compounds corpus (Horbach et al., 2016), the English Magpie corpus (Haagsma et al., 2020), and the English, Portuguese and Galician Semeval 2022 task 2 corpora (Tayyar Madabushi et al., 2022). The PARSEME collection could be extended to include literal readings (Savary et al., 2019), and this was explored for German (Ehren et al., 2020).

4 Pre-processing and post-processing

Due to the variety of tagging methods, there is often need for a conversion step between the MWE labelling schemes used in the ST data and that preferred by models. This conversion step is reported to various degrees; omission of reporting can pose a problem for replicability.

BIO-style encoding and sequence tagging BIO-style encoding is frequently preferred for sequence tagging tasks. Common practice for both named-entity recognition (NER) systems and MWEI systems is to label tokens in the input data with one of these three labels, 'B' (begin), 'I' (inside), or 'O' (outside). While tolerably effective for capturing sequences of MWE tokens, it fails to capture discontinuous, nesting, or overlapping MWEs.

Schneider et al. (2014) experimented with 4 different tagging schemes based on BIO-style encoding; the 8 positional tags including BbIiOo_˜, where the lower-case counterparts 'o', 'b', and 'i' are additionally introduced for tagging nested MWEs, and '_' and '˜' to discriminate among strong (idiomatic) and weak (compositional) MWEs. Example 1 demonstrates how the nested expressions ***leaves a lot to be desired*** are annotated with this scheme. This tagset was adopted in DiMSUM (Schneider et al., 2016).

(1)	The	staff	**leaves**	*a*	*lot*	**to**	**be**	**desired**	.
	O	O	B	b	i_	I_	I_	I_	0

PARSEME annotation (Ramisch et al., 2018a) took a more generalised approach to annotating verbal MWEs in different languages. In their scheme, each MWE token takes a consecutive numerical index in the sentence and – for the initial token in an MWE – its category. A token can have multiple labels, separated with semicolons, if it belongs to more than one MWEs in the sentence. For example, the overlapping expressions ***did study*** *and* ***research*** would be annotated as in Example 2.

(2)	I	**did**	a	lot	of	**study**	and	**research**	.
	*	1:LVC;2:LVC	*	*	*	1	*	2	*

In this paper, we refer to the PARSEME label scheme as "CUPT", which is also the name of the

[3] Unseen MWE: multiset of lemmas not annotated in train.

a tabular data format in which the corpora are released (Ramisch et al., 2018a).[4]

4.1 From ST corpora to system data (pre-processing)

Pre-processing steps can include cleaning the data (e.g. removing long sentences, noisy tokens, or special characters). This step also includes any necessary conversion from ST format to whatever format is required for the prediction of MWEs. Of the 27 ST papers, 12 use some form of IO- or BIO-style encoding, while 7 of the 9 non-ST papers use a similar encoding. Among these 12+9 papers, 12 explicitly account for gaps in the MWE sequences, using a particular token to mark these (e.g. 'G' (gap), 'o').

Nested MWEs are handled with the *gappy 1-level* scheme developed by Schneider et al. (2014) or other variants (i.e. *bigappy-unicrossy* scheme developed by Berk et al. (2019)), however, overlapping MWEs such as the case in Example 2 above are only partially handled by *bigappy-unicrossy* and not handled by *gappy 1-level*. Such cases are rare in the corpora, and as such do not greatly impact the data. One paper (Walsh et al., 2022) attempts to address this problem of overlapping or shared-token expressions by modifying the BIO-style encoding, while another paper (Taslimipoor and Rohanian, 2018) appends multiple categories separated by a semicolon, similar to the CUPT-style encoding.

Other methods employed by systems include the extraction of dependency trees or other sub-graph constructions, or multisets of lemmas.[5] To capture MWE annotations; such methods make use of the tree structure to attend to discontinuities and nesting. Waszczuk (2018) describes a pre-processing step to reattach case dependents to their grandparents, so that MWEs of certain categories (e.g. inherently adpositional verbs) are connected. To handle overlaps, they train one model per MWE category and combine their outputs at post-processing.[6]

Most papers do not explicitly mention their strategy to deal with overlapping MWEs. When mentioned, overlapping MWE annotations are either ignored (Zampieri et al., 2022a), duplicated into separate sentences (Zampieri et al., 2018), or handled by the tagging scheme (Yirmibeşoğlu and Güngör, 2020).

4.2 From system output to ST evaluation (post-processing)

Post-processing steps may require conversion of the labels used during prediction into the ST format to allow for evaluation and comparison with other systems in the ST. 13 ST papers and 5 non-ST papers explicitly describe the post-processing steps taken to perform this conversion. 5 ST papers and 1 non-ST paper did not require this conversion step, with the remaining 9 ST papers and 3 non-ST papers not reporting the methodology applied for this step; this may pose a problem for reproducibility. We explore some of the common methods of label processing below.

Conditional random fields Given their ability to observe relationships between labels in a sequence and consider future relationships when observing a pattern, conditional random fields (CRFs) have seen successful application in sequence-labelling tasks such as named-entity recognition, POS-tagging, and MWEI. One of the advantages of CRFs is that they can be applied to both feature-based (symbolic) and continuous models, as an extra layer on top of standard neural architectures (LSTMs or pre-trained transformers). However, since CRFs in neural models are trained using back-propagation, there is no guarantee that they will generate valid label sequences, potentially requiring heuristics to fix the label sequence in converting BIO-like labels into MWE annotations. In our sample, 8 out of 36 system papers report using CRF to predict labels.

BIO-style conversion Reversing the conversion from BIO-style to ST format requires making decisions regarding the grouping of predicted labels, i.e. to which MWE should each predicted label be assigned? With IO-style or binary encoding, grouping continuous predicted MWE labels together may be straightforward, although this can be more complicated when MWEs directly follow each other, with no gaps in between. A BIO-style scheme for predicting labels addresses this problem, as I-labelled tokens can be assumed to belong with the preceding B-labelled token. However, there remains the issue of how to assign I-labelled tokens that may belong to one of several preceding B-labelled tokens, as is the case with nested or overlapping MWEs. There is also the question of how

[4] https://multiword.sourceforge.net/cupt-format/

[5] Multiset: set allowing multiple instances of each element.

[6] This does not handle same-category overlaps, though.

to assign standalone I-labelled tokens. In our sample, a heuristic algorithm is frequently applied (7 of 36 papers), with tokens of the same predicted category grouped together, and standalone I-labelled tokens either filtered out or assigned to a new MWE group. A greedy-matching algorithm can be used to generate deep stacks of nested MWEs with gaps (Scherbakov et al., 2016). Alternatively, Viterbi decoding can be used to prevent invalid BIO sequences from being generated (Liu et al., 2021).

Dependency trees In systems where the MWEs are labelled through predicted dependency trees, conversion to CUPT format is relatively straightforward,[7] with all elements of an MWE assumed to be nodes in the same subtree. Waszczuk (2018) highlights the issue of segmenting MWEs within a dependency tree: their heuristic algorithm groups MWEs of the same category within the subtree. If a group contained two or more verbs, it was divided into the corresponding number of MWEs. Gombert and Bartsch (2020) use dependency trees to group MWEs as a post-processing step.

5 Evaluation metrics

Evaluation strategies for structured tagging tasks are less straightforward than that of classification. System performance is determined based on the correct prediction for sets of labels (e.g. for all tokens in ***raining cats and dogs***). The strict matching between the labels of all components of an MWE in the gold data and its correspondents in the predicted data is measured using MWE-based precision, recall and F1 measures in PARSEME. The same measures are referred to as *exact match* in DiMSUM. Nevertheless, in order to reward systems for partially correct predictions, PARSEME uses token-based precision, recall and F1 measures and DiMSUM (Schneider et al., 2016) introduces link-based measures which are computed based on links (correct use of tags) between consecutive tokens in an expression.[8] 20 out of 21 PARSEME ST papers focused on reporting MWE-based F1 (with the focus of 6 PARSEME 1.2 papers being on unseen expressions only), and only one system (Pasquer et al., 2020a) which was designed for predicting seen MWEs reported MWE-based precision on unseen expressions only.[9] All 6 DiMSUM papers reported linked-based F1, with four of them reporting P and R as well.

Standard machine learning approaches optimize systems towards the best F1-measure. Depending on the target task, precision or recall might be more beneficial. Gombert and Bartsch (2020) boost MWE-based precision by modifying the output of their transformer-based system by filtering out the predictions that involve tokens that are not connected in dependency trees.

Focused measures Introducing focused measures in PARSEME 1.0, 1.1, 1.2 developed over time, motivated by related work. For example, Al Saied et al. (2018) showed the negative correlation between system performance and the number of unseen MWEs.

Seen/Unseen Identifying unseen expressions became the focus of PARSEME 1.2, resulting in interesting insights. Word embeddings trained on extra unannotated data (Yirmibeşoğlu and Güngör, 2020) proved successful in detecting unseen expressions and not surprisingly pre-trained language models (Taslimipoor et al., 2020; Kurfalı, 2020) were the best. While rule-based syntactic pattern-matching based on association measures (Pasquer et al., 2020a) failed at capturing unseen expressions, it showed promising results in detecting various forms of a seen MWE. All 6 PARSEME edition 1.2 papers and 3 papers from previous editions focused on reporting performance on unseen expressions

Diversity Evaluating a system's capability to identify variants of existent MWEs is possible thanks to one of PARSEME's additional focused measures. Only two PARSEME papers reported these focused measures. A more recent study by Lion-Bouton et al. (2022) expanded on the above analysis, and proposed two new measures, namely richness and evenness, for evaluating diversity in models' predictions. In the experiments on MWE identification with PARSEME datasets, they showed that F1-measure performance roughly correlates with the richness of models' predictions but not with their evenness.

[7] No DiMSUM ST paper applied this method.

[8] The linked-based measures only work for DiMSUM data, where the MWE tags exactly follow their tagging scheme in which there is no big O label in between MWE components and no single-token MWE. Single-token MWEs are allowed in PARSEME to account for tokenisation problems, e.g. Spanish ***abstenerse*** (lit. 'abstain oneself'), which occurs as such although ideally it should be tokenised as ***abstener␣se***.

[9] 4 PARSEME papers did not report precision and recall, but the reports of all PARSEME evaluation measures for all systems are available on the corresponding websites.

Discontinuity MWEs pose a unique challenge to NLP due to the discontinuity that often occurs between the words that make up the expression. This challenge distinguishes MWEs from other similar phrasal structures, such as keyphrases or multiword named entities, making their processing more difficult. PARSEME's STs introduce additional evaluation measures focused on discontinuity. Five out of 27 studies on PARSEME datasets reported results on discontinuous MWEs separately. Most of them use dependency parse grammatical structure to identify the relationships between constituents of an MWE (Waszczuk, 2018; Moreau et al., 2018). Rohanian et al. (2019) propose a model which benefits from combining attention mechanism with graph convolutional network to improve identifying discontinuous MWEs. We believe that these focused measures can be generalized to other NLP tasks to alleviate more thorough evaluation.

6 Hypothesis testing and significance

System (or model) comparison has been one of the most important methodological tools, driving progress in NLP for the last 30 years. In this paradigm, we conclude that system A is superior to system B if it obtains a better evaluation score than system B on some given test set(s). The previous sections discussed data (§ 3) and evaluation metrics (§ 5) usually employed in the context of MWE identification. However, several papers throughout the decades have shown that there is a probability that this conclusion is false in general, because the test set is a limited-size sample of the actual language (text) on which the systems will be applied in production (Yeh, 2000; Berg-Kirkpatrick et al., 2012; Dror et al., 2018). Fortunately, statistic tools can estimate this probability given the characteristics of the test set, and in particular its size.

In a nutshell, *hypothesis testing* can be used to assume no difference between two systems as the null hypothesis to reject. Then, a statistical method can be used to estimate the *p-value*, that is, the probability of type-I error.[10] In other words, a p-value estimates the probability of wrongly rejecting the null hypothesis (i.e. concluding that the systems are indeed different) when there is actually no difference between the systems. One can consider that the difference between the systems is *statistically significant* if the p-value is lower than a confidence threshold (usually set to 0.05).Then, if we claim that system A is superior to B, there is a probability of at most 5% that this conclusion is wrong.

In MWE identification, comparison is based on precision, recall, and F-score, which prevents the use of simple parametric tests like Student's t-test (Yeh, 2000). Thus, non-parametric tests such as the bootstrap (Berg-Kirkpatrick et al., 2012) should be employed. However, our survey showed that p-values were reported for only 2 papers. The DiMSUM ST paper compares system predictions using the non-parametric McNemar's test. The official ST ranking shows three systems tied in first position since their results are not significantly different from each other. However, as discussed by Dror et al. (2018), this test is not very powerful, and this result may fall into type-II error, that is, not being able to reject the null hypothesis when it is actually true. Then, Hosseini et al. (2016) report significance using randomized approximation, which is a more appropriate test in this case since it is both non-parametric and powerful.

Significance analysis Given the lack of systematic significance analysis in our paper sample, we propose a new tool and a first analysis of the system predictions of the PARSEME ST 1.2.

We have re-implemented the ST evaluation script using the `cupt` library.[11] On top of it, we have added an option to compare two systems, estimating the p-value of their difference for all calculated metrics (global and phenomenon specific). P-values are estimated using the bootstrap method which resamples k=10,000 new test sets with replacement from the original test set.[12] The p-value is estimated as the relative frequency of extreme results, that is, the proportion of samples for which the difference between the system scores is at least twice as large as the difference observed on the whole test set. Our tool is available at `https://gitlab.com/parseme/significance`.

In practice, significance is more relevant when the differences between systems are small and/or test sets are small. This is the case for many languages and system pairs in the 1.2 edition of the PARSEME ST.[13] Our analyses were performed on each language individually, running the signifi-

[10] Confidence intervals are an alternative, but p-value seems to be preferred in the NLP literature.

[11] `https://gitlab.com/parseme/cupt-lib`

[12] Our implementation is based on the pseudo-code provided in Berg-Kirkpatrick et al. (2012). We resample test sizes with the same number of sentences as the original one.

[13] `https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2/system-results`

Systems	Open track						Closed track
		MTLB-STRUCT	TRAVIS-multi	HMSid	Seen2Unseen	FipsCo	Seen2Seen
	F1	**0.4309**	**0.3776**	**0.3739**	**0.2483**	**0.1883**	**0.0354**
TRAVIS-mono	**0.4837**	0.03	0.0	0.0	0.0	0.0	-
MTLB-STRUCT	**0.4309**		0.012	0.015	0.0	0.0	-
TRAVIS-multi	**0.3776**			0.447	0.0	0.0	-
HMSid	**0.3739**				0.0	0.0	-
Seen2Unseen	**0.2483**					0.01	-
ERMI	**0.252**	-	-	-	-	-	0.0

Table 1: p-value of the *MWE-based F1* score for results on *Unseen-in-train* MWEs in French. Non-significant results for $\alpha = 0.05$ are underlined.

Systems	Open track				Closed track
		TRAVIS-multi	Seen2Unseen	TRAVIS-mono	ERMI
	F1	**0.6911**	**0.6892**	**0.6709**	**0.6308**
MTLB-STRUCT	**0.7158**	0.025	0.038	0.0	-
TRAVIS-multi	**0.6911**		0.464	0.081	-
Seen2Unseen	**0.6892**			0.103	-
Seen2Seen	**0.7068**	-	-	-	0.0

Table 2: p-value of the the *MWE-based F1* score for results on *global* MWEs in Swedish. Non-significant results for $\alpha = 0.05$ are underlined.

cance tool on all possible system pairs submitted to the same track (open, closed). For each of these pairs, we calculated the 3 p-values (precision, recall, F-score) for each of the evaluation metrics (MWE-based, Unseen-in-train, etc.)

The results table contains 2,728 p-values in total, which we cannot exhaustively present here. Thus, only a sample of the results is gathered here, trying to cover test sets of different sizes, since sample size is known to influence the significance of results. In Table 1, we observe the behavior of the p-value between the unseen-in-train F-scores of systems, and on a language that had a large dataset, that is, French (1,359 MWEs). Results show that on the represented metric, (here, global MWE-based F-score), most systems are significantly different, with a p-value lower than the 0.05 threshold. However, the difference between Travis-multi and HMSid is not deemed significant, so we cannot conclude that the former is better than the latter.

In Table 2, we look at the global MWE score for another language, Swedish, which test set is much smaller (969 MWEs). Here, we observe that Seen2Unseen, Travis-multi and Travis-mono are not significantly different from each other, although some absolute differences in F-scores are larger than for French. Out of all comparisons made, 783 p-values fall above the 0.05 threshold, so potentially up to 29% of the system predictions are not significantly different from each other. Appendix A presents further examples of significance values.

Our analysis is not exhaustive, and other MWE identification papers did report significance in the past, e.g. Constant et al. (2016). Nonetheless, our analyses show that this methodological precaution is mostly neglected in the field. We hope that our survey can contribute to raising awareness on this issue for future publications.

7 Error analysis

Error analysis, when conducted properly, can help to identify particularly challenging cases for MWEI, whether because of intrinsic properties of the MWEs, the dataset, or the language, or because of weaknesses in the model, as demonstrated by the survey. 33 out of 40 papers carried out some degree of error analysis; certain properties of MWEs, languages, or corpus phenomena are investigated in particular. Comparisons of model performance across languages (sometimes including examination of the linguistic features or MWE categories particular to that language) are carried out in 11 papers (Simkó et al., 2017; Boros et al., 2017), while reporting the model results across the focused measures highlighted in § 5 are carried out in 15 papers. The PARSEME 1.1 and 1.2 papers usually report and discuss focused metrics, as these metrics were implemented in the ST evaluation scripts (Waszczuk, 2018; Berk et al., 2018a).

Analyses tended to take one of two forms: example-based analysis reporting individual instances where the model performed better or worse than usual (Klyueva et al., 2017; Walsh et al., 2022), and automatic metrics aggregated across particular properties or phenomena. Among the focused metrics, some papers pay special attention to discontinuities (Björne and Salakoski, 2016; Moreau et al., 2018; Berk et al., 2018a; Rohanian et al., 2019) and seen/unseen MWEs (Maldonado et al., 2017; Zampieri et al., 2018; Taslimipoor and Rohanian, 2018). Some studies analyse the model's features and modules via ablation experiments (Scherbakov et al., 2016; Tang et al., 2016; Stodden et al., 2018; Pasquer et al., 2020a). Cross-language performance was also discussed, especially in the first editions of PARSEME (Simkó et al., 2017; Boros et al., 2017). More original aspects discussed less often include POS sequence patterns (Cordeiro et al., 2016; Tang et al., 2016), the use of external lexicons (Kirilin et al., 2016), syntactic dependencies between components (Pasquer et al., 2018; Moreau et al., 2018), pre-trained embedding representations (Zampieri et al., 2019), and tagging schemes, as discussed in § 4 (Zampieri et al., 2022b).

In short, although quite heterogeneous, error analyses are usually present in MWEI papers, and tend to uncover interesting research questions for future work.

8 Conclusions and open issues

This paper provides a survey on experimental conditions reported and discussed in recent works on identifying MWEs. Analysis of the details of methodological choices by authors helps researchers and practitioners understand the performance of different models and identify areas for improvement. While STs help benchmark many of such experimental designs and evaluation criteria, tight schedules and less attention to task description papers cause many such details still to be neglected.

This survey focuses on two shared tasks on identifying MWEs and consequent systems designed based on their task definitions, datasets, and evaluations. As common-sense best practices, we advocate reporting on experimental choices such as corpus constitutions and selections, pre- and post-processing, evaluation metrics and significance testing of performance, and some error analysis performed in related work. We encourage the introduction of focused measures that facilitate error analysis, as is done in the later PARSEME editions. For statistical significance testing, we propose a tool that can automatically run such analyses on standard PARSEME-formatted predictions.

However, our analyses are not exhaustive and there are other methodological details to be discussed in the papers. One aspect that we only skim over in our discussion of the use of dev sets is hyper-parameter tuning. Which hyper-parameters were tuned, on which selection of the datasets, and what strategy (if any) was taken (e.g. grid search, random, etc.) are aspects that only very few of the papers clearly reported, and future work should encourage authors to report these.

Currently, most evaluation techniques are automatic. One open issue is whether there is a place in which manual evaluation of detected MWEs should be performed, (e.g. in the context of downstream tasks). New evaluation protocols can be considered in the future, towards answering other questions, e.g. whether some categories of MWEs are more important than others. We expect that our survey can contribute to the gradual adoption of methodological standards and best practices, both for shared tasks and independent research work in our community.

Acknowledgements

This work has been funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01), and by the Irish Research Council and CampusFrance, via an Ulysses PHC grant (45358QD). We thank Marius Thorre and Silvio Cordeiro for their contribution to the development of the significance software, Agata Savary for her feedback on an initial version of this paper, and the anonymous reviewers for their valuable suggestions.

References

Hazem Al Saied, Marie Candito, and Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Language Science Press., Berlin.

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

Verginica Barbu Mititelu, Mihaela Cristescu, and Mihaela Onofrei. 2019. The Romanian corpus annotated with verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 13–21, Florence, Italy. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018a. Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 248–253, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gözde Berk, Berna Erden, and Tunga Güngör. 2019. Representing overlaps in sequence labeling tasks with a novel tagging scheme: bigappy-unicrossy. In *Computational Linguistics and Intelligent Text Processing*, pages 622–635. Springer International Publishing.

Gözde Berk, Berna Erden, and Tunga Güngör. 2018b. Turkish verbal multiword expressions corpus. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Jari Björne and Tapio Salakoski. 2016. UTU at SemEval-2016 task 10: Binary classification for expression detection (BCED). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 925–930, San Diego, California. Association for Computational Linguistics.

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

Jean-Pierre Colson. 2020. HMSid and HMSid2 at PARSEME shared task 2020: Computational corpus linguistics and unseen-in-training MWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 119–123, online. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, 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.

Matthieu Constant, Joseph Le Roux, and Nadi Tomeh. 2016. Deep lexical segmentation and syntactic parsing in the easy-first dependency framework. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1095–1101, San Diego, California. Association for Computational Linguistics.

Matthieu Constant and Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 161–171, Berlin, Germany. Association for Computational Linguistics.

Matthieu Constant, Anthony Sigogne, and Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 204–212, Jeju Island, Korea. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop on Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. 2016. UFRGS&LIF at SemEval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 910–917, San Diego, California. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220, Online. Association for Computational Linguistics.

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

Fabienne Fritzinger, Marion Weller, and Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Sebastian Gombert and Sabine Bartsch. 2020. MultiVitaminBooster at PARSEME shared task 2020: Combining window- and dependency-based features with multilingual contextualised word embeddings for VMWE detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 149–155, online. Association for Computational Linguistics.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Andrea Horbach, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner, and Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 836–841, Portorož, Slovenia. European Language Resources Association (ELRA).

Mohammad Javad Hosseini, Noah A. Smith, and Su-In Lee. 2016. UW-CSE at SemEval-2016 task 10: Detecting multiword expressions and supersenses using double-chained conditional random fields. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 931–936, San Diego, California. Association for Computational Linguistics.

Uxoa Iñurrieta, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar, and Iñaki Alegria. 2018. Verbal multiword expressions in Basque corpora. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 86–95, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Menghan Jiang, Natalia Klyueva, Hongzhi Xu, and Chu-Ren Huang. 2018. Annotating Chinese light verb constructions according to PARSEME guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angelika Kirilin, Felix Krauss, and Yannick Versley. 2016. ICL-HD at SemEval-2016 task 10: Improving the detection of minimal semantic units and their meanings with an ontology and word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 937–945, San Diego, California. Association for Computational Linguistics.

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. 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.

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. Evaluating diversity of multiword expressions in annotated text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nelson F. Liu, Daniel Hershcovich, Michael Kranzlein, and Nathan Schneider. 2021. Lexical semantic recognition. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 49–56, Online. Association for Computational Linguistics.

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

Alfredo Maldonado and Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. 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 149–175. Language Science Press., Berlin.

Johanna Monti and Maria Pia di Buono. 2019. PARSEME-It: an Italian corpus annotated with verbal multiword expressions. *IJCoL*, 5:61–93.

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. Association for Computational Linguistics.

István Nagy T. and Veronika Vincze. 2014. VPCTagger: Detecting verb-particle constructions with syntax-based methods. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 17–25, Gothenburg, Sweden. Association for Computational Linguistics.

Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the PARSEME Turkish corpus of verbal multiword expressions. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 100–104, Marseille, France. European Language Resources Association.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2018. If you've seen some, you've seen them all: Identifying variants of multiword expressions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2582–2594, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020a. Seen2Unseen at PARSEME shared task 2020: All roads do not lead to unseen verb-noun VMWEs. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 124–129, online. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2020b. 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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018a. 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, Renata Ramisch, Leonardo Zilio, Aline Villavicencio, and Silvio Cordeiro. 2018b. A corpus study of verbal multiword expressions in Brazilian Portuguese. In *Computational Processing of the Portuguese Language 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings*, Lecture Notes in Artificial Intelligence, Cham, Switzerland. Springer International Publishing. `https://link.springer.com/chapter/10.1007/978-3-319-99722-3_3`.

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.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. 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 2692–2698, Minneapolis, Minnesota. Association for Computational Linguistics.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Silvio Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. 2019. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical 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.

Agata Savary and Jakub Waszczuk. 2020. Polish corpus of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 32–43, online. Association for Computational Linguistics.

Andreas Scherbakov, Ekaterina Vylomova, Fei Liu, and Timothy Baldwin. 2016. VectorWeavers at SemEval-2016 task 10: From incremental meaning to semantic unit (phrase by phrase). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 946–952, San Diego, California. Association for Computational Linguistics.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

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.

Yutaro Shigeto, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kose, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung, and Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 139–144, Atlanta, Georgia, USA. Association for Computational Linguistics.

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

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Xin Tang, Fei Li, and Donghong Ji. 2016. WHUNlp at SemEval-2016 task DiMSUM: A pilot study in detecting minimal semantic units and their meanings using supervised models. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 918–924, San Diego, California. 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.

Shiva Taslimipoor and Omid Rohanian. 2018. SHOMA at Parseme shared task on automatic identification of VMWEs: Neural multiword expression tagging with high generalisation. *arXiv preprint arXiv:1809.03056*.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295, Hissar, Bulgaria. Association for Computational Linguistics.

Veronika Vincze, István Nagy T., and Richárd Farkas. 2013. Identifying English and Hungarian light verb constructions: A contrastive approach. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Sofia, Bulgaria. Association for Computational Linguistics.

Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider, and Clarissa Somers. 2018. Constructing an annotated corpus of verbal MWEs for English. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 193–200, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online. Association for Computational Linguistics.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2022. A BERT's eye view: Identification of Irish multiword expressions using pre-trained language models.

In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 89–99, Marseille, France. European Language Resources Association.

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. Association for Computational Linguistics.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Zeynep Yirmibeşoğlu and Tunga Güngör. 2020. ERMI at PARSEME shared task 2020: Embedding-rich multiword expression identification. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 130–135, online. Association for Computational Linguistics.

Nicolas Zampieri, Carlos Ramisch, and Geraldine Damnati. 2019. The impact of word representations on sequential neural MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 169–175, Florence, Italy. Association for Computational Linguistics.

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. 2022a. Identification des expressions polylexicales dans les tweets (identification of multiword expressions in tweets). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 365–373, Avignon, France. ATALA.

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. 2022b. Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202–210, Marseille, France. European Language Resources Association.

Nicolas Zampieri, Manon Scholivet, Carlos Ramisch, and Benoit Favre. 2018. Veyn at PARSEME shared task 2018: Recurrent neural networks for VMWE identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 290–296, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Further significance analyses

Here, we present two further samples of our significance tool output for the results of the PARSEME 1.2 shared task. Table 3 shows the p-values for French considering the global MWE score (the main paper text shows the analysis for *Unseen-in-train* MWEs in Table 1). In Table 4 we show the analysis for a language with a very small test set, Irish, containing 436 annotated MWEs. In both cases, we observe small F-score variations between systems that are not deemed significant. Thus, one cannot say that Travis-multi (F1=0.7689) is better than Seen2Unseen (F1=0.7677) for the French global MWE measure. The same applies for the difference between Seen2Unseen (F1=0.3058) and MTLB-struct (F1=0.3007) for the Irish global MWE-based score.

Systems		Open track					Closed track
		MTLB-STRUCT	TRAVIS-multi	Seen2Unseen	HMSid	FipsCo	ERMI
	F1	**0.7942**	**0.7689**	**0.7677**	**0.6579**	**0.5067**	**0.6141**
TRAVIS-mono	**0.826**	0.0	0.0	0.0	0.0	0.0	-
MTLB-STRUCT	**0.7942**		0.003	0.009	0.0	0.0	-
TRAVIS-multi	**0.7689**			0.47	0.0	0.0	-
Seen2Unseen	**0.7677**				0.0	0.0	-
HMSid	**0.6579**					0.0	-
Seen2Seen	**0.7863**	-	-	-	-	-	0.0

Table 3: P-value of the the *MWE-based F1* score for results on *global* MWEs in French. Non-significant results for $\alpha = 0.05$ are underlined.

Systems		Open track		Closed track
		MTLB-STRUCT	TRAVIS-multi	ERMI
	F1	**0.3007**	**0.0717**	**0.1958**
Seen2Unseen	**0.3058**	0.423	0.0	-
MTLB-STRUCT	**0.3007**		0.0	-
Seen2Seen	**0.2689**	-	-	0.004

Table 4: P-values of the *MWE-based F1* score for results on *global* MWEs in Irish. Non-significant results for $\alpha = 0.05$ are underlined.