A Novel Dataset and Three Ways to Approach Automatic Metaphor Detection in German Religious Online Forums

Sebastian Reimann, Tatjana Scheffler
Ruhr University Bochum
Department for German Language and Literature
{firstname.lastname}@rub.de

Abstract

In recent years, automatic metaphor detection has received considerable attention within NLP. However, the largest share of research, including most datasets annotated for metaphor, has concentrated on English and a limited set of genres. Automatic metaphor detection for a genre like religious online communication, which is particularly rich in metaphor, remains understudied, in particular since annotated data for this genre is lacking in the first place. This paper aims to close these gaps by offering a novel dataset of posts from German online forums annotated for metaphor, which opens up new research opportunities for automatic metaphor detection for German. Moreover, we present an in-depth exploration in which we evaluate the suitability of different strategies to overcome the relative lack of training data for this task by comparing cross-lingual and cross-genre transfer strategies with the use of LLM prompting. We find that fine-tuning encoder-only language models outperforms the prompting-based approach, that different architectures based on contextual embeddings indeed exhibit considerable differences in their behavior and that smaller in-genre data may be preferable for certain use cases over fine-tuning on larger datasets from different genres.

Keywords: metaphor, figurative language, cross-lingual transfer, online forums, social media

1. Introduction

Metaphoric language is more than just a stylistic tool. It rather points to underlying mappings between semantic domains that structure the way humans perceive the world (Lakoff and Johnson, 1980). For some genres, metaphor is of particular importance, since it serves to describe what otherwise cannot be put into words. For example, in religious communication, transcendent and supernatural entities like God must be described using metaphors from the physical world, such as through describing God as a *father* and believers as *children* (Krech et al., 2023). Religious communication (together with other specialized discourse such as science communication) consequently represents a particularly fruitful use case for the application of automatic metaphor detection methods. Such solutions would also enable potentially new interdisciplinary and quantitative perspectives for the study of religion. However, while automatic metaphor detection has indeed received more and more attention in recent years, the best performing approaches (Choi et al., 2021; Babieno et al., 2022; Li et al., 2023) were almost exclusively applied in scenarios where large amounts of training data that matched the test data in terms of both language and genre were available; particularly focusing on the VU Amsterdam Metaphor Corpus (Steen et al., 2010) (VUAMC), which consists of texts from news, everyday conversation, light fiction and academic discourse. The application to datasets from other genres, as well as cross-lingual applications that involving testing on data in other languages than English, remains scarce. Particularly for German, there is a lack of resources. Existing metaphor datasets for German are either very small and machine-translated versions of English datasets (Berger et al., 2024), not yet publicly available (Egg and Kordoni, 2023) or focus on one specific word class such as particle verbs only (Köper and Schulte im Walde, 2016). Additionally, the study of Reimann and Scheffler (2024) is the only work applying metaphor detection to religious texts.

More recently, there has been a shift from supervised finetuning to zero- and few-shot prompting of generative Large Language Models (LLMs) for many NLP tasks. In theory, this presents new opportunities in cases where data is scarce. However, metaphor detection appears to be a task that still poses problems for generative LLMs (Chen et al., 2024; Liang et al., 2025; Reimann and Scheffler, 2025a,b), which raises questions on how to deal with the data shortage for our specific use case of metaphor detection in Christian online forums. Can LLM prompting be a viable option for scenarios where training data does not match the target test data in language and/or genre or are supervised transfer learning scenarios the better option?

More specifically, to answer the aforementioned issue, we make the following contributions in this paper:

1. We make a novel dataset obtained from German Christian online forums publicly available[1].

[1] https://github.com/SFB-1475/

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 106–118
11 May 2026.

This dataset is, to the best of our knowledge, the largest publicly available German dataset annotated for metaphor and will thus counteract the lack of high-quality datasets for this particular language and genre combination.

2. We explore cross-lingual transfer potentials for German metaphor detection with multilingual, encoder-only language models and English as a source language. We find that such transfer indeed works if the underlying cross-domain mapping is common in both languages.

3. We share the first study that explicitly contrasts cross-lingual, cross-genre transfer and prompting based approaches for automatic metaphor detection. In a supervised setting, we find that using data from similar genres but different languages may indeed be preferred, even if the datasets are smaller and cross-lingual transfer is necessary.

2. Previous Work

In recent years, the most successful body of research on automatic metaphor detection (Mao et al., 2019; Choi et al., 2021; Babieno et al., 2022; Zhang and Liu, 2023) used contextual word embeddings to model the procedures for manual metaphor identification, in particular the Metaphor Identification Procedure (Pragglejaz Group, 2007) (MIP) concerned with the contrast between the meaning of a word in its context and its more concrete basic meaning, and the Selectional Preference Violation (SPV) approach (Wilks, 1975) concerned with the semantic difference between a word and its context. However, these methods for automatic metaphor detection focused almost exclusively on English.

The earliest cross-lingual approach was put forward by Tsvetkov et al. (2014) who use vectors that incorporate features for abstractness, imageability and supersenses in order to detect metaphorical subject-verb-object and adjective-noun constructions with a random forest classifier. Non-English examples are translated before the are vectorized. The F1-scores achieved by their approach remain relatively constant across all four languages (English, Spanish, Russian and Farsi), demonstrating potential for cross-lingual transfer.

Despite the promising early results of Tsvetkov et al. (2014), cross-lingual metaphor detection has been relatively underexplored until recently. Besides several other probing experiments, Aghazadeh et al. (2022) tested the cross-lingual transfer capabilities of pretrained language models for metaphor detection. For these cross-lingual experiments, they used data from the multilingual LCC Corpus (Mohler et al., 2016) containing sentences in English, Spanish, Russian and Farsi. For each language, they finetuned XLM-RoBERTa (XLM-R) on equal portions of the corpus and tested them on data from all languages in the corpus. Although training and testing on the same language always yielded the highest accuracy, for all cross-lingual transfer settings, the models outperformed a random baseline, suggesting that some transfer occurred successfully.

Sanchez-Bayona and Agerri (2022) present CoMeta, a large Spanish corpus annotated via MIPVU, the same guidelines used for the annotation of the VUAMC. They carry out cross-lingual metaphor detection experiments, where they finetune the most successful models from monolingual experiments (XLM-R for the evaluation on Spanish and DeBERTa for the evaluation on English) in a cross-lingual transfer setting, from Spanish to English and vice versa. Training on Spanish and evaluating on English expectedly underperformed the monolingual results. However, finetuning on English and evaluating on Spanish even outperformed the monolingual Spanish setting, which they attribute to the larger size of the English training set compared to its Spanish counterpart.

For metaphor detection for Slovene, Klemen and Robnik-Šikonja (2023) also included a series of cross-lingual experiments, besides monolingual experiments, where they train on either English data from the VUAMC or a combination of English and Slovene data and test on Slovene, using XLM-R and a trilingual model pretrained on English, Slovene and Croatian data (Ulčar and Robnik-Šikonja, 2020). However, they find that English data only had a minor effect on performance.

Berger et al. (2024) used a version of the English metaphor corpus of Gordon et al. (2015) that was translated into German and reannotated via MIPVU to explore the cross-lingual transfer for metaphor detection. They approached the task in two ways: as a sequence classification task providing one label for each word in a sentence and a classification task providing a label for the entire sentence on whether it contains a metaphoric word. In all experiments, they used VUA for training and evaluated both the original English data from Gordon et al. (2015) and their German translations. For the sequence classification task the performance was low overall. For sentence classification, they report more positive results, with SBERT only underperforming a monolingual evaluation by 5 points in F1 (66% vs 61&).

Hülsing and Schulte Im Walde (2024) apply cross-lingual transfer to verb metaphor detection in low-resource scenarios with English as a source language and German (using the particle verb data of Köper and Schulte im Walde (2016)), Latin and

C04-Cross-Lingual-MD

Russian as target languages. They train multilingual BERT in both a zero-shot setting with only English data involved in training as well as in a few-shot setting with 20 target language examples and with the multilingual adapter MAD-X on English and evaluate its performance on German, Latin and Russian. They report mostly satisfactory results, with F1-scores ranging from 60% to 87%, with the best performances on Russian with the MAD-X adapter. Introducing the target language examples, surprisingly, did not bring any improvement.

3. Data

3.1. English

For our English data, we rely on preexisting datasets annotated for metaphor via MIPVU. In order to have a dataset that corresponds in terms of genre with our German data, we use the data of Reimann and Scheffler (2024), consisting of posts from two Christian subreddits.

Additionally, for the comparison of cross-lingual and cross-genre transfer, we include a large dataset from a different genre. Here, the logical choice is the version of the VUA corpus (Steen et al., 2010) used in the 2020 Metaphor Detection Shared Task (Leong et al., 2020). Both the Reddit dataset of Reimann and Scheffler (2024) and the version of VUA used in Leong et al. (2020) only included content words (nouns, verbs, adjectives and adverbs) in their metaphor annotations. The VUA corpus is available under a CC BY-SA 3.0 license and the Reddit data is available under the CC-BY-4.0 license.

3.2. German

For our German dataset, we searched the religious German online forum *jesus.de* for threads where the word *Metapher* ("metaphor") is mentioned, in order to capture as many examples as possible that contain metaphors that were deliberately used as such.

We annotated our data following the MIPVU protocol (Steen et al., 2010), in particular its variant for German (Herrmann et al., 2019), which makes additional recommendations on how to deal with German morphology and the most suitable German dictionaries (Duden and DWDS). We only annotate nouns, verbs, adjectives and adverbs. For its basic procedure, MIPVU asks the annotator to perform the following steps:

1. Read and understand the text
2. Divide the text into lexical units
3. On a word by word basis, define the contextual meaning (3a), decide if a more basic (i.e., more concrete, human-oriented) meaning exists in a corpus-based dictionary (3b), and decide if the two meanings are sufficiently distinct but still related by some sort of similarity (3c)

If these conditions are met, the word is considered to be a *indirect Metaphor-Related Word (MRW)*. Additionally, MIPVU covers so-called *direct* MRWs, which are metaphoric words that are part of a metaphoric comparison (e.g., "strong like a lion"). Steen et al. (2010) argue that they are used in their literal meaning and thus not covered by the standard procedure but they still express a mapping between two domains.

Four annotators were involved in the annotation process, a PhD student of computational linguistics and a PhD student of religious sciences, as well as two student assistants, one with a background in philosophy and English linguistics and one with a background in religious sciences. All annotators are native speakers of German. Initially, we carried out an annotation round to familiarize ourselves with the data as well as the annotation guidelines with all annotators and extensively discussed all cases of disagreement in the curation process.

One aspect that caused initial confusion among annotators were frequently occurring lexicalized expressions in German such as *es gibt* ("there is", lit. "it gives") and *es geht mir gut* ("I'm doing fine", lit. "it goes me fine"), which convey some sense of figurativeness and may possibly be related by metaphor when looking at them through the lens of historical linguistics. However, MIPVU emphasizes to assume a synchronic, modern-day perspective. From this, we argue that these uses of *gibt* and *geht* and their corresponding forms are not related by similarity with their more basic meanings.

The largest portion of our data was annotated by the two student assistants. Between the two student assistant annotators, we report a substantial agreement ($\kappa = 0.6$), close to the agreement reported in Sanchez-Bayona and Agerri (2022, $\kappa = 0.63$). Additional posts were annotated by students of German linguistics within a seminar on metaphor. In all cases, the first author of this paper took care of the data curation, in close discussion with the annotators to resolve disagreements. In all scenarios, INCEpTION (Klie et al., 2018) was used as the main tool for annotation.

An overview of the datasets used in our experiments is given in Table 1.

4. Experimental Setup

Our main goal is to investigate methods to overcome the scarcity of data that matches our use case of posts from German religious online forums. The largest portion of this paper is dedicated to the investigation of cross-lingual transfer potentials for

Dataset	Lang.	Tokens	MRWs (%)
DE_CHR	DE	14,222	2,734 (19.22%)
EN_CHR	EN	14,981	3,170 (21.16%)
EN_VUA	EN	72,611	13,070 (18.00%)

Table 1: Overview of our data.

automatic metaphor detection within the context of genre differences. Consequently, in our experiments we train and evaluate on various combinations of English and German data. Given the lack of research on automatic metaphor detection for German, we will use the German dataset as a test set and train the models on either the smaller in-domain Reddit data or the larger, but out-of-domain VUA20 data. However, in order to gain further insight on cross-lingual transfer for this task, we also use the German data set as training set, evaluate on the English Reddit data and compare these results to the results reported in Reimann and Scheffler (2024) for training on VUA and testing on Reddit. In addition to the supervised experiments, we will also further develop previous approaches to metaphor detection with generative LLMs with in-context learning and without the use of additional training or finetuning.

Supervised Transfer Baseline. In our experiments, for the supervised and transfer learning based approach we design a simple baseline, the token-based architecture that was already used as a baseline in the 2020 Metaphor Detection Shared Task (Leong et al., 2020). It uses the contextualized BERT embeddings of a sentence, which are given to a linear layer, where a softmax predicts a label for each token (here: metaphoric or not) (BASE_XML-R). Instead of the monolingual BERT model, we use the multilingual XLM-R by Conneau et al. (2020).

Transfer Experiments. As our main experiment, we test two architectures for metaphor detection that claim to be inspired by linguistic theories and modify them slightly to enable cross-lingual transfer. On the one hand, we use MelBERT (Choi et al., 2021). It uses two transformer encoders, one to encode the entire sentence and one for the target word only, to replicate MIP and SPV. For MIP in MelBERT, the contrast between basic and contextual meaning is modeled by concatenating the contextual embedding of the target word and the embedding of the word in isolation. For SPV, Choi et al. (2021) model the contrast between a word and its context by concatenating the contextual meaning of the word and the embedding of the [CLS] token, representing the entire sentence. Both concatenations are in the end given to a linear classifier. We, however, replace the monolingual RoBERTa used in the original MelBERT with XLM-R. We prefer the original MelBERT over modifications that model the basic meaning through dictionary entries (Zhang and Liu, 2022; Babieno et al., 2022) or extract it from literal examples in the training data (Li et al., 2023). These approaches rely on either language-specific resources or training and test data in the same language and are thus not suitable for cross-lingual experiments.

As a second model, we use AdMul (Zhang and Liu, 2023). It is also influenced by MIP but approaches this in a different way. In a multi-task learning framework, they combine metaphor detection with an auxiliary task called basic sense detection, predicting if the word is used in its basic meaning. For this auxiliary task, they transform the the Sense Disambiguation (WSD) dataset of Raganato et al. (2017) into a binary dataset, where word uses with the most common sense are labeled as used in their most basic sense. The model is trained on both tasks and a global discriminator aligns the representations. We replace the DeBERTa (He et al., 2021) model of AdMul with its multilingual counterpart. For the basic meaning disambiguation task, we additionally add the German fraction of XL-WSD (Pasini et al., 2021), a multilingual extension of the WSD framework by Raganato et al. (2017). For comparison purposes and and due to a lack of separate validation data for additional hyperparameter tuning, we used the optimal hyperparameters reported in Choi et al. (2021) and (Zhang and Liu, 2023).

In all experiments, we use the available models from HuggingFace (Wolf et al., 2020). We ran all models on a NVIDIA A30 Tensor Core GPU. In total, running the experiments took approximately 90 hours: 47 hours for running the LLM experiments and 43 hours for running the experiments with the XLM-R and DeBERTa based models.

LLM Prompting. Additionally, we further develop a prompting-based LLM method for automatic metaphor detection that does not make use of any additional training data for finetuning. For this we use the method of Reimann and Scheffler (2025b) as a starting point, in particular the version that uses a one-shot prompt with an example in the last prompt. We modify their series of prompts using the insights of Hicke and Kristensen-McLachlan (2024), who observed positive results with a prompt that replicates step 3b of MIPVU, the identification of a more basic meaning. Reimann and Scheffler (2025b) modeled this step with two prompts, one to ask the model to generate a dictionary entry and another to ask which, if any, may be consid-

ered more basic according to MIPVU. We replace these two prompts with the one used by Hicke and Kristensen-McLachlan (2024), asking the model if a more basic meaning is available and, if yes, to briefly define this meaning.We provide the prompts in Appendix A. Reducing this question on a more basic meaning to one prompt would make the procedure of Reimann and Scheffler (2025b) cheaper with respect to computational costs and potentially less prone to error propagation. For all runs of our LLM approach, we use the 8B version of LLaMa 3.1, which is explicitly labeled as multilingual and among the better performing models in Reimann and Scheffler (2025a). We aim for a lightweight approach that is relatively inexpensive in computational resources. Consequently , we prefer the lightweight 8B version of LLaMa 3.1 over its 70B version.

5. Results

Table 2 shows the results for all models. First, we can see that the fine-tuned encoder-only models again outperformed the one-shot LLM, even in scenarios where training and testing data were coming from different languages. Second, the results show that cross-lingual transfer for metaphor detection is to some extent possible, with F1-scores mostly above 60%. Regarding the choices of training data, finetuning the XLM-R based models on the much smaller in-genre dataset notably outperformed finetuning on VUA for German. For evaluation on the English Reddit data, the results are more nuanced. For the simple sequence classification approach and MelBERT, the differences (between cross-genre and cross-lingual transfer) in performance are small with a slightly better precision achieved by the models trained on the smaller in-genre dataset in German. This is surprising, since the results of Reimann and Scheffler (2024) would rather suggest improvements with respect to recall. However, the best performance on the English data are achieved by AdMul with German training data and again, improvements in precision playing a major role. Overall, cross-lingual transfer (from a small in-domain dataset) shows better performance over cross-genre transfer (from a large general dataset) even for the English test set.

Finally, in terms of differences between models, we mostly observe moderate improvements over the simple XLM-R baseline by the more elaborate MelBERT and AdMul architectures, which do not fully reflect the improvement that the models achieve in monolingual settings in the reported literature. Between the multilingually adapted versions of MelBERT and AdMul we observe only minor differences in F1. However, precision and recall signal that the models indeed behave differently. When finetuned on English data and evaluated on German, AdMul appears to overgeneralize the metaphor label given the comparatively low precision, compared to the high recall values. This does not seem to be a problem for MelBERT, since it rather struggles with finding metaphorical examples in the first place. These behavioral differences also come to light when comparing the different transfer settings. For AdMul, training on the German forum dataset (DE_CHR) led to notable improvements in precision over training on VUA, while recall remains constant. However, for MelBERT, a drop in recall can be observed. In conclusion, based on the metrics, it can be stated that for finetuning on smaller datasets, AdMul appears to be the best option, where it even outperforms all models finetuned on VUA for both datasets we evaluate on.

6. Error Analysis

In order to better understand the results reported in Section 5, we conduct an extensive error analysis, considering the metaphoric words that were most commonly missed by the models (false negatives) and the literal words that were most often falsely considered to be metaphoric (false positives). For all models, these results are shown in Table 3.

We can see multiple tendencies. When finetuned on VUA, several MRWs that are characteristic for religious language are not recognized. Among the top five false negatives for MelBERT, we have both *Vater* (“father”) as a metaphor for God and its corresponding genitive form *Vaters*, echoing the results of Reimann and Scheffler (2024), as well as *Willen* (“will”), utilised in the sense of “will of God”. This is similar for AdMul, with *Vater*, as well as *Beziehung* (“relationship”), often referring to a relationship with God, and *Herrn* (“lord”) for God among the most common false negatives. Using the in-genre dataset resolves this for MelBERT and AdMul, where the number of unrecognized examples of *Vater* metaphors drops to three and zero, respectively. Interestingly, the baseline model did not profit that much from seeing the metaphorical sense of *Vater* referring to God. Smaller improvements can be seen, with 11 false negatives compared to 26 when finetuning on VUA.

In the scenario of training on the English Reddit data and testing on the German forum data, we, however, notice a sharp increase of unrecognized conventional metaphors such as *machen* (“to make” or “to do”), *klar* (“clear”) and *sagen* (“to say”). The latter may be linked to the modality of online forums. When looking up *sagen* in the German dictionary Duden, we would, on the one hand, find a sense describing the bodily action of articulating words and a more abstract sense that just describes the act of expressing or formulating something. When

Training Data	Model	DE_CHR			EN_CHR		
		P	R	F1	P	R	F1
None	BASE_LLaMa	20	63	31	22	93	36
DE_CHR	BASE_XLM-R	-	-	-	72	55	63
	MelBERT	-	-	-	68	53	59
	AdMul	-	-	-	**74**	61	**67**
EN_CHR	BASE_XLM-R	60	67	63	-	-	-
	MelBERT	**64**	62	63	-	-	-
	AdMul	56	**78**	**65**	-	-	-
EN_VUA	BASE_XLM-R	56	50	53	67	55	60
	MelBERT	56	52	54	67	54	60
	AdMul	41	72	54	57	**67**	61

Table 2: Overview of precision, recall and F1-score of the models with different combinations of training (rows) and test (columns) data.

	EN_VUA -> DE_CHR		EN_CHR -> DE_CHR		DE_CHR -> EN_CHR	
	FP	FN	FP	FN	FP	FN
Base XLM-R	da 104	Willen 26	geht 28	klar 14	father 9	thing(s) 90
	geht 33	Vater 26	lassen 12	sagen 13	see 8	way 44
	hier 26	Vaters 14	finde 12	Willen 11	start 8	feel 26
	lassen 10	Dinge 14	bleiben 11	Beziehung 11	keep 6	part 16
	Situation 9	sagen 13	wählen	Vater 11	say 6	called 15
Mel-BERT	da 108	Vater 26	geht 29	machen 14	father 10	thing(s) 83
	hier 39	Willen 26	lässt 12	sagen 13	start 9	way 40
	geht 35	Vaters 14	lassen 9	klar 11	say 8	feel 26
	gibt 15	sagen 13	Gefahrenabwehr 8	Ansatz 10	find 7	away 20
	lassen 14	Dinge 10	Pocken 7	Ansicht 9	keep 6	back 17
Ad-Mul	da 138	Vater 25	geht 25	sagen 12	father 17	feel 27
	dann 52	Beziehung 19	lassen 20	Willen 10	born 12	thing(s) 24
	gibt 39	Willen 18	Gewalt 15	Sinn 8	has 12	get 20
	hier 39	Vaters 14	bleiben 12	Eindruck 7	still 12	called 15
	mutiert 36	Herrn 8	Gebote 9	Sump 7	mother 9	elect 14

Table 3: Most frequent false positives and false negatives produced by the models across transfer scenarios.

referring to something in a forum post like in Example 1, the second sense is more suitable. However, the former sense represents a more basic meaning according to MIPVU that is also sufficiently distinct and which is arguably related by similarity, thus it would fulfill both criteria of MIPVU to be considered metaphorical.

(1) Und wenn Du nun sagst : Beleg es mir – dann kann ich nur sagen : kann ich nicht .
"And if you now say : Prove it to me – then I can only say : I can't ."

Ansatz, in the sense of "attempt" or "approach" represents an error that may be explained via cross-linguistic differences. In German, *Ansatz* may also refer to the base of something mechanical like a pipe. In the Duden, we find both this meaning and *Ansatz* referring to an attempt. As structural similarities may be drawn between the two concepts, and as the former represents a more concrete, thus basic meaning according to MIPVU, this renders *Ansatz* ("attempt, approach") to be an MRW according to MIPVU, even though it is a very conventionalized one. In English, there is no word expressing *attempt* in a metaphorical way, which would explain why models trained on English data never consider the immaterial sense of *Ansatz* to be metaphoric.

Similar observations can be made when looking at the false positives, i.e., words wrongly marked as metaphorical. For all models, using the German forum data in training and the English Reddit data in evaluation led to overgeneralization regarding family terms such as *father* and *mother*, where even literal usages of these terms were often considered metaphoric. Another striking observation regarding false positives can be made for the scenario using VUA as training data, which explains the previously observed drops in precision. Here, the adverb *da* ("there") occurs as the most frequent false negative for all models. Similar to English *there*, *da* may

occur in a spatial sense, referring to a location, as well as a temporal sense, referring to a point in time, such as the use of *da* in Example 2, or referring to situations like in Example 3. It may indeed be argued that the spatial use of *da* represents a more basic meaning according to MIPVU and that the uses of *da* referring to times and situations are related by some sort of similarity. However, none of the human annotators actually considered *da* to be metaphoric, thus raising doubts if this similarity is, from the perspective of a German native speaker, still apparent for the modern day language user.

(2) Es gibt Zeiten, da ist es egal [...].
"There are times when it does not matter [...]."

(3) Wozu brauche ich da einen eigenen Willen?
"For what do I need my own will there?"

In VUA, we can indeed find 27 instances of English *there* that were labeled as MRWs. Among these are Example 4 with *there* expressing a temporal meaning and Example 5 referring to a situation. Our results thus strongly suggest that, from these examples, the models learned that *da* in such instances may be considered metaphoric.

(4) The building society will be staffing a mortgage desk at each auction, and says buyers could arrange finance there and then, subject of course to proof of income and status.

(5) Seriously, I'd fucking have it out of there, everything I own.

Another aspect that needs to be considered is the impact of the auxiliary basic sense detection task in AdMul, which represents also the biggest architectural difference between the models. For this, we look at the five words, for which we see the biggest improvement in terms of recall when choosing AdMul over MelBERT in the in-genre cross-lingual transfer scenario and then look at the output of the basic sense detection task of AdMul for these words. For English, they are *thing*, *way*, *feel*, *away* and *back* and for German those are *machen* ("make"), *sagen* ("say"), *klar* ("clear"), *Ansatz* ("approach") and *Ansicht* ("view"). According to the linguistic theory, the metaphoric sense is never the more basic sense of a word and thus, for metaphoric words, the model should not consider them to be used in their basic meaning. Indeed, in metaphoric instances of the aforementioned words, the model predicted that they are not used in the basic sense, suggesting that it may have indeed learned valuable information for metaphor detection from the auxiliary task.

7. Discussion

From our results, we can see that architectures inspired by linguistic procedures for metaphor detection are to some extent also an option for cross-lingual metaphor detection. They also clearly outperformed our LLM-based approach, suggesting that, as long as there is training data available, such supervised approaches may be preferred. For MelBERT, we do not observe the improvements reported in Choi et al. (2021) over the simple sequence classification baseline. However, AdMul provided the strongest results overall, even though it sacrificed precision in the scenarios that involved transfer from English to German. We hypothesize that the predictions of AdMul are more closely dependent on what it saw in finetuning since our error analysis in the previous section has shown that the most frequent cases of false positives it produced were mostly the same as the Baseline and MelBERT but with a notably higher frequency.

Our results moreover underline the findings of Reimann and Scheffler (2024) that genre differences and consequently the metaphors and domain mappings represented in the training data are a major factor in metaphor detection. Since it reflects similar patterns with respect to metaphors describing God as a human, we can also see a notably better performance on the German forum data when using the small dataset from Christian subreddits for finetuning, compared to the large VUA dataset. Vice versa, finetuning on the German dataset and testing on the English Reddit data did not result in large improvements overall, which may however be expected, given its size and the fact that linguistic metaphors are still relatively language-specific. However, despite these constraints, two of the three models performed better on EN_CHR when trained on the smaller German data and the error analysis again showed patterns where the model did better on metaphors specific to religious communication.

The choice of whether to prefer cross-lingual or cross-genre transfer against a lack of perfectly suitable training data remains thus an issue depending on the specific use case. In cases where the detection of genre-specific metaphors (i.e., family terms related to God) is much more important than the detection of generally frequent, heavily conventionalized metaphors, it may be reasonable to prefer in-genre data over in-language data.

Finally, another issue raised by our results concerns the annotation process. In the example of *da*, the German annotators did not see an underlying metaphorical mapping. This also extends to the use of *there* in the English Reddit corpus, which was annotated by German speakers, as well. One hypothesis is that in both annotation projects, the

annotators were explicitly told to focus on content words and thus dismissed these examples since they resemble temporal prepositions (e.g. *at 4 pm*) expressing a TIME IS SPACE mapping very closely. On the other hand, some metaphorical mappings may be much more easy to grasp for native speakers of one language compared to native speakers of other languages. Thus, focusing on native speakers of only one language altogether would introduce unwanted biases. However, a further exploration of this question goes beyond the scope of this paper.

8. Conclusion

We collected a novel dataset of posts from Christian German online forums and annotated it for metaphor via the MIPVU procedure. We used this dataset to evaluate state-of-the-art methods for automatic metaphor detection (MelBERT and AdMul) in a cross-lingual setting. To do this, we finetuned these models on two English datasets, one smaller in-genre dataset and a larger dataset from different genres. We also used the German dataset as a training set and evaluated on the English dataset from the same genre. We compared these results to a monolingual English, cross-genre transfer setting. We also had a deeper look into our results with a comprehensive error analysis.

From these results, we conclude that cross-lingual transfer between German and English for metaphor detection is possible for metaphors where the underlying cross-domain mapping is represented in both source and target language. Moreover,in line with previous research on metaphor detection, we stress that architectures that employ contextualized embeddings and supervised learning may be preferred over zero- and few-shot approaches with LLMs as long as annotated training data is available. Regarding the specific model choice, AdMul outperformed both MelBERT and the baseline. We observed that the auxiliary task in AdMul may have helped it to learn from finetuning on smaller dataset. This is particularly relevant, given that the genre of the training data plays a vital role, to the extent that in some cases, smaller in-genre training datasets, even if they are in a different source language may be better suited than in-language data from a different domain.

For future work, we suggest, on the one hand, to explore a wider range of languages for cross-lingual transfer, especially from more dissimilar language pairings than English and German. We also suggest to consider the backgrounds of annotators more closely, especially their native language, and explore how these experiences would influence their annotations. Several other NLP tasks already benefited from including multiple perspectives instead of an aggregated gold standard (Cabitza et al., 2023) and our findings suggest that metaphor detection may be a prime example for this.

9. Limitations

One limiting aspect that prevents us from making more general statements on cross-lingual transfer and metaphor detection is the fact that we only considered two relatively related languages. Moreover, although we were aiming for a computationally cheap approach, we are aware that using larger models or additional finetuning of generative LLMs may have further improved the LLM performance here. Finally, all our annotators were native speakers of German. As we discuss in the paper, this may have brought in biases with regards to the perception of metaphors.

10. Ethical Considerations

Regarding the resource-hungry nature of generative LLMs, we limited our usage of such models to a minimum and preferred smaller models. Given that religious beliefs represent particularly sensitive personal information, we made sure that the data we used did not contain information that may link certain opinions to individual people. Our dataset also does not contain usernames. The student assistants did their annotation work within a fixed work contract and were paid according to public payscales. All annotators were informed that their annotations will be used as training data for metaphor detection.

11. Bibliographical References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language 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, Dublin, Ireland. Association for Computational Linguistics.

Mateusz Babieno, Masashi Takeshita, Dusan Radisavljevic, Rafal Rzepka, and Kenji Araki. 2022. MIss RoBERTa WiLDe: Metaphor Identification Using Masked Language Model with Wiktionary Lexical Definitions. *Applied Sciences*, 12(4).

Maria Berger, Nieke Kiwitt, and Sebastian Reimann. 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, Torino, Italia. ELRA and ICCL.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Puli Chen, Cheng Yang, and Qingbao Huang. 2024. Merely judging metaphor is not enough: Research on reasonable metaphor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5850–5860, Miami, Florida, USA. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. 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.

Markus Egg and Valia Kordoni. 2023. A corpus of metaphors as register markers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 220–226, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonathan Gordon, Jerry Hobbs, Jonathan May, Michael Mohler, Fabrizio Morbini, Bryan Rink, Marc Tomlinson, and Suzanne Wertheim. 2015. A corpus of rich metaphor annotation. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 56–66, Denver, Colorado. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

J. Berenike Herrmann, Karola Woll, and Aletta G. Dorst. 2019. Chapter 6. Linguistic metaphor identification in German. In Susan Nacey, Aletta G. Dorst, Tina Krennmayr, and W. Gudrun Reijnierse, editors, *MIPVU around the world*, pages 113–135. John Benjamins Publishing Company.

Rebecca M. M. Hicke and Ross Deans Kristensen-McLachlan. 2024. Science is exploration: Computational frontiers for conceptual metaphor theory. In *Proceedings of the Computational Humanities Research Conference 2024*.

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.

Matej Klemen and Marko Robnik-Šikonja. 2023. Neural metaphor detection for slovene. In *Selected papers from the CLARIN Annual Conference 2022*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Maximilian Köper and Sabine Schulte im Walde. 2016. Distinguishing literal and non-literal usage of German particle verbs. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California. Association for Computational Linguistics.

Volkhard Krech, Tim Karis, and Frederik Elwert. 2023. Metaphors of religion. a conceptual framework. *Metaphor Papers*, 1.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. Univ. of Chicago Press, Chicago [u.a.].

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Yucheng Li, Shun Wang, Chenghua Lin, and Frank Guerin. 2023. Metaphor detection via explicit basic meanings modelling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 91–100, Toronto, Canada. Association for Computational Linguistics.

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.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).

Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13648–13656.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Sebastian Reimann and Tatjana Scheffler. 2024. Metaphors in online religious communication: A detailed dataset and cross-genre metaphor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11236–11246, Torino, Italia. ELRA and ICCL.

Sebastian Reimann and Tatjana Scheffler. 2025a. The struggles of large language models with zero- and few-shot (extended) metaphor detection. *Journal for Language Technology and Computational Linguistics*, 38(2):97–109.

Sebastian Reimann and Tatjana Scheffler. 2025b. 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.

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*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

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.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest bert and crosloengual bert. In *Text, Speech, and Dialogue*, pages 104–111, Cham. Springer International Publishing.

Yorick Wilks. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1):53–74.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

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.

Shenglong Zhang and Ying Liu. 2023. Adversarial multi-task learning for end-to-end metaphor

detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1483–1497, Toronto, Canada. Association for Computational Linguistics.

A. Prompts

Table 4 gives an overview over the prompts used for metaphor detection on the English data and Table 5 shows the prompts used for metaphor detection on the German data, which are the translated equivalent of the English prompt but with a different example for the final prompt (*Weg* (“way") instead of *journey*).

Step	Prompt
MIPVU Step 3a	In one sentence, describe the meaning of the given word in the context of the given post as general as possible. Word: [WORD] Post: [POST]
MIPVU Step 3b	For the given word, determine if it has a more basic contemporary meaning in other contexts than the one in the given post. For our purposes, basic meanings tend to be: - more concrete; what they evoke is easier to imagine, see, here, feel, smell, and taste - related to bodily action - more precise (as opposed to vague) - historically older Basic meanings are not necessarily the most frequent meanings of the word. If such a more basic meaning can be identified, output a brief definition of this basic meaning, otherwise just output 'No.' Word: [WORD] Post: [POST]
MIPVU Step 3c	Can you see a similarity between the senses 1 and 2? 'Similarity' may also mean that the two senses denote distinct concepts that share certain aspects, functions or features. The following example for the word 'journey' illustrates this: journey: Sense 1: "an occasion when you travel from one place to another, especially over a long distance" Sense 2: "a long and often difficult process by which someone or something changes and develops" Answer: Yes. Sense 1 and Sense 2 are similar because in both senses refer to something that takes a longer period of time. Answer with 'yes' or 'no' followed by a brief explanation. Sense 1: [CONTEXTUAL SENSE] Sense 2: [MORE BASIC SENSE]

Table 4: Overview over the English prompts used.

Step	Prompt
MIPVU Step 3a	Beschreibe in einem Satz die Bedeutung des gegebenen Wortes im Kontext des gegebenen Posts. Bleibe dabei so generell wie möglich. Wort: [WORD] Post: [POST]
MIPVU Step 3b	Entscheide für das gegebene Wort, ob es eine grundlegendere, moderne Bedeutung in anderen Kontexten besitzt, als die im gegebenen Post. Für unsere Zwecke ist eine grundlegendere Bedeutung eine Bedeutung, die: - konkreter ist; das heißt, das was sie evoziert kann man sich leichter vorstellen und kann man sehen, hören, riechen, fühlen oder schmecken - die mit körperlichen Handlungen verbunden ist - die präziser ist - die älter ist Eine grundlegendere Bedeutung muss nicht zwangsweise die häufigste Bedeutung des Wortes sein. Falls so eine Bedeutung identifiziert werden kann, dann definiere diese kurz. Andernfalls, falls dem nicht so ist, antworte mit "Nein.". Wort:[WORD] Post:[POST]
MIPVU Step 3c	"Gibt es eine Ähnlichkeit zwischen den beiden Bedeutungen 1 und 2? "Ähnlichkeit" in diesem Zusammenhang, meint auch, dass die beiden Bedeutungen sich auf unterschiedliche Konzepte beziehen, aber bestimmte Aspekte, Funktionen oder Merkmale teilen. Das folgende Beispiel für das Wort "Weg" beschreibt dies genauer: Weg: Bedeutung 1: "Strecke, die zurückzulegen ist, um an ein bestimmtes Ziel zu kommen" Bedeutung 2: "Art und Weise, in der jemand vorgeht, um ein bestimmtes Ziel zu erreichen; Möglichkeit, Methode zur Lösung von etwas" Antwort: Ja, die Bedeutungen 1 und 2 sind sich ähnlich, da Bedeutungen ein Ziel am Ende steht. Antworte mit "ja" oder "nein", gefolgt von einer kurzen Erklärung. Bedeutung 1: [CONTEXTUAL SENSE] Bedeutung 2: [MORE BASIC SENSE]

Table 5: Overview over the German prompts.