Attention on Multiword Expressions: A Multilingual Study of BERT-based Models with Regard to Idiomaticity and Microsyntax

Iuliia Zaitova, Vitalii Hirak, Badr M. Abdullah,
Dietrich Klakow, Bernd Möbius, Tania Avgustinova
Saarland University, Germany
izaitova@lsv.uni-saarland.de

Abstract

This study analyzes the attention patterns of fine-tuned encoder-only models based on the BERT architecture (BERT-based models) towards two distinct types of Multiword Expressions (MWEs): idioms and microsyntactic units (MSUs). Idioms present challenges in semantic non-compositionality, whereas MSUs demonstrate unconventional syntactic behavior that does not conform to standard grammatical categorizations. We aim to understand whether fine-tuning BERT-based models on specific tasks influences their attention to MWEs, and how this attention differs between semantic and syntactic tasks. We examine attention scores to MWEs in both pre-trained and fine-tuned BERT-based models. We utilize monolingual models and datasets in six Indo-European languages — English, German, Dutch, Polish, Russian, and Ukrainian. Our results show that fine-tuning significantly influences how models allocate attention to MWEs. Specifically, models fine-tuned on semantic tasks tend to distribute attention to idiomatic expressions more evenly across layers. Models fine-tuned on syntactic tasks show an increase in attention to MSUs in the lower layers, corresponding with syntactic processing requirements.

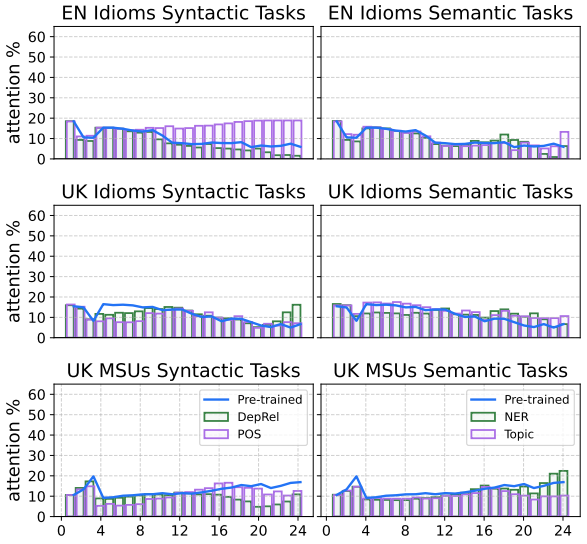

Figure 1: Layer-wise attention distribution in BERT-based models for idioms and microsyntactic units (MSUs). Results shown for English (EN) and Ukrainian (UK). Models fine-tuned on syntactic tasks (Dependency Relation Classification – DepRel, Part-of-Speech Tagging – POS) are on the left, and semantic tasks (Named Entity Recognition – NER, Topic Classification – Topic) are on the right. The y-axis shows the percentage of attention scores directed *from other sentence tokens towards Multiword Expressions*, with higher percentages indicating stronger attention focus.

1 Introduction

Attention mechanisms in Natural Language Processing (NLP) enhance the ability of models to focus on relevant aspects of input data by adjusting the model's focus dynamically. Such capabilities are foundational in models like BERT (Devlin et al., 2019), which utilize attention mechanisms to manage deep contextual understanding and complex linguistic phenomena.

Multiword Expressions (MWE) consist of two or more (lexical) components and function as a single unit. This is characteristic of idioms, collocations, and formulaic expressions, all of which exhibit a degree of semantic cohesion that distinguishes them from literal combinations of the contained words (Masini, 2019; Shwartz and Dagan, 2019). Warren (2005) notes that such expressions are pervasive in language use, often outnumbering single words. In this study, we are specifically interested in two distinct MWE types, each presenting its unique challenges: idioms and microsyntactic units.

Idioms, such as *spill the beans* or *das Handtuch werfen* ("throw in the towel") in German, embody the challenge of semantic non-compositionality. These expressions, characterized by meanings that cannot be deduced simply from their literal components, require a deeper understanding of language beyond its surface structure. The complexity of

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4083–4092

idioms poses significant challenges for NLP systems, as they must navigate these non-literal, often figurative expressions to achieve accurate understanding and generation of language (Baldwin and Kim, 2010; Fadaee et al., 2018; Zeng and Bhat, 2021; Tan and Jiang, 2021).

Microsyntactic units (MSU), such as *all the same* or тем не менее (translit.: "tem ne menee", "nevertheless") in Russian, fall out of standard grammatical categorization and analysis, due to syntactically unpredictable behavior (Iomdin, 2015). For instance, the MSU *all the same* can function in multiple ways: it may express persistence despite a fact, as in 'She was kind, but all the same she terrified me', or indifference, as in *It is all the same to me whether you stay or go* (Iomdin, 2016). Avgustinova and Iomdin (2019) highlight the role of MSUs in challenging the boundaries between lexicon and grammar, noting the need for advanced parsing strategies that can adapt to their unique syntactic structure.

In this paper, we analyze the attention patterns of both pre-trained and fine-tuned encoder-only models based on the BERT architecture (BERT-based models) towards two distinct MWE types: idioms and MSUs. We adopt the approach of Jang et al. (2024), who examine whether the attention scores in BERT change during the fine-tuning process for downstream tasks. Extending their methodology to MWEs allows us to contrast how these models handle the semantic non-compositionality in idioms against the syntactic non-conformity in MSUs. To validate our results, we employ monolingual models and datasets in six Indo-European languages coming from two language groups: Slavic and Germanic. The languages are English (**EN**), German (**DE**), Dutch (**NL**), Polish (**PL**), Russian (**RU**), and Ukrainian (**UK**). The choice of these languages is motivated by extensive linguistic resources and pre-trained models available for them, which is essential for our analyses. While we acknowledge that the language sample is not typologically diverse and does not include languages outside of Indo-European language family, leveraging these specific languages allows for a more controlled analysis within the scope of our research.

Our results indicate that fine-tuning enhances the models' ability to appropriately focus attention on MWEs, with distinct patterns observed between idioms and MSUs. Models fine-tuned on semantic tasks show a more even distribution of attention to idiomatic expressions, reflecting the need for integrating information across layers. In contrast, models fine-tuned on syntactic tasks exhibit increased attention to MSUs in lower to middle layers, according to syntactic processing requirements. These findings suggest that attention mechanisms in BERT-based models adapt during fine-tuning to better handle different types of MWEs, aligning with the linguistic characteristics of idioms and MSUs.

To facilitate further research, we have made our datasets and fine-tuned models publicly available at `github.com/IuliiaZaitova/mwe-attention`.

2 Related Work

Interpreting BERT's Attention. While the attention mechanisms in BERT-based models have been extensively studied, some works stand out for their relevance to our current study.

Tenney et al. (2019) investigate how linguistic information is represented across different layers within BERT, discovering that the model architecture implicitly mirrors the classical NLP pipeline, contradicting the often criticized "black box" nature of such models. Their findings indicate that lower layers of BERT are better at encoding local syntactic information, while higher layers progressively engage with more complex semantic processing. In addition, they observe that syntactic information is more localizable, with weights related to syntactic tasks tending to be focused on a few layers, while information related to semantic tasks is generally spread across the entire network. These findings suggests that despite BERT's holistic training approach, it may maintain an interpretable and hierarchical structure.

Jang et al. (2024) specifically examined how BERT's attention scores vary with lexical categories during the fine-tuning process on downstream tasks. Their study explores the model's behavior during the fine-tuning phase on the GLUE benchmark tasks (Wang et al., 2018), hypothesizing that BERT's attention mechanism is selectively sensitive to the lexical category of tokens — with increased attention to content words for semantic tasks and to function words for syntactic tasks. Their findings confirm that BERT's attention is not uniformly distributed but is instead strategically adjusted to emphasize relevant lexical categories based on the task, demonstrating a certain level of linguistic adaptability.

MWE Processing in Language Models. Prior

studies have explored both the extent to which Language Models understand MWEs and their different types (Kurfalı, 2020; Walsh et al., 2022; Miletić and Walde, 2024; Dankers et al., 2022; Rambelli et al., 2023; Tian et al., 2023). Further research has investigated how these models can be fine-tuned to improve their performance on the classification of MWEs (Boisson et al., 2022; Avram et al., 2023; Bui and Savary, 2024).

A work that has a particular relevance is Zaitova et al. (2023). This study proposes an approach to detect MSUs by using cosine similarity retrieved from five Word Embedding Models (WEMs), and evaluates how well these models capture syntactic idiosyncrasies. The results demonstrate the effectiveness of WEMs in capturing MSUs across six Slavic languages. Additionally, it shows that WEMs adapted for syntax-based tasks consistently outperform other WEMs at the task.

In spite of these contributions, there is still a gap in understanding how BERT-based models attend to different types of MWEs, particularly in a multilingual context. Previous studies often focus on a single language or do not consider different types of MWEs. Our study addresses these limitations by analyzing how fine-tuning affects attention to MWEs in BERT-based models across six languages.

3 Methodology

3.1 Datasets

We conducted our experiments with BERT-based models using idiom and MSU datasets. While the idiom dataset includes two groups of Indo-European languages – Germanic (DE, EN, NL) and Slavic (PL, RU, UK), the MSU dataset only includes the Slavic languages (PL, RU, UK).

An example of the data in all languages is given in Table 1. To ensure a fair comparison between idioms and MSUs, we selected or created a subset of 227 idioms in context for each language where possible, matching the number of MSUs in the MSU dataset to allow for balanced analyses.

3.1.1 Microsyntactic Unit Dataset

The Slavic MSU dataset (Zaitova et al., 2023)[1] was compiled using the list of MSUs provided in the Russian National Corpus(rus, 2003–2023). The selection process focused on the most frequently occurring MSUs, resulting in a total of 227 instances for Russian.

These 227 Russian MSUs are accompanied by their translational equivalents and parallel bilingual context sentences across five Slavic languages. The translational equivalents were manually sourced from the parallel sub-corpora of the Russian National Corpus and the Czech National Corpus (Machálek, 2020), generating parallel sets for the analysis. In our study, we only use the MSU sets for Polish, Russian, and Ukrainian.

3.1.2 Idiom Dataset

English, German, Dutch, and Polish Idiom Dataset. For the languages EN, DE, NL, and PL, we used the ID10M dataset (Tedeschi et al., 2022), which provides idiom annotations in 10 languages. The dataset was developed as part of a complete framework for idiom identification in several languages. It includes automatically created training and development data with idioms, their context sentences, and their annotations. Additionally, the test sets for four languages were curated manually. Among the languages in our analysis, the test sets were available for EN and DE.

Russian Idiom Dataset. For Russian, we used 85 idioms from the dataset by Aharodnik et al. (2018). The remaining 142 idioms in context were manually retrieved from the "Academic Dictionary of Russian Phraseology" (Baranov and Dobrovolsky, 2015), ensuring comprehensive coverage. The resulting idioms were proofread by two native speakers (a 28 year old female, and a 24 year old male), who are also professional linguists.

Ukrainian Idiom Dataset. For Ukrainian, due to the absence of pre-existing idiom datasets, we created our own dataset by generating idioms with OpenAI's ChatGPT, specifically using the GPT-4 model. Each idiom was subsequently verified by a native Ukrainian speaker (age: 24; gender: male), who is also a professional linguist to ensure its accuracy.

3.2 Models

We used six large 24-layer encoder-only transformer models based on the BERT architecture trained using only the masked language modeling objective on monolingual texts (BERT-based models). Specifically, for English we utilized BERT-large-cased, for German – GBERT-large (Chan et al., 2020), for Dutch – RobBERT-large (Delobelle and

[1] https://huggingface.co/datasets/izaitova/slavic_fixed_expressions

Lang	Type	Sentence	English Translation
EN	Idiom	They covered the whole field **from A to Z** in eight classes.	-
DE	Idiom	**Ihre große Liebe** sei Jonathon, sagt Sarah.	Sarah then says that Jonathon is her **great love**.
NL	Idiom	**Af en toe** verzorgde ze nog een gastoptreden.	**Now and then**, she would make a guest appearance.
PL	Idiom	**Ca ly czas** by lo mi zimno z nim.	I was **constantly** cold because of him.
PL	MSU	**Z trudem** kojarzy l i pojmowa l, co do niego mówia.	He **barely** understood what was said to him.
UK	Idiom	Для нього робота — це альфа і омега всього життя.	For him, work is the **be-all and end-all** of life.
UK	MSU	Я все ще сподівався на банальну аварію.	I was **still** hoping it was just a mundane accident.
RU	Idiom	За что позор за позором валится на мою голову?	Why does disgrace after disgrace **fall on my head**?
RU	MSU	Я всё время думал о тебе, день и ночь.	I thought about you **all the time**, day and night.

Table 1: Idioms and microsyntactic units in context.

Remy, 2023), for Polish – HerBERT-large-cased (Mroczkowski et al., 2021), for Russian – ruBERT-large (Zmitrovich et al., 2023), and for Ukrainian – Liberta-large (Haltiuk and Smywiński-Pohl, 2024).

3.3 Fine-tuning Process

To analyze how task-specific training affects the attention mechanisms of pre-trained models when processing idioms and MSUs, we fine-tuned each model on two syntactic and two semantic NLP tasks. The selection of tasks was guided not only by the availability of fine-tuning datasets in the studied languages, but also by their relevance to linguistic properties of MWEs.

3.3.1 Syntactic Tasks

Dependency Relation Classification (DepRel) predicts the dependency relation tag for each token in a sentence. This task assesses the model's ability to understand syntactic relationships between tokens, which is crucial for parsing the often non-standard structures of MWEs. For DepRel, we used datasets based on Universal Dependencies (Nivre et al., 2020).

Part-of-Speech (POS) Tagging assigns grammatical categories to each token, testing the model's grasp of syntactic roles and morphological forms. POS tagging is essential for handling tokens that may have unconventional syntactic functions in idiomatic versus literal contexts. Datasets for this task are also sourced from the Universal Dependencies.

3.3.2 Semantic Tasks

Named Entity Recognition (NER) identifies and classifies entities in texts into categories like people, organizations, and locations, evaluating the model's ability to extract semantic information. The task of NER is relevant to our study goal because MWEs are often culture-specific and can include named entities. We utilized the WikiANN dataset (Pan et al., 2017) for this task.

Topic Classification (Topic) assigns sentences to predefined topics based on content, requiring understanding of broader semantic context. Topic classification is relevant since MWEs often require broader context to be interpreted correctly. The SIB-200 dataset (Adelani et al., 2023) was used for this task.

3.3.3 Training and Evaluation

The datasets were divided into training, development, and test sets, with sizes varying depending on the language and dataset. Table 2 shows the number of training samples used for fine-tuning the models on each downstream task across the six languages studied.

All models were fine-tuned for 10 epochs using the Hugging Face Transformers library (Wolf et al., 2020), with the development set for validation and hyperparameter tuning. We evaluated the models on the test set using the F1 score to ensure consistent comparison across tasks. The models achieved competitive performance, with F1 scores exceeding 0.75 across all tasks and languages.

Lang	DepRel	POS	NER	Topic
EN	5000	7000	5000	701
DE	5000	7000	5000	701
NL	5000	7000	5000	701
PL	5000	7000	5000	701
RU	5000	5400	7000	701
UK	5496	5000	7000	701

Table 2: Number of training samples for fine-tuning by task.

4 Experimental Setup

4.1 Data Preprocessing and Setup

Each context sentence with an MWE is tokenized using the pre-trained model tokenizer (e.g., BERT-large-cased tokenizer for English). This ensures consistency and accuracy in mapping MWE tokens to attention vectors. After tokenization, the inputs are fed into both the pre-trained and the fine-tuned models, and the attention outputs are extracted for analysis.

To maintain alignment between the tokens of the MWEs and the attention weights, we carefully handle cases where MWEs are split into subword tokens. We aggregate the attention weights corresponding to all subword tokens that compose an MWE, treating them as a single unit in our analysis.

4.2 Attention Extraction

For each model, we extract the multi-head attention matrices from all layers during the forward pass. These matrices represent the attention weights that each token in the input sequence assigns to every other token, including itself. Specifically, for a model with L layers and H attention heads per layer, we obtain $L \times H$ attention matrices per input sequence.

To simplify our analysis, we average the attention matrices across the H heads in each layer. This results in a single averaged attention matrix per layer of size $T \times T$, where T is the length of the tokenized input sequence. By examining these averaged matrices, we can analyze how attention is distributed across the model's layers.

4.3 Attention Analysis

To quantify the attention patterns related to MWEs, we adopt and extend the metrics proposed by Jang et al. (2024). Specifically, we additionally study the layer-wise changes in attention distribution across MWE categories both within MWE and from context to MWE.

Our analysis focuses on the following aspects:

- **Attention from Context to MWEs**: For each layer in the model, we compute the average attention scores directed from all other tokens in the sentence towards tokens of MWEs.
- **Attention within MWEs**: We also analyze the attention among tokens within MWE, calculating the average attention that MWE tokens pay to one another at each layer.
- **Impact of Fine-tuning**: We compare the pre-trained models to fine-tuned models and assess how fine-tuning on syntactic or semantic tasks changes the model's attention to MWEs. Moreover, we analyze the direction of changes in attention scores (positive or negative) for all models and both categories of MWEs.
- **Attention Differences between types of MWEs**: We look at the attention distributions for idioms and MSUs to compare the patterns for two types of MWEs.
- **Language-Specific Attention Patterns**: We examine how different language-specific BERT-based models process MWEs to identify similarities and differences in attention patterns across languages and model architectures.

5 Results and Discussion

In this section, we present our analysis of the attention patterns by the pre-trained and fine-tuned BERT-based models when processing two types of MWEs: idioms and MSUs across different layers.

5.1 Pre-trained vs. Fine-tuned Models

Figure 1 shows the percentage of attention scores directed towards MWEs by layer in PL, UK, and EN. The figure compares the pre-trained models with models fine-tuned on syntactic and semantic tasks. As can be seen from the figure, in the pre-trained models, the attention directed towards both idioms and MSUs is more uniform across middle and upper layers, indicating a tendency towards general-purpose representation. The models fine-tuned on both syntactic and semantic tasks demonstrate noticeably higher attention peaks. This implies that fine-tuning improves the model's ability to concentrate on MWEs in general.

Task	EN			DE			NL			PL			RU			UK		
	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
Pre-trained	1	4	5	3	2	1	**12**	3	7	3	2	5	1	2	**10**	**4**	6	5
DepRel	1	5	4	3	2	1	5	3	6	3	2	5	1	3	2	1	**11**	**12**
POS	1	**11**	4	3	2	5	**12**	7	3	3	2	1	1	2	**9**	1	2	**12**
NER	1	5	6	3	6	2	**12**	7	6	3	2	5	1	2	3	1	2	**12**
Topic	1	4	5	3	1	2	3	7	5	3	2	5	1	**9**	2	7	5	4

(a) Idioms

Task	EN			DE			NL			PL			RU			UK		
	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
Pre-trained	–	–	–	–	–	–	–	–	–	3	2	1	2	4	3	3	2	**12**
DepRel	–	–	–	–	–	–	–	–	–	3	2	1	2	3	5	3	2	**12**
POS	–	–	–	–	–	–	–	–	–	3	2	1	**11**	2	**12**	3	2	**12**
NER	–	–	–	–	–	–	–	–	–	3	2	1	2	3	6	3	2	1
Topic	–	–	–	–	–	–	–	–	–	3	2	1	2	4	3	3	2	1

(b) Microsyntactic Units

Table 3: Top three layers with highest attention percentage allocated to idioms and microsyntactic units (MSUs) across six languages. The layers with the highest attention percentages are labeled as T1 (highest), T2 (second highest), and T3 (third highest). Lower layers (1-8) are marked in blue color and middle layers (9-16) in yellow color and bold font to illustrate where the model focuses its attention.

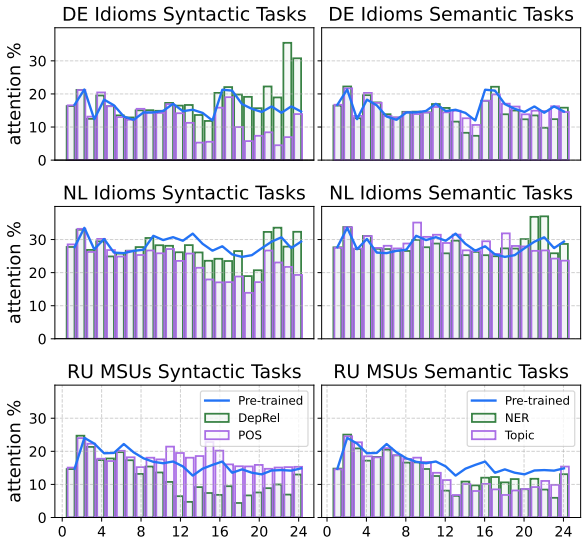

Figure 2: Layer-wise attention distribution within Multiword Expressions (MWEs) in German (DE), Dutch (NL), and Russian (RU) BERT-based models. The graphs show attention distribution within tokens of idioms and microsyntactic units (MSUs), comparing pre-trained models with those fine-tuned on syntactic tasks (Dependency Relation Classification – DepRel, Part-of-Speech Tagging – POS) and semantic tasks (Named Entity Recognition – NER, Topic Classification – Topic). The y-axis represents the percentage of attention between tokens within the same MWE.

In addition to attention from other tokens in the sentence to MWEs, we analyzed the attention scores within the tokens of MWEs. Figure 2 illustrates the percentage of attention scores within idioms and MSUs by layers in the DE, NL, and RU models. Here, we can again observe stronger attention peaks in the fine-tuned models as compared to the pre-trained models.

According to attention scores both within MWE and from context to MWE, models fine-tuned on semantic tasks – NER and Topic – show an increase in attention to MWEs in the higher layers compared to the pre-trained model. This is consistent across all six languages studied except for RU.

Figure 3 shows the differences in the percentage of average attention values from the pre-trained model across layers for two types of MWEs in RU. From the figure, we can see that fine-tuning leads to mostly decreased attention to both idioms and MSUs across most layers, and particularly in higher layers. Fine-tuning on syntactic tasks (Topic and DepRel) leads to a sharp decrease in middle to upper layers, except for POS task when processing MSUs. Fine-tuning on semantic tasks (NER and Topic) results in mixed changes. For both NER and Topic, attention to idioms and MSUs decreases in most layers but slightly increases in some lower to

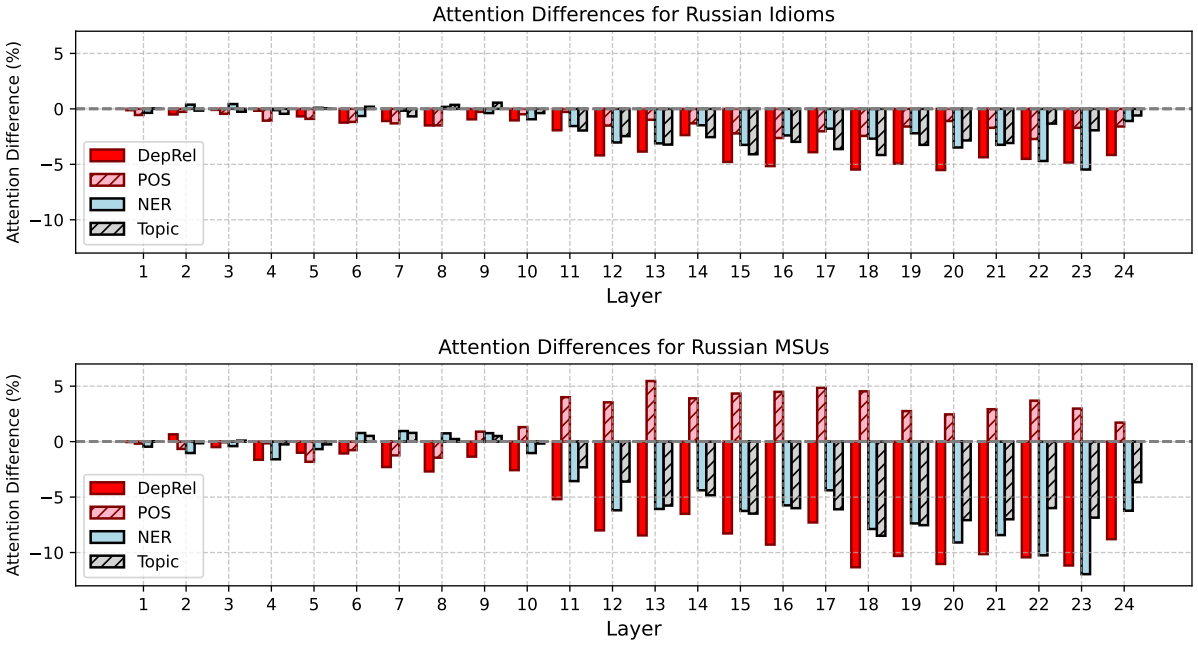

Figure 3: Differences in Attention Patterns for Idioms and Microsyntactic Units (MSUs) in Russian. The bars show layer-wise changes in attention percentage after fine-tuning on syntactic tasks (Dependency Relation Classification – DepRel, Part-of-Speech Tagging – POS) and semantic tasks (Named Entity Recognition – NER, Topic Classification – Topic). Positive values indicate increased attention and negative values show decreased attention compared to the pre-trained model.

middle layers.

These mixed changes suggest that fine-tuning on NER and Topic classification may cause the model to redistribute attention in a more nuanced way, possibly balancing between capturing named entity-specific information and broader contextual cues.

The consistent decrease in attention across most layers and tasks may indicate that fine-tuning reduces the model's sensitivity to MWEs. The changes in attention are especially pronounced in higher layers. Since higher layers are often associated with capturing abstract and semantic information, as mentioned in Section 2, their decreased attention suggests a possible trade-off between task performance and the model's ability to handle MWEs.

5.2 Idioms vs. Microsyntactic Units

Regardless of whether the model is only pre-trained or has been fine-tuned, a consistent pattern where attention peaks at the lower layers (3-4) when processing MSUs is observed across all models for all languages.

The presence of attention peaks in the lower layers for MWEs, and especially for MSUs in both figures, suggests that these units are more closely associated with syntactic processing. As stated before, lower layers of neural networks are typically more focused on syntactic features of a language since they process more structural aspects of the input data before passing higher-level semantic information to upper layers.

For MSUs, attention percentage is in general more varied across layers in models fine-tuned on syntactic tasks, while semantic tasks lead to a flatter, more uniform distribution. In Figure 3, we can see that attention to MSUs increases in middle and upper layers when fine-tuned on the POS task. The other syntactic task, DepRel, does not show the same pattern. In contrast, the attention scores for MSUs drop even more compared to idioms, except for a slight increase in layer 2. This could be because DepRel relies more on understanding grammatical relations rather than the specific syntactic categories like POS.

5.3 Comparison between Language Groups

With the Germanic languages (EN, DE, NL), the models produce more uniform attention patterns towards idioms within MWE and from context to MWE. This may be attributed to lower morphological complexity in Germanic languages, allowing the models to adapt more uniformly during fine-tuning.

In contrast, the models for Slavic languages (PL, RU, UK), especially for PL, display more varied attention patterns, as can be seen in Figure 1. PL also stands out from all other languages by the presence of a large attention peak in the lower layers, exactly at the layers where the attention drops for other languages. Since this is only observed when processing PL idioms and not PL MSUs, such anomaly is more likely to be related to the dataset rather than the language or the model.

Table 3 provides an overview of top three layers with highest attention percentage towards idioms and MSUs in six languages. The table shows that, in general, for EN, NL, RU, and UK, the layers with highest attention to MWEs are spread across both lower and middle layers. For DE and PL, all MWEs receive high attention in the lower layers. We can see, however, that MSUs are mostly attended to in lower layers as opposed to idioms, which is expected given that idioms typically attract more attention in higher layers.

6 Conclusion

In this study, we analyzed the attention patterns of fine-tuned BERT-based models in relation to idioms and MSUs across six Indo-European languages from two language groups. By extending the methodology of Jang et al. (2024) to MWEs, we demonstrated that fine-tuning on syntactic and semantic tasks significantly affects how models allocate attention to different types of MWEs.

Our results indicate that:

- In general, models fine-tuned on syntactic tasks exhibit increased attention to MSUs in lower to middle layers, in accordance with syntactic processing requirements.
- Models fine-tuned on semantic tasks show a tendency to distribute attention more evenly across layers, which could reflect a need for integrating information across different layers.
- Cross-linguistic differences exist both between Germanic and Slavic languages, as well as across languages of the same language group. This underscores the complexity of how transformer models manage attention distribution.
- While there is a general trend towards decreased attention in syntactic tasks and more evenly distributed attention in semantic tasks, the anomalies highlight the non-uniform behavior of attention mechanisms in BERT-based models.

These findings suggest that attention mechanisms in transformer models adapt during fine-tuning to better handle complex linguistic phenomena, according to the linguistic properties of the target language and task.

Future work could explore larger and more diverse datasets, different model architectures, and additional languages to build upon our findings and expand the understanding of how neural models process MWEs.

Limitations

While our study provides valuable insights into how BERT-based models process MWEs across different languages and tasks, several limitations should be acknowledged.

Our research utilized idiom datasets obtained from various sources, which differ in composition and definition. The concept of an idiom itself lacks a precise, universally accepted definition, which might lead to inconsistencies in interpreting the results. This variability could affect the generalizability of our findings, as the models' performance might be influenced by the idiosyncrasies of the datasets used. Our dataset of microsyntactic units includes only the Slavic language groups, which could also bias the analysis.

Moreover, while our study focuses on semantic tasks like Named Entity Recognition and Topic Classification, which require semantic understanding, these tasks may involve only shallow semantic processing. This limitation could affect the extent to which fine-tuning on these tasks enhances the model's attention to idioms. The performance of BERT-based models on other semantic NLP tasks, such as summarization and paraphrasing, could provide additional insights into their ability to handle MWEs.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

References

2003–2023. Russian National Corpus. Accessed 25.09.2024.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a russian idiom-annotated corpus. In *International Conference on Language Resources and Evaluation*.

Tania Avgustinova and Leonid Iomdin. 2019. *Towards a Typology of Microsyntactic Constructions*, volume 11755 of Lecture Notes in Computer Science. Springer, Cham., pages 15–30.

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, Second Edition*, pages 267–292. Chapman and Hall/CRC.

A Baranov and D Dobrovolsky, editors. 2015. *Academic Dictionary of Russian Phraseology*, 2 edition. LEKSRUS, Moscow.

Joanne Boisson, Jose Camacho-Collados, and Luis Espinosa-Anke. 2022. CardiffNLP-metaphor at SemEval-2022 task 2: Targeted fine-tuning of transformer-based language models for idiomaticity detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 169–177, Seattle, United States. Association for Computational Linguistics.

Van-Tuan Bui and Agata Savary. 2024. Cross-type French multiword expression identification with pre-trained masked language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4198–4204, Torino, Italia. ELRA and ICCL.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

P Delobelle and F Remy. 2023. Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion.

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.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mykola Haltiuk and Aleksander Smywiński-Pohl. 2024. LiBERTa: Advancing Ukrainian language modeling through pre-training from scratch. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 120–128, Torino, Italia. ELRA and ICCL.

Leonid Iomdin. 2015. Microsyntactic constructions formed by the Russian word raz. *SLAVIA c ˇasopis pro slovanskou filologii*, 84(3).

Leonid Iomdin. 2016. Microsyntactic phenomena as a computational linguistics issue. In *Proceedings of the Workshop on Grammar and Lexicon: interactions and interfaces (GramLex)*, pages 8–17, Osaka, Japan. The COLING 2016 Organizing Committee.

Dongjun Jang, Sungjoo Byun, and Hyopil Shin. 2024. A study on how attention scores in the BERT model are aware of lexical categories in syntactic and semantic tasks on the GLUE benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1684–1689, Torino, Italia. ELRA and ICCL.

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.

Tomáš Machálek. 2020. Kontext: Advanced and flexible corpus query interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France. European Language Resources Association.

Francesca Masini. 2019. Multi-word expressions and morphology. Oxford University Press, Oxford.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kyiv, Ukraine. Association for Computational Linguistics.

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

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Giulia Rambelli, Emmanuele Chersoni, Marco S. G. Senaldi, Philippe Blache, and Alessandro Lenci. 2023. Are frequent phrases directly retrieved like idioms? an investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 87–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *Preprint*, arXiv:1905.05950.

Ye Tian, Isobel James, and Hye Son. 2023. How are idioms processed inside transformer language models? In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 174–179, Toronto, Canada. 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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Beatrice Warren. 2005. A model of idiomaticity. *Nordic Journal of English Studies*, 4:35–54.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Iuliia Zaitova, Irina Stenger, and Tania Avgustinova. 2023. Microsyntactic unit detection using word embedding models: Experiments on Slavic languages. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1265–1273, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic Expression Identification using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian. *Preprint*, arXiv:2309.10931.