

Metaphor Detection for Czech

Barbora Štěpánková
Research Project Report

Abstract

Metaphor detection is a challenging natural language processing task, especially for low-resource languages such as Czech, where large metaphor-annotated datasets are not available. This work investigates whether multilingual transformer-based models and cross-lingual transfer from languages with larger annotated datasets can be used for Czech metaphor detection. We evaluate several training settings using English, Spanish, Slovenian, and Czech metaphor annotations and analyze the effect of multilingual training data. The results show that multilingual models can transfer metaphor knowledge to Czech, but even a small amount of Czech annotated data substantially improves performance. We further explore how metaphor properties such as conventionality and translatability relate to model performance and present preliminary findings suggesting that these properties influence cross-lingual metaphor detection.

1 Introduction

Metaphor detection is an important task in natural language processing, as metaphorical language is common in everyday communication while remaining challenging for automatic systems. Metaphors occur across domains, languages, and levels of conventionality, making their identification difficult even for modern neural models. Reliable metaphor detection can support downstream applications such as sentiment analysis, machine translation, question answering, and literary analysis.

Most research on metaphor detection has focused on English, with several large-scale annotated datasets and neural approaches developed for the language (Steen et al., 2010; Su et al., 2020; Zhang and Liu, 2022; Reimann and Schefler, 2025). Other languages have received considerably less attention. Besides English, substantial datasets exist for Spanish (Sanchez-Bayona and

Agerri, 2022, 2026) and Slovenian (Antloga, 2020), while research on other languages (mainly Russian and German) often relies on smaller datasets or transfer learning approaches (Tsvetkov et al., 2013, 2014; Berger, 2022; Hülsing and Schulte Im Walde, 2024).

Czech represents a low-resource setting for metaphor detection. Although a Czech metaphor-annotated dataset following the MIPVU annotation procedure exists (Pavlas et al., 2018), it is substantially smaller than the available English, Spanish, and Slovenian resources. Building large annotated datasets is expensive, which motivates investigating whether multilingual models and transfer learning can be used to improve metaphor detection in languages with limited supervision.

A generally accepted view of metaphor describes it as a systematic mapping between conceptual domains, where one type of experience is understood in terms of another (Lakoff and Johnson, 1980). This suggests that some metaphorical structures may be shared across languages, providing a motivation for transferring metaphor detection knowledge between languages. However, languages can differ in how specific metaphors are expressed, conventionalized, or lexicalized, making cross-lingual transfer a non-trivial problem.

In this work, we adopt the operational definition of metaphor provided by the MIPVU annotation procedure (Steen et al., 2010) used in all datasets considered in this work. Under this framework, a lexical unit is annotated as metaphorical when its contextual meaning contrasts with a more basic meaning but can be understood in comparison with that basic meaning. The metaphor detection task is formulated as token classification, where each lexical unit receives a binary label indicating whether it is used metaphorically or not. Since MIPVU has been adapted for several languages (Nacey et al., 2019), it provides a consistent annotation framework for multilingual experiments.

We investigate whether multilingual transformer-based models can be used for metaphor detection in Czech with only a small amount of Czech metaphor-annotated data. We evaluate different training data compositions, including cross-lingual transfer from English, Slovenian, and Spanish, the influence of additional Czech data, and the effect of training data properties such as part of speech distribution. We also analyze which types of metaphors are easier for multilingual models to detect and explore the role of metaphor conventionality and cross-lingual translatability.

This work presents the following contributions:

- We evaluate multilingual transformer models for Czech metaphor detection under different cross-lingual transfer settings.
- We analyze how the amount and composition of training data influence Czech metaphor detection.
- We introduce a preliminary analysis of metaphor conventionality and translatability in a cross-lingual setting.

We made our code and data publicly available at <https://github.com/stepankovab/Czech-Metaphor-Detection>.

2 Data

Large manually annotated metaphor corpora are available only for a handful of languages. We chose to use three existing datasets annotated following the most widespread metaphor annotation schema MIPVU (Steen et al., 2010), despite it not being able to always capture the ideal representation of language metaphoricality. We also gathered a small amount of Czech data annotated following this schema. In this section, I will describe these datasets and the preprocessing steps taken. More detailed statistics are in Table 1. The graphs of percentage of metaphoricality based on the POS are in Figure 4.

2.1 Large Existing Datasets

English: VUA (Steen et al., 2010) The VU Amsterdam Metaphor Corpus is the first corpus annotated with MIPVU procedure and reaches inter-annotator agreement of Fleiss’ kappa of 0.85. It contains English academic texts, news, fiction, and speech transcriptions. The corpus has relatively

Language	Words	Sentences	Metaphors
English	237,247	16,202	10.9%
Spanish	116,976	3631	1.8%
Slovenian	259,881	13,963	6.22%
Czech General	4659	236	10.5%
Czech Poetry	685	43	17.2%

Table 1: Number of words, sentences, and the percentage of metaphors for each of the annotated datasets.

high percentage of metaphors of 10.9%, and annotates all POS: for example, 44.8% of adpositions are annotated as metaphorical. On average it has 14.6 words per sentence. The dataset is annotated for both metaphor and POS.

Slovenian: Komet (Antloga, 2020) The Komet dataset takes Slovenian texts in the news, fiction and online texts genres and annotates them with MIPVU procedure modified for Slovenian. It was annotated by a single annotator, not providing an agreement. It also annotates all POS, but provides less metaphorical text than VUA with only 6.2% of metaphorical words. The average sentence length is 18.6 words. The dataset was also annotated for both metaphor and POS.

Spanish: CoMeta (Sanchez-Bayona and Agerri, 2022) The CoMeta dataset uses MIPVU adapted to Spanish to annotate Spanish news, online texts and political speeches. CoMeta annotates only nouns, verbs and adjectives, resulting in only 1.8% of metaphors in the corpus. The inter-annotator agreement of Cohen’s kappa of 0.63 was reached, suggesting that CoMeta might be more complex dataset than the VUA. The sentences are longer than for the two previous datasets with average length of 32.2 words. The POS was added using UDPipe (Straka et al., 2016).

2.2 Czech datasets

In this section, we will describe the Czech MIPVU-annotated portion of data. There are five domains of the Czech data: academic texts, news, fiction, and speech transcriptions and poetry. These have been annotated using the MIPVU modified for Czech (Pavlas et al., 2018). We divide these into two groups: *Czech General* containing the academic texts, news, fiction, and speech transcriptions, as these do not introduce a domain transfer compared to the three existing dataset mentioned in Section 2.1, and *Czech Poetry*, as poetry was to

the best of our knowledge not present in any of the large existing datasets.

Czech: General The *General* section of the Czech data was annotated as part of the Czech MIPVU experiment (Pavlas et al., 2018), and was obtained from the authors based on personal communication. Extended filtering and postprocessing was carried out on the data to deduplicate the annotated texts and compile annotations from several annotators. For texts with three annotators available, the final annotation was the majority agreement. For texts with two annotators available, the cases with disagreement were resolved by the author of this report. The annotator agreement reported by Pavlas et al. (2018) was 0.7. POS was added using UDPipe.

Czech: Poetry The author of the report and two other students annotated 6 poems from Jiří Dědeček and František Gellner according to the guidelines by Pavlas et al. (2018). We achieved the Fleiss’ kappa of 0.53 signifying moderate agreement. The six poems consisted of 685 words and 118 annotated metaphors, showing very high metaphoricality (17.2%). Compared to the other annotated texts we examined, in Poetry, the majority of metaphors are nouns, verbs and adjectives.

Considering the moderate agreement as well as the background of the annotators, we decided to use the *Czech: General* dataset for the following experiments unless explicitly stated that the experiments were carried out on *Poetry* data.

3 Setups

In our work, metaphor detection is formulated as a token classification task. We experimented with several transformer-based models with support for languages of all datasets we use such as multilingual BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and modern multilingual BERT (Marone et al., 2025). As there have been several studies that multilingual models are not language neutral (Pires et al., 2019; Libovický et al., 2020) and biased towards English (Papadimitriou et al., 2023), we also experimented with Slavic BERT (Arkhipov et al., 2019). Considering the class imbalance of metaphor-annotated data, we tested weighted cross-entropy loss function with various weight parameters as well as focal loss

function (Lin et al., 2020). We also conducted an extensive hyper-parameter search.

In this section, we present several setups which give interesting insight into cross-lingual metaphor detection. Across these setups we change the training and testing data and unless stated otherwise, all experiments use the same model and hyperparameters: modern multilingual BERT (mmBERT) trained for 3 epochs with a batch size of 32, learning rate of $3e-5$, a weight decay of 0.01, a warmup ratio of 0.06 and standard cross-entropy loss. The AdamW optimizer with a linear learning rate scheduler is used.

Monolingual Training and testing data come from the same dataset. In this setup, we gradually increase the number of training sentences and report the F1 score on the same test dataset. This experiment is carried out for English, Slovenian and Spanish.

One language transfer In this setup, the experiments were carried out on a fixed train and test split of the *Czech: General* data (80% train and 20% test). In addition to the train portion of the Czech data (189 sentences) which is always included, we gradually increase the number of training sentences of one other language (English, Slovenian or Spanish) which are added before the Czech training sentences.

Unannotated Czech data As there is no large metaphor-annotated dataset for Czech, we experiment whether adding a number of unannotated Czech sentences could improve the performance of Czech metaphor detection. In this setup, we increase the number of Czech sentences with all words annotated as non-metaphorical in the training data, followed by the Czech train portion of 189 annotated sentences.

Training languages ratio In this setup, we control for the amount of training data by words rather than by sentences. We train each model on roughly 94k words, where a little less than 4k words are Czech and the remaining 90k are distributed in different ratios between English, Slovenian and Spanish. To utilize all Czech annotated data as well as to have a more robust evaluation, we use 5-fold evaluation for the Czech data, using 4 folds for training and 1 for testing.

All annotated data: zero-shot vs fine-tune In this setup we use all available annotated data in

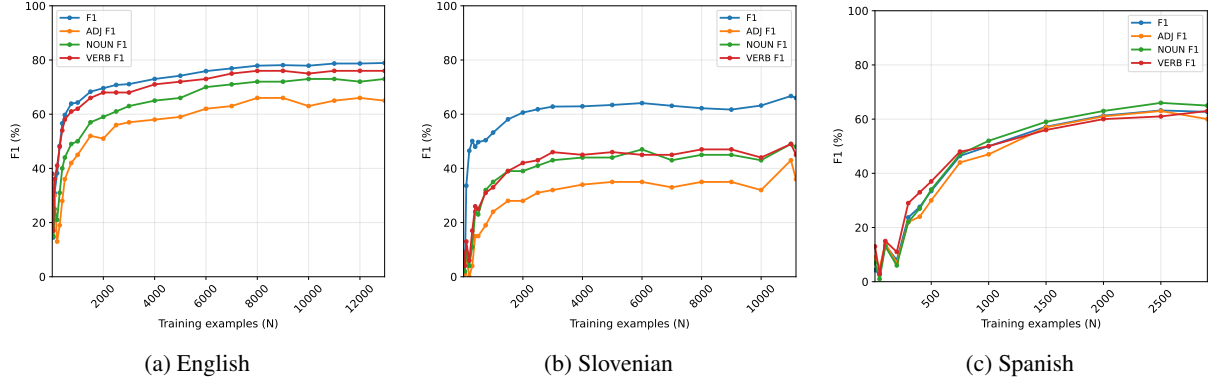


Figure 1: F1 score of monolingual metaphor detection models trained with increasing numbers of training sentences in English, Slovenian, and Spanish.

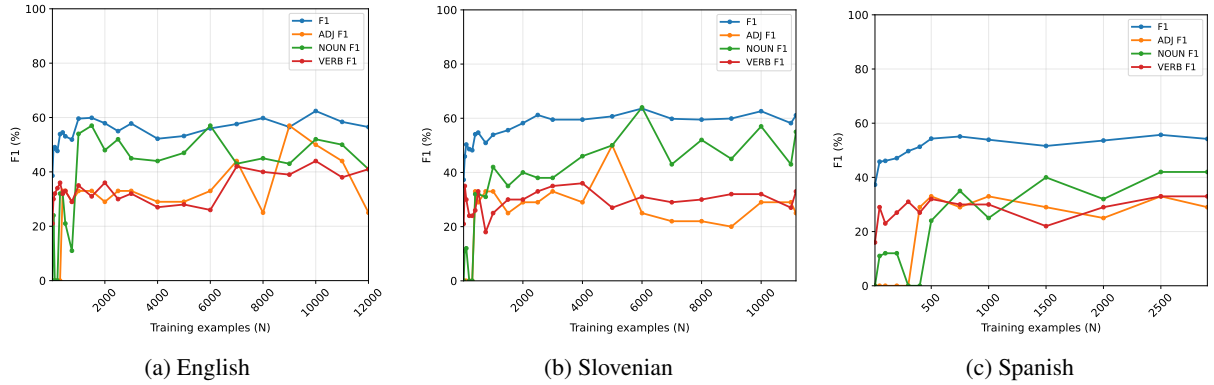


Figure 2: F1 score on Czech metaphor detection with increasing amounts of training data from English, Slovenian, or Spanish added to the Czech training data.

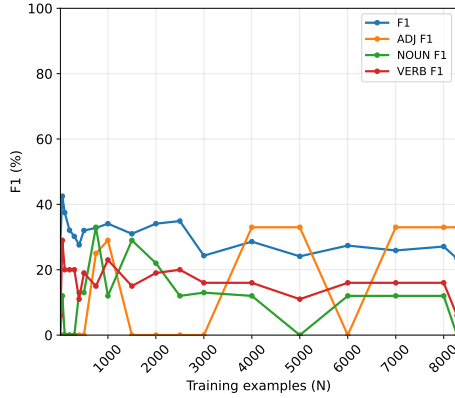


Figure 3: Effect of adding increasing amounts of unannotated Czech data before training on the annotated Czech metaphor dataset.

English, Slovenian and Spanish for training and all available annotated data in Czech for testing. In the zero-shot experiments, we do not use any Czech data for training, testing the model’s ability to transfer the metaphor knowledge on a new language. In the fine-tune experiments, we use the k-fold evaluation to add a portion of Czech data to the training

data, while still being able to evaluate the model on all Czech data. This tests the influence even a small amount of Czech data has on the results. For additional experiments about the best order of languages in the training data see Appendix A.

Part of speech Each dataset has different distribution of metaphorical words in regards to part of speech (POS). In this setup, we test how trainig on various POS influences the results. Specifically, we use all annotated data as trainig data and test on all Czech data, using the 5-fold evaluation. We compare the results when we leave all original annotations in the training and testing data and when we leave metaphor annotations only for nouns, verbs and adjectives.

Poetry Finally, we test the performance of the metaphor detection system on Czech Poetry, introducing cross-linguistic and cross-domain transfer. We use all available annotated data for training and the Czech Poetry dataset for testing.

Training Data Composition	F1	Precision	Recall	Adj F1	Noun F1	Verb F1
90k English, 4k Czech	0.61	0.71	0.55	0.32	0.47	0.44
90k Slovenian, 4k Czech	0.57	0.68	0.50	0.32	0.35	0.35
90k Spanish, 4k Czech	0.59	0.76	0.49	0.35	0.36	0.37
60k English, 15k Slovenian, 15k Spanish, 4k Czech	0.60	0.69	0.54	0.43	0.45	0.40
60k Slovenian, 15k English, 15k Spanish, 4k Czech	0.63	0.73	0.56	0.44	0.43	0.40
60k Spanish, 15k English, 15k Slovenian, 4k Czech	0.62	0.73	0.54	0.50	0.45	0.39
30k English, 30k Slovenian, 30k Spanish, 4k Czech	0.61	0.69	0.55	0.50	0.40	0.39

Table 2: Average results for different multilingual training data compositions. Each setup contains approximately 90k words from non-Czech languages and 4k Czech words.

4 Results

In this Section, we discuss the results of the various experimental setups presented in Section 3.

Monolingual In Figure 1, we can see that generally, with more training data the performance rises. When trained on less than 100 sentences, the performance is not above chance, with F1 score close to the expected random baseline. We can also see that the overall performance on all POS is better than individual performance on nouns, verbs and adjectives for English and Slovenian, while for Spanish, the F1 scores for the individual POS stay close to the overall F1 score as in the Spanish dataset there are no annotations for adpositions, which are considerably easier to detect for the model in English and Slovenian.

One language transfer When training on 189 sentences of the Czech training data only, the overall F1 score is 0.30 and for nouns, verbs and adjectives stays 0. Adding just 10 sentences from English, Slovenian or Spanish raises the overall F1 score to 0.39, 0.37 and 0.37 respectively, while also raising the verb F1 score to be non-zero. Contrary to the *Monolingual* setup, there is not a clear trend in the *One language setup* that with more data the model gets better after the initial few hundreds of sentences (see Figure 2). This might be related to the small amount of testing data, which is responsible for the overall experiment instability.

Unannotated Czech data While adding 10 additional unannotated sentences before the 189 annotated sentences did not change the F1 score, adding 50 sentences gave an F1 score of 0.43. After this peak, the F1 score dropped with additional unannotated sentences, while still maintaining quite stable non-zero F1 score for verbs and nouns (Figure 3)

Training language ratio This setup shows that training on a bigger mix of languages can be beneficial compared to training on a single language plus a small amount of Czech. In Table 2 we can see that while training on only Slovenian and Czech or only Spanish and Czech yields worse performance than training multilingually, especially in verb, noun and adjective F1 scores. On the other hand, training on only English and Czech yields the best achieved F1 score for nouns and verbs out of the experiments in this setup. Several factors may contribute to this observation, including English-centric multilingual representations, differences in annotations across datasets, and the limited size of the Czech evaluation data, even though for this setup, we use k-fold cross-validation to utilize all Czech data for testing.

All annotated data: zero-shot vs fine-tune We can see in Table 3 that there is a considerable improvement of performance on Czech data when even a relatively small amount of Czech data (less than 4k) are included in the training data (over 600k together), especially in noun and verb F1 scores. The zero-shot performance of 0.44 is still substantially above chance performance. In this setup, training on all data means that the approximate ratio between English, Slovenian and Spanish is 2:2:1. When we also compare the total F1 score of 0.64 to the scores in the *training language ratio* setup which was trained on 94k words only, the score is slightly better, but not drastically, being consistent with the findings of *one language transfer* setup.

5-fold evaluation To better understand the behavior of the model on Czech data, in Table 4 we can see each of the five folds of the Czech data evaluated individually under the *all annotated data: fine-tune* setup. We can see a large gap in performance based on which folds are used for training

Setup	F1	Prec.	Recall	Adj F1	Noun F1	Verb F1
Zero-shot	0.44	0.82	0.30	0.36	0.18	0.19
Fine-tune	0.64	0.72	0.57	0.50	0.46	0.45

Table 3: Comparison of zero-shot cross-lingual transfer and adding Czech annotated data during fine-tuning.

Fold	F1	Precision	Recall	Adj F1	Noun F1	Verb F1
4	0.7	0.79	0.63	0.57	0.48	0.48
1	0.7	0.72	0.67	0.80	0.40	0.50
3	0.61	0.67	0.55	0.25	0.48	0.49
2	0.6	0.68	0.54	0.40	0.40	0.50
0	0.59	0.75	0.48	0.48	0.54	0.26
All	0.64	0.72	0.57	0.50	0.46	0.45

Table 4: Performance across individual cross-validation folds and evaluation on the full Czech dataset using all annotated training data.

and which one is used for testing. This highlights the instability of the data, and how some folds contain more challenging metaphorical expressions for the model to recognize than others.

Part of speech Table 5 shows that having metaphor annotations in the training data on all POS instead of only on nouns, verbs and adjectives improves the performance of the model on two of these classes (adjectives and verbs) while the performance on nouns drops only slightly in comparison. We offer two possible explanations for this: first, having more words annotated as metaphor provides the model with more signal to learn what exactly a metaphor is, helping to distinguish between all metaphors independently of the POS. Second, having more words annotated as metaphorical gives lower class imbalance, making it easier for the model to learn to say that a word is metaphorical.

Poetry The results on the Czech Poetry dataset are above chance, but quite low (see Table 6). This result points towards the difficulty of this dataset, which is supported by the moderate inter-annotator agreement. The domain shift from prose to poetry is likely another contributing factor.

5 Translatability of Metaphors

During the experiments, we observed that some metaphor types were consistently easier for the models to detect than others. Manual inspection suggested that this behavior may be related to metaphor conventionality and cross-lingual translatability. We define *translatability* as the possi-

Train + Test Data	Adj F1	Noun F1	Verb F1
All POS	0.50	0.46	0.45
Adj/Noun/Verb only	0.38	0.49	0.37

Table 5: Performance when training and evaluating on all metaphorical POS compared to training and evaluating only on adjective, noun, and verb metaphors.

Test Data	F1	Prec.	Recall	Adj F1	Noun F1	Verb F1
Poetry	0.41	0.37	0.47	0.46	0.40	0.57

Table 6: Metaphor detection performance on the Czech Poetry dataset.

bility of directly translating a metaphorical expression into English while preserving its metaphorical meaning. Conventionality describes whether a metaphor is established in the language or represents a more creative usage. While conventionality has received attention in previous work (Do Dinh et al., 2018; Neidlein et al., 2020; Reimann and Scheffler, 2024; Liang et al., 2025), to the best of our knowledge, translatability has not been studied in a computational setting.

In this section, we first evaluate transferring metaphor predictions through machine translation and word alignment. We then introduce an additional annotation layer of the Czech dataset for conventionality and translatability and analyze how these properties relate to model performance.

5.1 Translation and back alignment

In the previous experiments, we observed that models trained and evaluated on English data generally achieved the best performance. This can be caused by several factors, including the quality and size of the available English data, the fact that MIPVU was originally designed for English, or the bias of multilingual models towards English. Motivated by these observations, we tested whether metaphor predictions obtained from an English-trained model could be transferred to Czech through machine translation.

In this experiment, Czech texts were translated into English, predictions were obtained from an English-trained metaphor detection model, and these predictions were projected back to Czech using awesome-align (Dou and Neubig, 2021) word alignment. The preliminary results show that this approach with an F1 score of 0.31 does not reach the performance of a model trained on all available annotated data and evaluated directly on Czech (Ta-

Setup	F1	Precision	Recall
All annotated training data → Czech data	0.64	0.72	0.57
English training data → translated Czech data → alignment	0.31	0.38	0.26

Table 7: Comparison of metaphor detection on Czech data and predictions on Czech data transferred from English through machine translation and word alignment.

Status	Czech	English translation
Detected	<i>Že nás [nezaslepí] chamtivost</i>	<i>so that greed does not [blind] us</i>
Undetected	<i>ceny opravdu [platily]</i>	<i>the prices were actually [valid]</i>
Incorrectly detected	<i>obsluha kamen [čítala] tři lidi</i>	<i>it [took] three people to operate the stove</i>

Table 8: Examples of Czech metaphors that were detected and undetected by the English-trained model.

ble 7). However, the agreement between the two sets of predictions is higher ($F1 = 0.41$), suggesting that the approaches identify a partially overlapping set of metaphors. This may indicate that metaphors whose meaning is preserved during translation are easier to transfer across languages.

5.2 Manual Evaluation

To better understand the behavior of the models, we performed a manual error analysis of the predictions. We inspected examples correctly detected by the English-trained model, examples missed by the model, and cases where the model predictions appeared incorrect. We also compared errors across the different experimental setups. For examples, see Table 8.

The analysis suggested that metaphor detection performance is related to individual metaphor properties, particularly conventionality and cross-lingual translatability. Since translatability has received little attention in computational metaphor detection, we introduce an additional annotation layer for the Czech dataset covering these two properties.

5.3 Czech dataset annotation for Conventionality and Translatability

To analyze the relation between metaphor properties and model performance, we extended the Czech General dataset with annotations for conventionality and translatability. Conventionality was annotated as either *conventional* or *novel*. Although the distinction is not always clear-cut (Do Dinh et al., 2018), previous work suggests that novel metaphors are more difficult to detect automatically (Neidlein et al., 2020; Reimann and Scheffler,

2024).

We define translatability as the possibility of directly translating the metaphorical expression into English while preserving its metaphorical meaning. This definition does not include cases where a conceptual equivalent exists in English but requires different wording. A limitation of this definition is that a metaphor can preserve its meaning across languages while changing its degree of conventionality.

The author of this report annotated all metaphorical nouns, verbs, and adjectives in the Czech General dataset for conventionality/novelty and direct translatability to English. Other metaphorical POS were not annotated. Besides conventional and novel metaphors, a third category, *not a metaphor*, was introduced for cases that were originally annotated as metaphorical but were considered non-metaphorical during this annotation process. These cases consisted mainly of metonymies and several expressions used in their literal meaning.

The annotation was performed by a single annotator, therefore inter-annotator agreement is not available. A more extensive annotation with multiple annotators is left for future work.

5.4 Preliminary Findings

Table 9 shows preliminary results based on the annotations described in Section 5.3. The *not a metaphor* category is rarely detected by the models. In particular, models without Czech training data almost never identify these cases, suggesting that some of these examples may rely on Czech-specific annotations.

Novel metaphors are detected less frequently than conventional metaphors, similarly to previous

Training setting	Conventional		Novel		Not a Metaphor	
	DT	NDT	DT	NDT	DT	NDT
All annotated data	49/71	30/106	3/20	5/12	2/23	0/3
30k English, 30k Slovenian, 30k Spanish, 4k Czech	41/71	34/106	1/20	4/12	3/23	0/3
90k English, 4k Czech	37/71	32/106	3/20	7/12	5/23	1/3
90k Slovenian, 4k Czech	38/71	19/106	3/20	2/12	1/23	0/3
30k English, 30k Slovenian, 30k Spanish	16/71	11/106	1/20	1/12	0/23	0/3
90k English	22/71	15/106	2/20	1/12	1/23	0/3

Table 9: Number of correctly detected metaphors out of the total number of annotated metaphors in each category. DT = directly translatable, NDT = not directly translatable.

observations (Neidlein et al., 2020; Reimann and Scheffler, 2024). However, the number of novel metaphors in our dataset is too small to draw reliable conclusions. Most directly translatable novel metaphors come from a single text about animal athletes, where the metaphorical meaning of expressions such as *sprinter* depends on context introduced several sentences earlier. This highlights a limitation of the models’ context range.

Conventional metaphors provide the most informative comparison, as this is the only category large enough to observe systematic differences between training settings. Directly translatable conventional metaphors are detected substantially more often than non-directly translatable ones. When Czech data are included in training, performance on conventional metaphors increases in both categories, suggesting that Czech training data provide useful information for detecting both directly and non-directly translatable metaphors. The similar performance of the English + Czech and Slovenian + Czech models on directly translatable conventional metaphors suggests that this advantage may be related more to multilingual model properties, including English bias, than to English training data alone.

Finally, we observe that many metaphors annotated as not directly translatable have conceptual equivalents in English, but require different wording. Since previous work suggests that semantic similarity and synonyms can support metaphor detection (Neidlein et al., 2020), these cases may still be easier for multilingual models despite not being directly translatable at the word level.

5.5 Future Work

The analysis presented in this section is intended as the basis for a separate research paper. Future work will focus on creating higher-quality anno-

tations for Czech and extending the analysis to other languages. One possible direction is using LLMs to create initial annotations for additional languages, followed by validation and correction. This would allow a more language-independent analysis of which metaphor types are easier for multilingual models to detect. To the best of our knowledge, cross-lingual differences in metaphor detection based on translatability have not yet been studied.

6 Conclusion

This work explored metaphor detection in Czech using multilingual transformer-based models and cross-lingual transfer from languages with larger metaphor-annotated datasets. The results show that multilingual models can transfer metaphor knowledge to Czech, and that even a small amount of Czech annotated data substantially improves performance. The best results were achieved by combining multilingual training data with Czech annotated data, while direct transfer from other languages remains limited. The analysis also suggests that metaphor properties such as conventionality and translatability influence which metaphors are detected by the models.

The current performance is not sufficient for reliable use of metaphor detection in the EduPo educational application. However, the experiments provide insight into which types of data and training setups are useful for improving Czech metaphor detection. The analysis of conventionality and translatability will be further developed into a separate study focusing on how these properties affect cross-lingual metaphor detection.

References

- Špela Antloga. 2020. [Korpus metafor komet 1.0](#). In *Conference on Language Technologies and Digital Humanities*.
- Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. [Tuning multilingual transformers for language-specific named entity recognition](#). In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.
- Maria Berger. 2022. [Transfer learning parallel metaphor using bilingual embeddings](#). In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 13–23, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- 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.
- Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. [Weeding out conventionalized metaphors: A corpus of novel metaphor annotations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.
- Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Anna Hülsing and Sabine Schulte Im Walde. 2024. [Cross-lingual metaphor detection for low-resource languages](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 22–34, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.
- Jiahui Liang, Aletta G. Dorst, Jelena Prokic, and Stephan Raaijmakers. 2025. [Using gpt-4 for conventional metaphor detection in english news texts](#). *Computational Linguistics in the Netherlands Journal*, 14:307–341.
- Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. [On the language neutrality of pre-trained multilingual representations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. [Focal loss for dense object detection](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. [mmbert: A modern multilingual encoder with annealed language learning](#). *Preprint*, arXiv:2509.06888.
- Susan Nacey, Aletta G. Dorst, Tina Krennmayr, and W. Gudrun Reijnierse, editors. 2019. *Metaphor Identification in Multiple Languages: MIPVU around the world*, volume 22 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam and Philadelphia.
- Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. [An analysis of language models for metaphor recognition](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. [Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.
- Dalibor Pavlas, Ondřej Vrabec, and Jiří Kozmér. 2018. [Applying mipvu metaphor identification procedure on czech](#). In *Proceedings of the Workshop on Annotation in Digital Humanities*, pages 37–40, Sofia.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Sebastian Reimann and Tatjana Scheffler. 2024. [When is a metaphor actually novel? annotating metaphor novelty in the context of automatic metaphor detection](#). In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.
- Sebastian Reimann and Tatjana Scheffler. 2025. [Using large language models to perform MIPVU-inspired automatic metaphor detection](#). In *Proceedings of the*

2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pages 10–21, Vienna, Austria. Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. [Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection](#). In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2026. [Meta4XNLI: A cross-lingual parallel corpus for metaphor detection and interpretation](#). *Computational Linguistics*, 52(1):191–235.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. [A Method for Linguistic Metaphor Identification: From MIP to MIPVU](#). Number 14 in *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam.

Milan Straka, Jan Hajič, and Jana Straková. 2016. [UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Chuangdong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiye Li, Rongbo Wang, and Zhiqun Chen. 2020. [DeepMet: A reading comprehension paradigm for token-level metaphor detection](#). In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. [Metaphor detection with cross-lingual model transfer](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. [Cross-lingual metaphor detection using common semantic features](#). In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia. Association for Computational Linguistics.

Shenglong Zhang and Ying Liu. 2022. [Metaphor detection via linguistics enhanced Siamese network](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4149–4159, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

A Additional Experiments

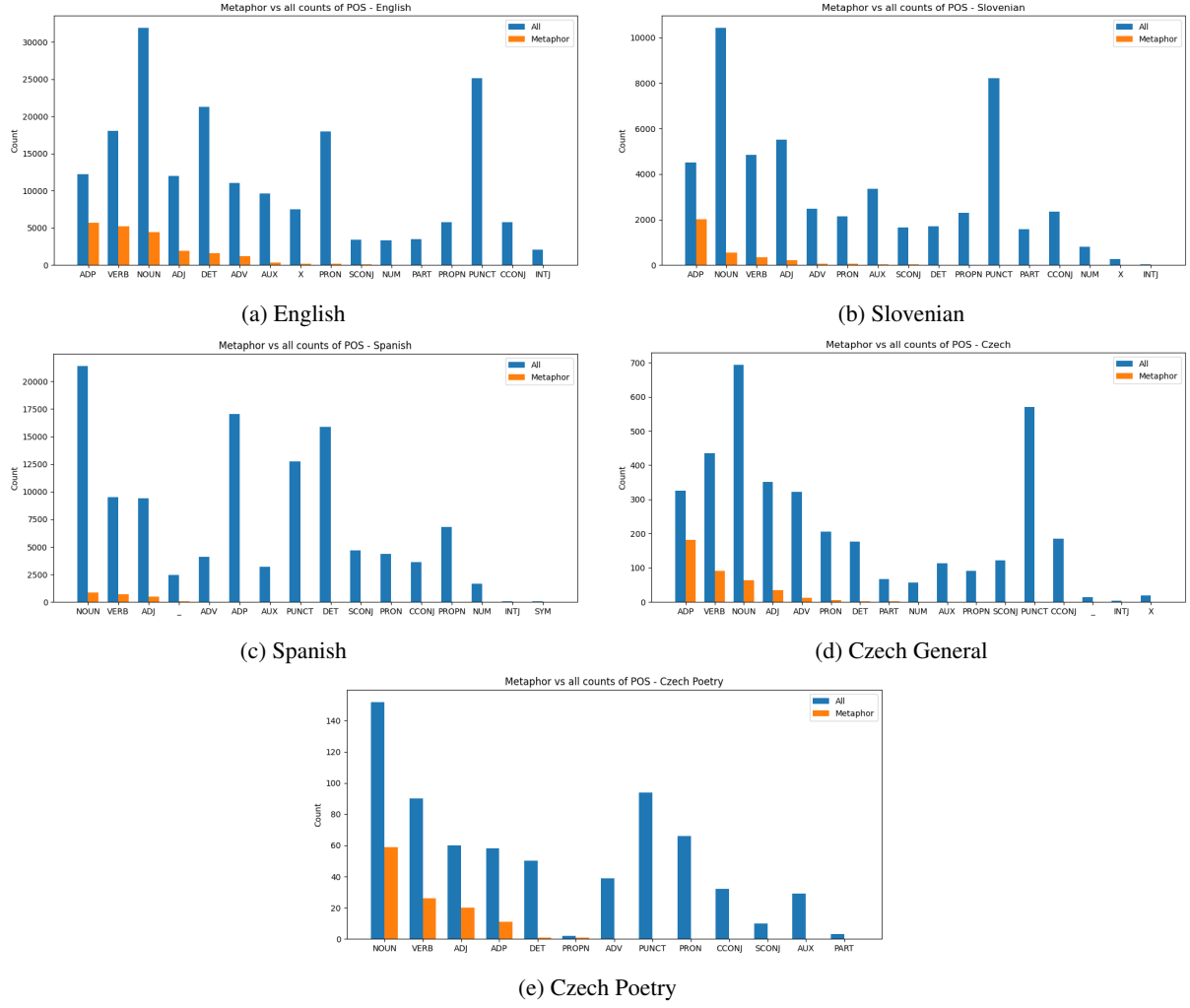


Figure 4: data pos percentage

Epochs	F1	Precision	Recall	Adj F1	Noun F1	Verb F1
7	0.663	0.706	0.625	0.36	0.54	0.49
5	0.648	0.699	0.604	0.40	0.59	0.39
9	0.632	0.705	0.573	0.50	0.58	0.41
8	0.628	0.711	0.562	0.29	0.44	0.34
4	0.626	0.663	0.594	0.40	0.53	0.39
6	0.623	0.732	0.542	0.25	0.50	0.34
3	0.607	0.581	0.635	0.33	0.48	0.41
10	0.604	0.699	0.531	0.25	0.36	0.43
2	0.565	0.446	0.771	0.30	0.33	0.57
1	0.505	0.371	0.792	0.31	0.24	0.53

Table 10: Hyperparameter search results for mmBERT with weighted loss (weight = 0.9) and different numbers of training epochs.

Epochs	F1	Precision	Recall	Adj F1	Noun F1	Verb F1
3	0.656	0.678	0.635	0.40	0.55	0.45
2	0.636	0.700	0.583	0.44	0.45	0.44
5	0.636	0.700	0.583	0.44	0.52	0.39
4	0.632	0.720	0.562	0.29	0.42	0.40
1	0.599	0.770	0.490	0.29	0.33	0.37

Table 11: Hyperparameter search results for mmBERT with weighted loss (weight = 0.5) and different numbers of training epochs.

Training language order	F1	Precision	Recall	Adj F1	Noun F1	Verb F1
en, sl, cs, es	0.674	0.763	0.604	0.29	0.62	0.46
es, en, sl, cs	0.670	0.709	0.635	0.40	0.56	0.48
en, sl, es, cs	0.659	0.740	0.594	0.25	0.56	0.47
en, cs, sl, es	0.595	0.694	0.521	0.29	0.36	0.40
sl, cs, es, en	0.588	0.676	0.521	0.25	0.48	0.37

Table 12: Selected language orders for multilingual training. Language abbreviations: en = English, sl = Slovenian, es = Spanish, cs = Czech.