PMI MWE Scorer at PARSEME 2.0 Subtask 1: identifying multi-word expressions using pointwise mutual information and universal dependencies

Syntax-Aware PMI for Multi-Word Expression Identification Using Universal Dependencies

Anna Bogdanova and Ileana Bucur
Eberhard Karls Universität Tübingen
Tübingen, Germany

Abstract

Multi-word expressions (MWEs) remain a challenge for NLP systems due to their syntactic variability and non-compositional semantics, that is why this issue was proposed as shared task within Unidive organization. With increasing popularity of large language models (LLM), it is important to continue researching alternative solutions. One of the classical approaches for identifying MWEs is calculating Pointwise Mutual Information (PMI), but this is a purely statistical approach that cannot reveal the links between words in natural text. To fix this issue, we propose this paper with a simple syntax-aware PMI method that leverages Universal Dependency (UD) trees (Nivre et al., 2016) to model co-occurrence between syntactically related words. By computing PMI over dependency-linked word pairs and aggregating these scores, we aim to improve surface-based methods. Unlike expectations, our experiment shows that the classical statistical approach gets better results in partially identifying MWEs. Still, this approach is aimed to find a balance between lightweight calculations as opposed to LLMs and precision in results.

1 Introduction

Multi-word expressions (MWEs), such as *in spite of*, *take place*, or *make sense*, are pervasive in natural language and pose long-standing challenges for NLP systems. Their meaning does not equal to the sum of meanings of its elements, and their surface forms may vary syntactically, making them difficult to identify using purely lexical or contextual cues.

PMI is calculated for each pair of adjacent words, which is not suitable for this kind of task because parts of multiword expressions can be located far away from each other in a sentence. It is possible to use a skip-gram, but if we have a look at an example sentence:

(1) *She turned the proposal that had been debated for months by several committees down.*

we would see that the phrasal verb and its particle are located 11 words away from each other, and classical PMI would not identify an MWE in this example sentence.

To deal with this issue as part of our participation in the shared task, we propose using dependency analysis based on Universal Dependencies (UD) as a basis for calculating the PMI score. As opposed to classical PMI, two words are considered a pair not if they are adjacent in a sentence, but only if they are connected by an edge of any type. In this scenario, turned and down form a pair because the particle depends on its verb.

UD provides cross-linguistically consistent dependency relations that explicitly encode syntactic relationships independent of surface adjacency. Furthermore, UD is language-independent. To mitigate unnecessary variability in PMI pairs for languages with rich morphology, we utilize UD lemmas rather than surface forms. In conclusion, our research provides insights into

- Analysis of the impact of UD on MWE detection precision
- Computation complexity difference between adjacency based PMI and UD-based PMI scoring.

2 Method

Our approach extends traditional PMI-based MWE identification by redefining word co-occurrence in terms of syntactic dependency relations rather than linear adjacency. While classical PMI relies on surface proximity to approximate lexical association, such an assumption is often violated by multi-word expressions whose components may be syntactically related but linearly distant. To address this limitation, we exploit Universal Dependency (UD)

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 165–169

parsers to define co-occurrence over syntactic structure instead of surface order, while preserving the lightweight and language-independent nature of PMI-based methods.

2.1 Pointwise Mutual Information

Pointwise Mutual Information (PMI) is a widely used association measure that quantifies the strength of co-occurrence between two lexical items relative to their independent occurrence probabilities. Formally, PMI between two words w_1 and w_2 is defined as:

$$PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1)\,P(w_2)}$$

where $P(w_1, w_2)$ denotes the joint probability of the two words co-occurring under a given definition of co-occurrence, and $P(w_1)$ and $P(w_2)$ are their individual probabilities.

To mitigate sparsity and avoid undefined values caused by zero-frequency events (Jurafsky and Martin, 2025), we apply Laplace smoothing to all probability estimates, resulting in the following formulation:

$$\log \frac{P(w_1, w_2) + 1}{(P(w_1) + 1)(P(w_2) + 1)}$$

All counts are computed over lemmatized tokens in order to reduce morphological variation across languages. Punctuation symbols and stop-words are excluded from the computation, as they are unlikely to contribute to the identification of meaningful lexicalized expressions. We assume that few or no MWEs contain stop-words; thus, their exclusion results in a negligible number of false negatives.

2.2 Surface PMI Baseline

As a baseline, we compute PMI for adjacent word pairs occurring within a fixed window of size two. Under this setting, two lemmas are considered to co-occur if they appear consecutively in the corpus, and wheneer two lemmas appear together, the counter increments.

2.3 Dependency-Based PMI

In the dependency-based variant, the PMI formula remains unchanged, but the definition of co-occurrence is altered. Two lemmas are considered to form a co-occurring pair if they are directly connected by an edge in the Universal Dependency representation of a sentence, regardless of their linear distance.

In this manner, syntactically related words may form a pair even when separated by multiple intervening tokens. All dependency relations are treated uniformly, without restricting the computation to specific relation types.

Both the proposed method and classical PMI approaches are language-agnostic, but the dependency-based approach is considered to be more linguistically aware.

2.4 MWE Identification

To identify MWEs, all word pairs are ranked according to their PMI scores. Pairs exceeding a predefined threshold are marked as potential MWEs. Determining an optimal threshold is a significant challenge, as PMI values are not directly comparable across languages or datasets and are sensitive to corpus size and distributional properties. Rather than imposing a single threshold, we report results for multiple percentile thresholds.

3 Experimental Setup

3.1 Dataset

We evaluate our approach on a dataset provided by shared task organisers (Savary et al., 2023).

3.2 Evaluation Metrics

To evaluate our system, we adopted the official metrics of the original shared task.

The performance is assessed in terms of precision, recall, and F1-score at both the MWE level (exact match) and the token level (allowing partial matches). Diversity is measured through richness, Shannon evenness, and Shannon–Weaver entropy to capture the variety and balance of the identified MWE types. For each language, a distinct set of values is provided. In this study, we focus on the macro-averaged F1-score as our primary evaluation metric.

3.3 Baselines

We compare two PMI-based baselines that differ exclusively in how word co-occurrence is defined, while sharing the same scoring function, preprocessing pipeline, and thresholding strategy. This design allows us to isolate the effect of syntactic information on MWE identification, independently of other modeling choices.

4 Results and Analysis

The following are two tables Table 1 and Table 2 with F1 scores for every language for different thresholds. The first table contains scores for PMI approach enchanced with UD, and the second table presents scores for classical PMI.

Lang	Type	50	75	90	95	99
UK	MWE-based	0.0062	0.0046	0.0014	0.0003	0.0000
	Token-based	0.1173	0.0915	0.0500	0.0302	0.0055
EGY	MWE-based	0.0000	0.0000	0.0000	0.0000	0.0000
	Token-based	0.1006	0.0533	0.0155	0.0000	0.0000
EL	MWE-based	0.0089	0.0043	0.0022	0.0000	0.0000
	Token-based	0.0803	0.0594	0.0364	0.0302	0.0097
FA	MWE-based	0.0511	0.0446	0.0387	0.0288	0.0053
	Token-based	0.2062	0.1397	0.0771	0.0427	0.0041
FR	MWE-based	0.0108	0.0077	0.0033	0.0032	0.0013
	Token-based	0.1970	0.1231	0.0519	0.0257	0.0011
HE	MWE-based	0.0036	0.0022	0.0012	0.0004	0.0000
	Token-based	0.0916	0.0594	0.0330	0.0200	0.0051
JA	MWE-based	0.0714	0.0598	0.0387	0.0277	0.0081
	Token-based	0.2682	0.1844	0.0943	0.0534	0.0077
LV	MWE-based	0.0034	0.0021	0.0013	0.0011	0.0000
	Token-based	0.0394	0.0280	0.0199	0.0164	0.0056
NL	MWE-based	0.0730	0.0773	0.0000	0.0000	0.0000
	Token-based	0.2588	0.1939	0.0319	0.0089	0.0000
PL	MWE-based	0.0025	0.0017	0.0007	0.0003	0.0004
	Token-based	0.0879	0.0640	0.0415	0.0355	0.0288
PT	MWE-based	0.0033	0.0000	0.0000	0.0000	0.0000
	Token-based	0.0945	0.0693	0.0432	0.0107	0.0000
RO	MWE-based	0.0029	0.0020	0.0014	0.0009	0.0000
	Token-based	0.0885	0.0460	0.0134	0.0047	0.0014
SV	MWE-based	0.0015	0.0006	0.0000	0.0000	0.0000
	Token-based	0.0740	0.0481	0.0200	0.0091	0.0014
SR	MWE-based	0.0297	0.0160	0.0075	0.0034	0.0000
	Token-based	0.1781	0.1209	0.0633	0.0328	0.0046
KA	MWE-based	0.0443	0.0480	0.0265	0.0138	0.0027
	Token-based	0.1417	0.0960	0.0371	0.0161	0.0023

Table 1: F1 scores of the UD-based PMI approach across different thresholds for MWE-based and token-based evaluation.

From Tables 1 and 2 we suggest 3 observations:

- Scores for full MWE matches are extremely low for both approaches
- MWE-based results are slightly, but insignificantly better in UD approach
- Token-based results are better with lower threshold percentiles in classical PMI approach.

From observations above we can interfere the following:

- This suggests that while UD features may provide some benefit in helping capture MWEs, they do not dramatically improve full match accuracy. This could imply that syntactic information alone might not be sufficient to solve the full match problem and other factors might need to be considered.
- Stronger performance of classical PMI approach on lower thresholds suggests that raw statistics favour word pairs that overlap with MWEs, even if they do not recover the full expression.

As for the time consumed, on average for different languages UD approach takes 7% time more on the same machine. We expected UD approach to have drastically higher timings, and we are pleased to find that the increase in computing time is insignificant. Still, the results in UD approach did not pay off the time consumed.

5 Related Work

The identification of multi-word expressions (MWEs) has evolved from purely statistical association measures to complex neural architectures.

Statistical and Association Measures Traditional unsupervised methods rely on association measures (AMs) such as PMI or log-likelihood (Gries, 2018) to quantify lexical affinity. Evert (2005) provides a comprehensive foundational framework for these measures, noting their effectiveness for adjacent co-occurrences. However, Pecina (2008) demonstrated that no single AM is universal, suggesting that ranking performance varies significantly depending on the MWE category and language. While these methods are interpretable and cross-linguistic, they often fail to capture rare or syntactically flexible MWEs—a limitation we aim to address using syntactic graphs.

Syntax-aware Identification The transition from surface-based windows to syntactic adjacency was extensively explored by Seretan (2011), who argued that MWEs are primarily syntactic units and should be extracted from parsed corpora. By leveraging the Universal Dependencies (UD) framework (Nivre et al., 2016), researchers have sought to create cross-linguistically consistent identification pipelines. This is particularly relevant for shared tasks like PARSEME (Savary et al., 2023; Ramisch et al., 2018), where verbal MWEs often exhibit long-distance dependencies and interleaving components. To ensure high-quality syntactic input, modern pipelines often rely on robust parsers such as UDPipe 2.0 (Straka, 2018).

Neural and Hybrid Methods Contemporary research heavily utilizes Transformer-based models and contextual embeddings, such as Multilingual

Lang	Type	50	75	90	95	99
UK	MWE-based	0.0103	0.0057	0.0041	0.0035	0.0033
	Token-based	0.1279	0.1342	0.1376	0.1392	0.1406
EGY	MWE-based	0.0018	0.0000	0.0000	0.0000	0.0000
	Token-based	0.0986	0.0915	0.0911	0.0915	0.0919
EL	MWE-based	0.0049	0.0045	0.0027	0.0017	0.0011
	Token-based	0.0852	0.0903	0.0917	0.0915	0.0922
FA	MWE-based	0.0519	0.0478	0.0471	0.0470	0.0467
	Token-based	0.2571	0.2465	0.2433	0.2427	0.2419
FR	MWE-based	0.0188	0.0160	0.0151	0.0148	0.0146
	Token-based	0.2213	0.2207	0.2225	0.2229	0.2232
HE	MWE-based	0.0047	0.0029	0.0025	0.0024	0.0024
	Token-based	0.1196	0.1206	0.1217	0.1220	0.1223
JA	MWE-based	0.0287	0.0287	0.0306	0.0314	0.0313
	Token-based	0.2008	0.2017	0.2013	0.2007	0.1976
LV	MWE-based	0.0056	0.0022	0.0012	0.0012	0.0011
	Token-based	0.0627	0.0643	0.0696	0.0717	0.0731
NL	MWE-based	0.0457	0.0539	0.0435	0.0437	0.0506
	Token-based	0.2460	0.2624	0.2525	0.2566	0.2610
PL	MWE-based	0.0045	0.0033	0.0031	0.0031	0.0031
	Token-based	0.1200	0.1247	0.1277	0.1291	0.1305
PT	MWE-based	0.0154	0.0092	0.0083	0.0083	0.0083
	Token-based	0.0909	0.0876	0.0885	0.0890	0.0892
RO	MWE-based	0.0051	0.0031	0.0029	0.0028	0.0028
	Token-based	0.1490	0.1485	0.1490	0.1490	0.1490
SV	MWE-based	0.0269	0.0166	0.0079	0.0054	0.0036
	Token-based	0.1830	0.1806	0.1747	0.1724	0.1710
SR	MWE-based	0.0371	0.0401	0.0383	0.0374	0.0372
	Token-based	0.1534	0.1613	0.1656	0.1681	0.1711
KA	MWE-based	0.0011	0.0008	0.0007	0.0007	0.0007
	Token-based	0.0151	0.0128	0.0127	0.0128	0.0129

Table 2: F1 scores of the classical PMI approach (without UD) across different thresholds for MWE-based and token-based evaluation.

BERT (Avram et al., 2023), to capture semantic and contextual nuances. While these neural methods achieve high performance, Taslimipoor and Rohanian (2019) noted that they come with significant computational costs and a lack of interpretability compared to traditional statistical approaches.

Rule and Lexicon-Based Approaches Other paradigms include rule-based approaches (Cordeiro et al., 2016), which provide high precision for well-defined patterns but suffer from low recall and require intensive manual effort. Similarly, lexicon-based approaches (Mititelu et al., 2024) ensure reliable detection of known expressions by checking against idiom dictionaries, though their coverage is naturally limited by the dictionaries in use. Our work seeks a middle ground, maintaining the lightweight nature of PMI while introducing the linguistic awareness provided by UD trees.

6 Limitations

The approach relies heavily on the accuracy of dependency parsing. Parsing errors directly affect PMI estimates.

7 Conclusion

We presented a simple syntax-aware PMI method for MWE identification that leverages Universal Dependency trees. By redefining co-occurrence in terms of syntactic relations, our approach is meant to capture non-adjacent and syntactically flexible MWEs that surface-based methods miss. Despite the idea, in practice no significant difference was found.

In the future this work could be complemented with mixture of other approaches, for example rating PMI score higher if MWE follows a rule from rule-based approach or appears in a dictionary from lexicon-based approach

Acknowledgments

We would like to express our sincere gratitude to our supervisor, Prof. Çağrı Çöltekin, for their invaluable guidance and support throughout this research. We also extend our gratitude to the organizers of the PARSEME shared task for making the dataset available and for their efforts in coordinating the shared task.

References

Andrei Avram and 1 others. 2023. Multilingual BERT for multiword expression identification across 14 languages. *arXiv preprint arXiv:2306.10419*.

Ana Cordeiro and 1 others. 2016. Rule-based approaches for multiword expression identification. In

Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval), pages 1140–1145.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Stuttgart: University of Stuttgart.

Stefan Th. Gries. 2018. Multiword expressions: A corpus-driven approach. In *Proceedings of the Workshop on Multiword Expressions (MWE)*, pages 1–15.

Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 24, 2025.

Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stanković, and Christian Chiarcos. 2024. Multiword expressions between the corpus and the lexicon: Universality, idiosyncrasy, and the lexicon-corpus interface. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 147–153, Torino, Italia. ELRA and ICCL.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and 1 others. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*.

Pavel Pecina. 2008. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop on Multiword Expressions*, pages 54–57.

Carlos Ramisch and 1 others. 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 Inflection (LaTeCH-CLFL-MWE)*, pages 222–240.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Violeta Seretan. 2011. *Syntax-based collocation extraction*. Springer Science & Business Media.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 shared task: Tagging in context, more syntactic features, and shared structures. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Shiva Taslimipoor and Omid Rohanian. 2019. Investigating BERT for multilingual NLP tasks: The case of multiword expressions. In *Proceedings of the 15th Workshop on Multiword Expressions (MWE 2019)*, pages 157–163.