# Transfer-Learning German Metaphors Inspired by Second Language Acquisition

**Maria Berger**
Ruhr University Bochum
maria.berger-a2l@rub.de

## Abstract

A major part of figurative meaning prediction is based on English language training corpora. One strategy to apply techniques to languages other than English lies in applying transfer learning techniques to correct this imbalance. However, in previous studies, we learned that the bilingual representations of current transformer models are incapable of encoding the deep semantic knowledge necessary for a transfer learning step, especially for metaphor prediction. Hence, inspired by second language acquisition, we attempt to improve German metaphor prediction in transfer learning by modifying the context windows of our input samples to align with lower readability indices achieving up to 13% higher F1 score.

## 1 Introduction

Figurative language detection is one of the most crucial tasks in the current digital conversational landscape. However, computationally, it remains also one of the most challenging tasks. Comprehensive resources to train computational models for figurative language detection are generally rare. Further, most existing work is performed on English language textual data. Some works investigate metaphor recognition in languages other than English (Sanchez-Bayona and Agerri, 2022; Aghazadeh et al., 2022).

We focus on applying and testing transfer learning techniques to continuously correct for this imbalance in figurative language prediction. We think that, due to the conceptual nature of metaphors (Lakoff and Johnson, 1980), it is possible to transfer metaphoric meaning given a sufficient amount of data that is capable of encoding this conceptual nature.

The study in this paper is designed as follows: First, we address the motivations of this research by presenting the readability indices of the predicted test samples of a prior study (Berger et al., 2024). Then, we modify the test samples according to these insights by trimming the observed contexts. This means, shortening the input. Last, we re-apply the multi-lingually pre-trained transformer models to determine how the sample modification affects the performance of the multilingual classifiers.

## 2 Related work

Tsvetkov et al. (2013, 2014) use lexical-semantic word features as well as bilingual dictionaries in several languages as input data for transfer learning to recognize metaphorical expressions across languages. Also, using syntactic patterns or abstractness scores is a common technique to identify or analyze metaphoric expressions (Tsvetkov et al., 2013; Clausen and Nastase, 2019).

Clausen and Nastase (2019) investigate the effect of text simplification on linguistic metaphor preservation (Wolska and Clausen, 2017; Clausen and Nastase, 2019). The authors provide an analysis of parallel text data that are simplified for different grade levels identifying whether metaphors are either preserved, rephrased, or dropped. They also investigate which features are capable of discriminating on whether a metaphor is preserved or dropped and determine that age-of-acquisition scores, imagine-ability, and concreteness scores are useful in the tasks.

Berger et al. (2024) perform a comprehensive study on applying transfer learning to German metaphor prediction, framing the problem as both a sequence labeling and sentence classification task. Several pre-trained transformers are fine-tuned on English metaphor-labeled data and tested regarding their capabilities to identify metaphors cross-lingually. However, multilingual



*Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 48–54

classifiers perform only moderately well, because the cross-lingual semantic knowledge that these models need to be capable of encoding appears to be hidden deep within the semantic representation of a language.

## 3 Methodology

We already learned that computational approaches work well in semantically "coarse-grained" tasks such as semantic similarity prediction (Kenter and De Rijke, 2015; Moritz and Steding, 2018; Wang et al., 2020) or authorship attribution (Benzebouchi et al., 2018), because they are well capable to distinguish the meaning of a word in different contexts. In figurative language identification, contextual representation is also a good input for a classifier to predict whether a word is meant figuratively or literally (Bizzoni and Lappin, 2017; Bizzoni and Ghanimifard, 2018; Liu et al., 2020).

Transfer learning typically makes use of a well-resourced source language to train a classifier on, afterwards, the trained model is applied to predict metaphors in a low/less-resourced language. However, there are two major problems to make figurative language prediction work cross-lingually: first, only a few larger (lexicon-dependent) annotated datasets for training in the source language are available[1]; second, the translation models of today's transformers are incapable to encode the deep semantic knowledge required for transfer identification of figurative language (Berger et al., 2024).

As syntax is the structural representation of meaning, one can carefully state that sentences of more complex syntax usually also entail more complex semantics. As such, "adding" tokens to a string also often (not always) means to "add" semantics to the meaning of a phrase or sentence. This can be partially validated by Batiukova and Pustejovsky (2013) who investigate the role of compositionality and lexical semantics in determining informativeness at the phrasal level.

As we understand that the transformer models may need to be presented with "easier" (shorter, less complex) samples because this is the case when learning a new language, we attempt to improve German metaphor prediction in transfer learning by modifying the context windows of our input samples to align with lower readability indices. In particular, trimming the context of a potential metaphoric expression can aid in the model's focus on nearby domain-related context while long-distance context may be less relevant, and the preserving of the sentence's global meaning possibly plays a subordinated role. To backup the latter assumption, we also test a more sophisticated technique by applying Klöser et al. (2024)'s GPT2.0-based text simplifier—to the best of our knowledge the only model, applicable for German language text—to our test data.

### 3.1 Transformer models and data (re-)used

We first recapture the zero-shot transferred results from a former study (Berger et al., 2024) that applies multilingual transformers mBERT (Devlin et al., 2018), XLM-RoBERTa (Liu et al., 2019), and sentence transformers (SBERT) (Reimers and Gurevych, 2019) to predict German metaphors from a small German language test set.[2]

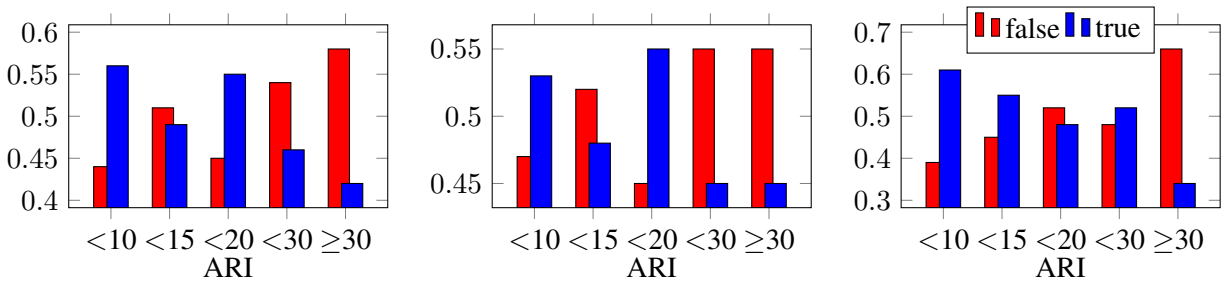
The pretrained transformer models were fine-tuned on the established English language VUA metaphor corpus (Steen et al., 2010) and tested on a smaller German metaphor dataset (Berger et al., 2024).[3] The task was designed as a sentence classification problem—inspired by Gao et al. (2018)'s embeddings approach—whereas every input was accompanied by the position of a verb in the sentence and the label whether that verb was used metaphorically (1) or literally (0) in the given context.

Note: Typically, metaphoric meaning predication normally is designed by token labeling or a word classification problem, not a sentence classification problem. Linguistically, however, it is common sense to identify a metaphor based on its source (image provider) and target (recipient of an image). The German test data that we use is a derivative of a linguistically annotated English language metaphor corpus that initially was annotated for lexical representatives of a metaphor

[1] The VUA Metaphor corpus (Steen et al., 2010), the TroFi corpus (Birke and Sarkar, 2006), and the MOH datasets (Mohammad et al., 2016) are among the larger ones.

[2] SBERT is an enhancement of the traditional BERT model, but it is specialized for problems of semantic similarity from sequential input (sentence) embeddings. It is only about half the size of the other two models and also trains/tests much faster.

[3] We use training:validation:testing data splits from the VUA corpus according to Gao et al. (2018) (15,516:1,724:5,873). These do not represent the most recent version of the VUA corpus, but enables us to compare our results with earlier results. As such, Gao et al. (2018) reach 58.9% F1 (acc. 69.1%) and 69.7% F1 (acc. 81.4%) with both their classifiers in a mono-lingual setup.




(a) mBERT (b) XLM-RoBERTa (c) SBERT

Figure 1: correct (blue) and wrong (red) prediction ratios grouped by 6 different ARI ranges

source and its target. This means, it does not provide labeled metaphoric meaning on the token level. But for trans-lingual metaphor classification (Gao et al., 2018), we can better compare our results with previous results based on this sentence classification set-up.

We use freely available bi-directional encoder representations from transformers instead of the emerging LLMs, because pre-trained BERT models are well-investigated and easily applicable for niche tasks such as transfer-learning for figurative language prediction in German.

### 3.2 Applying automated readability index

The grade level by Smith and Senter (1967), also known as automated readability index (ARI) is a well-performing measure for text complexity as it considers word and sentence complexity. We start by grouping correct and wrong predictions by the automated readability index (ARI) (Smith and Senter, 1967). We define five groups ranging the ARI below 10, lower than 15, lower than 20, lower than 30 and higher or equal to 30. These ranges approximately align with elementary school students (ARI $<$ 13), junior high school students (13-19 ARI), senior high school students (ARI 20-27), college students ($<$=28).[4] Figure 1 shows the predictions of the multilingual transformer models mBERT, XLM-RoBERTa and SBERT according to the groups of ARI scores the classification samples belong to. All of them show a strong correlation between false predictions and ARI scores. SBERT shows the most uniform curves.

### 3.3 Modifying input representations

As a pre-processing step, we simply trim our testing data to only allow a window of up to 3, 5 and 10 tokens to both sides of the given verb index, which results in a text snippet of 7, 11, 21 tokens respectively. This way we "simplify" our samples in a computationally easy manner. Because neural models that encode semantics of sentences as input representations cannot "understand" syntax—even though they can cope with it (de Dios-Flores et al., 2023)—it does not matter that our simplification approach ignores the tree-like structure of actual sentences. For typical neural classifiers, important features are mostly given by the contextual, especially sequential representations. Hence, trimming the surrounding longer-distance context makes the model stress close-context relations, which is especially important in figurative meaning prediction. Tab. 1 shows examples to illustrate how trimming modifies input samples.

Running Klöser et al. (2024)'s simplifier on our data removes metaphoric expressions or does not return a representation at all in almost 80% of the samples. Hence, we test metaphor prediction for the remaining 193 samples only.

## 4 Results and Discussion

Tab. 2 shows that the best performance increase can be reached with trimming the contextual span for the input representation to 11 tokens. Also, a context of 7 enables the models to drastically increase on performance while allowing a window of 10 to each side still results in an increase of up to 6% in F1 (c.f., upper part of Tab. 2). The neural simplification (that also preserves a sentence's meaning) achieves up to 9% increase in F1 returning the second highest F1 score.

Looking at samples from the SBERT output, we find that limiting the context can help the model to better focus on local meaning. For example in the form of not labeling words as figurative that actually are not used figuratively. While the following

[4]We refer to Smith and Senter (1967)'s grade level (GL). However, ARI score is more often used are common sense (see also Sec. 5).

| text | label | predicted | window |
|---|---|---|---|
| [...] auf der glücklichen Seite des Schweinetrogs stehen, **schmeckt** Demokratie ziemlich süß. | 1 | 0 | orig. |
| Seite des Schweinetrogs stehen , **schmeckt** Demokratie ziemlich süß . | 1 | 1 | 5 |
| *[...] on the lucky side of the pork trough, democracy **tastes** pretty sweet.* | | | |

Table 1: Sample sentences next to predictions; label 1: metaphorically meant; 0: literally meant

| model | approach | precision | recall | f1-score (+increase) | accuracy |
|---|---|---|---|---|---|
| mBERT | original | 58 | 46 | 52 | 50 |
| XLM-RoBERTa | sentence | 58 | 44 | 50 | 50 |
| SBERT | length | 57 | 65 | **61** | 51 |
| mBERT | | 67 | 62 | 65 (+13) | 61 |
| XLM-RoBERTa | window 3 | 65 | 60 | 62 (+12) | 59 |
| SBERT | | 67 | 72 | **69 (+8)** | 63 |
| mBERT | | 66 | 62 | 64 (+12) | 60 |
| XLM-RoBERTa | window 5 | 66 | 59 | 63 (+13) | 59 |
| SBERT | | 66 | 80 | **72 (+11)** | 65 |
| mBERT | | 63 | 53 | 58 (+6) | 55 |
| XLM-RoBERTa | window 10 | 60 | 50 | 55 (+5) | 52 |
| SBERT | | 60 | 73 | **66 (+5)** | 57 |
| mBERT | Klöser et al. (2024) | 70 | 51 | 59 (+7) | 52 |
| XLM-RoBERTa | simplified, | 71 | 40 | 51 (+1) | 48 |
| SBERT | 193 test samples | 70 | 70 | **70 (+9)** | 59 |
| mBERT | **fine-tuned** | 91 | 88 | **90** | 88 |
| XLM-RoBERTa | on **DE** metaphor, | 81 | 86 | 83 | 81 |
| SBERT | 98 test samples | 73 | 92 | 82 | 78 |
| mBERT | **fine-tuned** | 82 | 81 | 82 | 79 |
| XLM-RoBERTa | on **EN** metaphor, | 84 | 83 | **84** | 81 |
| SBERT | 908 test samples | 64 | 95 | 76 | 66 |

Table 2: precision, recall, f1, accuracy (%) according to a context of 7, 11, 21 tokens; trained on VUA corpus with train:val splits 15,516:1,724 tested on 908 DE language samples; upper part: original setup; mid part: input samples trimmed to window sizes and Klöser et al. (2024)'s simplification approach; lower part: fine-tuned on EN metaphor, splits: 1360:341:908 and fine-tuned on DE metaphor, splits: 720:90:98 (=908)

example was an FP before, it now is classified as TN:

*"Der Finanzmanager **erstellt**(TN) Finanzberichte [...]"* [The financial Manager **prepares** financial reports [...]]. Some could argue that "erstellt" might take the role of personification in the following context. This borderline example was labeled by SBERT as FP before. With the trimmed context, SBERT labels the examples as TN.

Regaeding TPs, the following example shows how SBERT can make better use of unusual relationships learned in the source language it was trained on. Hence, it interpretes the following example correctly in a figurative sense: *"[...] auf der glücklichen Seite des Schweinetrogs stehen, schmeckt(TP) Demokratie ziemlich süß."* [[...] on the lucky side of the pork trough, democracy **tastes** pretty sweet.]

For comparison, in the lower part of Tab. 2, we also list the results of fine-tuning in German, based on the 908 De language samples, which we split into train, validation and test sets (Berger et al., 2024). We can see that fine-tuning on target data and language after training in the VUA data brings the best results (Berger et al., 2024).

When we test whether fine-tuning on target domain English language data improves the test results in our German language data, we find a positive effect. Especially, XLM-RoBERTa shows the

ability to well generalize to language-independent data points when the source (training) and target (testing) domain remain the same. This can be explained by the dynamic masking process during RoBERTa's initial training process. However, in semantically challenging set-ups, this flexibly rather prevents RoBERTa from retrieving unknown items, as can be seen in the results of applying RoBERTa to our German metaphor data after only training on the VUA corpus (second line of Tab. 2).

| | | averaged ARI | |
|---|---|---|---|
| **model** | **window** | **correct** | **wrong** |
| mBERT | | 8.1 | 8.9 |
| XLM-RoBERTa | 3 | 7.8 | 9.4 |
| SBERT | | 8.1 | 9.1 |
| mBERT | | 8.6 | 9.3 |
| XLM-RoBERTa | 5 | 8.4 | 9.6 |
| SBERT | | 8.1 | 10.3 |
| mBERT | | 11.1 | 12.0 |
| XLM-RoBERTa | 10 | 11.1 | 12.0 |
| SBERT | | 11.1 | 12.2 |
| mBERT | | 13.4 | 13.5 |
| XLM-RoBERTa | simplified | 12.8 | 14.1 |
| SERT | | 13.5 | 13.5 |

Table 3: Averaged ARI score of the correct and wrong predictions after trimming

Table 3 shows the averaged ARI scores for the correct and wrong predictions of the three models. Almost every set-up shows that the averages of the ARI score are at least one point higher in the wrong predictions class compared to the correct predictions class. This inverse relationship between a model's ability to predict figurative language and ARI scores leads to the insight that certain lexical and textual properties—independent from the classifier—challenge the prediction of a verb's meaning in a given context. On the other hand, SBERT—our task-favorite—shows equal ARI scores in the simplification setup. It also is the model that reacts not as drastically to the trimming as the other models do. This hints us to investigate both more deeply, i) a model's translation representations, and ii) verbalization of metaphor in simpler sentence structures.

## 5 Remarks

**ARI was initially designed for English language text:** A possible weakness of this approach may be that the automated readability index (Smith and Senter, 1967) was originally designed to test students' capability to understand and comprehend the content of an English language text that also meets certain structural conditions. Because characters per word and words per sentence distribution differ across different languages, the grade-levels defined in 3.2 may not apply to our German language test data. However, Senter & Smith's score was used before to estimate the complexity reduction of text in languages other than English. For example in Moritz et al. (2016) and Tillman and Hagberg (2014). In the current study, we use the ARI score to obtain an understanding of prediction difficulty, and we think that applying the ARI score in this context is appropriate.

**Shortening is not simplification:** It is not always the case that a metaphor is difficult to extract because a sentence is syntactically complex, nor is it always true that humans understand shorter sentences better than longer ones. But, sentence simplification usually divides up complex content into many shorter sentences and this also improves metaphor recognition for a computational model. Further, our trimming approach is technically simple and streamlined and shows already good results. We further will elaborate on a quantitative approach that incorporates advanced syntax-tree rules into our window-trimming technique.

## 6 Conclusion

We demonstrated a computationally simple approach to correct input representation to make them shorter, hence, easier for the model to understand, because—as in second language acquisition, we learned that the translation representations of transformer models have some difficulty in "understanding" the deep semantics required for figurative meaning classification. We also applying a GPT-based simplifier. We achieve an increase of up to 13% (11-token context) and up to 9% (neural simpl.) in F1 and find that the sentence transformer models perform best in metaphor prediction. In future, we plan to apply didactically-informed approaches that utilize linguistic, comparative, and didactic knowledge while being applicable to quantitative methods as well.

## References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained lan-

guage models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050.

Olga Batiukova and James Pustejovsky. 2013. Informativeness constraints and compositionality. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 92–100.

Nacer Eddine Benzebouchi, Nabiha Azizi, Monther Aldwairi, and Nadir Farah. 2018. Multi-classifier system for authorship verification task using word embeddings. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Maria Berger, Sebastian Michael Reimann, and Nieke Marie Kiwitt. 2024. Applying transfer learning to german metaphor prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1383–1392.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European chapter of the association for computational linguistics*, pages 329–336.

Yuri Bizzoni and Mehdi Ghanimifard. 2018. Bigrams and bilstms two neural networks for sequential metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 91–101.

Yuri Bizzoni and Shalom Lappin. 2017. Deep learning of binary and gradient judgements for semantic paraphrase. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Yulia Clausen and Vivi Nastase. 2019. Metaphors in text simplification: To change or not to change, that is the question. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 423–434.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Iria de Dios-Flores, Juan Garcia Amboage, and Marcos García. 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 203–222.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. *arXiv preprint arXiv:1808.09653*.

Tom Kenter and Maarten De Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1411–1420.

Lars Klöser, Mika Beele, Jan-Niklas Schagen, and Bodo Kraft. 2024. German text simplification: Finetuning large language models with semi-synthetic data. *arXiv preprint arXiv:2402.10675*.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 250–255.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the fifth joint conference on lexical and computational semantics*, pages 23–33.

Maria Moritz, Barbara Pavlek, Greta Franzini, and Gregory Crane. 2016. Sentence shortening via morpho-syntactic annotated data in historical language learning. *Journal on Computing and Cultural Heritage (JOCCH)*, 9(1):1–9.

Maria Moritz and David Steding. 2018. Lexical and semantic features for cross-lingual text reuse classification: an experiment in english and latin paraphrases. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nils Reimers and Iryna Gurevych. 2019. https://aclanthology.org/D19-1410 Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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. Association for Computational Linguistics.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories (U.s.), Aerospace Medical Division,Wright-Patterson Air Force Base: 1–14. PMID 5302480. AMRL-TR-6620.

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, and Tina Krennmayr. 2010. Metaphor in usage. *Cognitive Linguistics 21–4*.

Robin Tillman and Ludvig Hagberg. 2014. Readability algorithms compability on multiple languages. KTH.

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.

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.

Congcong Wang, Paul Nulty, and David Lillis. 2020. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th international conference on natural language processing and information retrieval*, pages 37–46.

Magdalena Wolska and Yulia Clausen. 2017. Simplifying metaphorical language for young readers: A corpus study on news text. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 313–318.