Metaphors in Online Religious Communication: a Detailed Dataset and Cross-Genre Metaphor Detection

Sebastian Reimann and Tatjana Scheffler
CRC 1475 Metaphors of Religion
Ruhr University Bochum
{sebastian.reimann, tatjana.scheffler}@rub.de

Abstract

We present the first dataset of fine-grained metaphor annotations for texts from online religious communication, where figurative language plays a particularly important role. In addition to binary labels, metaphors are annotated for deliberateness, that is, whether they are communicated explicitly as metaphors, and we provide indicators for such deliberate use. We further show that cross-genre transfer metaphor detection (from the widely used VUA corpus to our Reddit data) leads to a drop in performance due to the shift in topic and metaphors from source domains that did not occur in the training data. We solve this issue by adding a small amount of in-genre data in fine-tuning, leading to notable performance increases of more than 5 points in F1. Moreover, religious communication has the tendency for extended metaphorical comparisons, which are problematic for current metaphor detection systems. Adding in-genre data had slightly positive effects but we argue that to solve this, architectures that consider larger spans of context are necessary.

Keywords: Corpus, Neural language models, Semantics, Social media processing

1. Introduction

Linguistic metaphors are expressions that relate two entities from different semantic domains by drawing on an implicit similarity between them. For example, Shakespeare's "Juliet is the sun" relates a person to a celestial object by alluding to their importance.

The automatic detection of such linguistic metaphors has received considerable attention within the field of computational linguistics with two shared tasks dedicated to it specifically (Leong et al., 2018, 2020). However, most previous research has focused only on the VUA corpus (Steen et al., 2010) and, consequently, on the genres represented in it (everyday conversation, news text, fiction, academic discourse) and argumentative texts as in Beigman Klebanov et al. (2018) and Mohler et al. (2016).

To the best of our knowledge, religious online forums are a textual domain (in the following, "genre", to avoid confusion with the semantic source and target domains a metaphor links) that has not been explored within the context of automatic metaphor detection. The identification of metaphors in such online forums however would be of interest from the angles of both computational linguistics and digital approaches to religious studies.

On the one hand, online communities in general appear to be a particularly fruitful place for the study of metaphors, as Del Tredici et al. (2019), who explored short-term meaning shift in football-related subreddits, identified metaphorization as one of three main reasons why words change their meaning in online communities. On the other hand, religious language has a particular tendency to use metaphors. Within the study of religion, it is often argued that it is impossible to speak about higher beings without the use of metaphors (Krech et al., 2023), which would consequently lead to a higher frequency of metaphorical language in texts about religion.

This tendency is partially reflected in the findings of Egg and Kordoni (2022) who provide a metaphor corpus consisting of German speeches, sermons, commentaries, light fiction, and debates. Among these genres, sermons contain the most non-conventional or novel and extended metaphors. These subtypes of metaphor are of particular interest as they represent cases of so-called *deliberate metaphor*, which encompasses metaphors that are purposefully communicated as metaphors. Within the context of metaphor and NLP, novel metaphors have in past research been shown to be problematic for metaphor detection systems (Neidlein et al., 2020), and there is even a general lack of research for extended metaphors and their automatic detection (Ge et al., 2023).

In this study we thus ask whether the findings of Egg and Kordoni (2022) for sermons apply to the more spontaneous text type of forum posts and whether forums, too, present such an ideal environment to find metaphors. Further, we apply current state-of-the-art automatic metaphor detection approaches to posts from religious online forums and we specifically discuss the role of deliberate metaphor in this context. Our paper makes the following contributions:

- We provide the first dataset of posts from religious online communities annotated for metaphors and their potential deliberateness.

We will make this dataset publicly available.[1]

- We explore how well current state-of-the-art metaphor detection systems fine-tuned on large-scale metaphor datasets from other genres generalize to data from religious online communities.
- We evaluate how well these models are able to find different kinds of deliberate metaphor in our data.
- We show that metaphor detection can be improved by a small amount of in-genre annotations.

2. Previous Work

2.1. Metaphor Annotation

Most efforts to annotate metaphors in natural text can ultimately be traced back to the Conceptual Metaphor Theory (CMT) by Lakoff and Johnson (1980). CMT defines metaphor as "understanding one concept in terms of another", such as when the concept of ARGUMENT may in parts be understood through the concept of WAR. Arguments may for example be *won*, claims can be *attacked* or may turn out to be *indefensible*. The conceptual metaphor that underlies this mapping would then be ARGUMENT IS WAR. Lakoff and Johnson (1980) further argue that such conceptual metaphors structure a large part of how humans think and how they perceive the world.

The Metaphor Identification Procedure Vrije Universiteit (MIPVU) (Steen et al., 2010) represents a method based on CMT to identify and annotate such aforementioned domain-mappings on the word level. It is concerned with the identification of metaphorically used words, so-called Metaphor-Related Words (MRWs). MRWs may either be used in a *direct* or *indirect* way. Steen et al. (2010) consider an MRW to be direct if the word is used in a literal sense but mapped onto a different domain as it is often the case in metaphorical comparisons. One such case would be *ferret* in (1), where the domain of animals is mapped to the domain of people by comparing an obnoxious person to a ferret. However, despite this domain shift, *ferret* is still used in its literal meaning. Such domain shifts are often signaled explicitly via lexical means such as *like* in this example. Steen et al. (2010) consider such signals to be *Metaphor Flags* (MFlag) and suggest to annotate them, too.

(1) He's like a ferret$_{mrw\_dir}$

For an MRW to be considered indirect, according to Steen et al. (2010), its meaning in context needs to differ from its usual literal meaning. More specifically, to identify indirect metaphors Steen et al. (2010) suggest to:

- read and understand the entire text and identify the contextual meaning of a word
- look up the word in the dictionary
- check if a more basic, i.e. either more concrete, specific or human-oriented, meaning can be found in the dictionary
- if a more basic meaning is listed, decide whether the basic and contextual meaning are sufficiently distinct and related by similarity

If the last point holds true, then the word is considered to be an indirect MRW. *Attacks* in example (2) would represent such a case. Here, the contextual meaning of *attacks* is some sort of criticism to a theory. If we look up *attack* in the Longman Dictionary of Contemporary English (LDOCE), we find the meaning "a strong and direct criticism of someone or something" as well as "the act of using weapons against an enemy in a war", where the second one would be more concrete and thus more basic. The two meanings have received separate numbered senses in the dictionary, which renders them sufficiently distinct, according to Steen et al. (2010). Moreover, a clear similarity can be seen between criticizing something and attacking it in the classical sense, confirming the word *attacks* as an indirect MRW.

(2) The attacks$_{mrw\_indir}$ are based on empirical observation.

2.2. Metaphor Corpora

The first and most widely used resource of token-based annotations à la MIPVU is the VU Amsterdam Metaphor Corpus (VUAMC) provided by Steen et al. (2010). For the Metaphor Detection Shared Tasks in 2018 and 2020 (Leong et al., 2018, 2020), the VUAMC was transformed into a binary dataset (in the following: VUA20). All content words, that is adjectives, adverbs, nouns and verbs, with the label *MRW* were considered metaphoric, all others were marked as literal. 23% of the dataset were held out for testing, the rest was used for training. Besides VUA20, another subtask of the shared task focused on the TOEFL metaphor dataset (Beigman Klebanov et al., 2018), which consists of argumentative essays written by Japanese, Italian and Arabic learners of English. The TOEFL dataset however is much smaller in size than VUA20 and only metaphors relevant to the argumentation were annotated.

[1] https://github.com/SFB-1475/C04-Reddit-Annotations

More recently, MIPVU has been applied to produce datasets in languages other than English. Sanchez-Bayona and Agerri (2022) present CoMeta, a Spanish corpus of MIPVU annotation that comprises a training set of 93,341 and a test set of 23,774 tokens. The data for CoMeta stems on the one hand from newspaper texts of the two largest Spanish UD treebanks, and on the other hand from Spanish parliamentary debates.

Egg and Kordoni (2022) annotated German speeches, sermons, commentaries, light fiction, and debates. In addition to MIPVU and metaphor flag annotation, they additionally provided binary labels on conventionality of metaphors: All MRWs whose contextual sense was present in the dictionary were labeled conventionalized and those without a listed sense were marked non-conventionalized. They also annotated so-called *backgrounds*, which are expressions in the same clause of a metaphoric word that represent its target domain. Moreover, they allowed for chaining of metaphors with the same source-target mapping to enable the identification of extended metaphor.

Mohler et al. (2016) provide a dataset covering four languages (English, Spanish, Russian, Farsi), called LLC. The English data was taken from the ClueWeb09 dataset of language web pages and from the Debate Politics online forum. In contrast to VUAMC, in the LLC dataset, not all tokens were annotated with a binary label. Instead, in each sentence only one word pair (consisting of one word from the source and target domains) was labeled with a numerical rating ranging from 0 ("no metaphor") to 3 ("clear metaphor"). Mohler et al. (2016) state that the criteria on which they base their annotations are comparable to MIP (Pragglejaz Group, 2007), the predecessor of MIPVU, but do not further elaborate on their criteria.

2.3. Metaphor Detection and Cross-Genre Transfer

In recent years, automatic metaphor detection has mostly been carried out using pre-trained language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). The best-performing models often make use of context information (Su et al., 2020), architecture choices that model linguistic theories (Choi et al., 2021), or additional resources like dictionaries (Babieno et al., 2022). They achieve F1-scores between 70–75, when evaluated on training and test data from the same genre.

In order to evaluate multilingual embeddings for automatic metaphor detection, Berger (2022) trained on data from the VUAMC and tested on a German dataset related to American politics. However, the observed performance drops were most likely related to cross-lingual transfer issues rather than genre differences. To our knowledge, the only work so far that explicitly focused on transfer between different textual genres for metaphor detection was carried out by Aghazadeh et al. (2022), who, among probing the layers of different pre-trained language models and testing cross-lingual transfer for metaphor detection, also probed BERT when trained and tested on different datasets, namely the LCC dataset, TroFi (Birke and Sarkar, 2006) (which however does not distinguish between metaphors and other types of figurative language like metonymy) and VUA20. They found that the cross-dataset training outperforms a random baseline, however, especially transfer between the VUA20 dataset and TroFI and LCC fell short compared to in-genre approaches.

In our work, we apply metaphor annotation and detection to a new genre and medium, online religious discussions on Reddit.

3. Annotation

3.1. Data Collection

We searched a corpus of threads from the progressive Christian subreddit r/OpenChristian and conservative Christian subreddit r/TrueChristian for posts that contain family-related terms (e.g., *father*). We then selected the top results and annotated the entire threads in which the posts occurred. This was done on the one hand to obtain metaphors that come from more source domains than just the previously mentioned family domain. On the other hand, annotating entire threads would also make it possible to see how these metaphors and their meaning are discussed, which provides necessary context for the annotation of potentially deliberate metaphors. All data was tokenized using the SoMaJo tokenizer (Proisl and Uhrig, 2016) with the provided English Web Treebank model[2] and POS-tagged using SoMeWeTa (Proisl, 2018) with the English newspaper model provided by the authors[3].

3.2. MIPVU

First, we annotated metaphor-related words. We mostly followed the original guidelines of Steen et al. (2010) outlined in chapter 2.1, including the annotation of MFlags. As the online version of the MacMillan Dictionary (used by Steen et al. (2010) as a primary resource) was taken down in June 2023, we used the LDOCE, which was used by Steen et al. (2010) for a second opinion, as our main lexicographic resource. In line with Leong et al. (2020), we explicitly focused on content words

[2] https://github.com/tsproisl/SoMaJo

[3] https://github.com/tsproisl/SoMeWeTa

only (nouns, verbs, adjectives, adverbs). In total, four people were working on the MIPVU annotation of our data: a PhD student with a background in computational linguistics, a PhD student with a background in religious studies, a student assistant with a background in English linguistics and philosophy and a student assistant with a background in religious studies.

Initially, two threads were annotated by all four annotators in close collaboration and discussion, and any conflicts between the perspectives of computational linguistics and religious studies were resolved. Importantly, we decided to treat words that attribute human characteristics to transcendental entities (e.g., the devil, God) as MRWs since we argue that they also represent a cross-domain mapping via some sort of similarity. This issue is further discussed in the section of religious online communication in Artemov et al. (forthcoming).

In subsequent stages of the annotation process, each thread was annotated by pairs of two annotators, mostly the two student assistants. Disagreements were resolved by discussion. The results of the MIPVU annotation are presented in Table 1.

Subreddit	Tokens	MRWs
r/OpenChristian	8,422	1,557
r/TrueChristian	8,118	1,966
all	16,540	3,523

Table 1: Results of the MIPVU annotation

For the student assistants, we saw a substantial inter-annotator agreement of $\kappa = 0.60$, which is in line with the agreement of $\kappa = 0.63$ reported in (Sanchez-Bayona and Agerri, 2022) for Spanish texts and the $\kappa = 0.70$ for the annotation of transcripts of spoken language, the most similar texts to ours in Steen et al. (2010).

One systematic aspect of religious language that led to some disagreement is that religious words such as *heaven* or *hell* may look tempting at first glance to be annotated as metaphorical but, when considering the dictionary, they are actually used in their most basic sense according to MIPVU.

3.3. DMIP

After applying MIPVU, we used the Deliberate Metaphor Identification Procedure (DMIP) (Reijnierse et al., 2018) to annotate whether an MRW may have been used in a deliberate fashion (i.e., communicated "as metaphor"). The authors of DMIP state that their main intention is to provide a method for the identification of *potentially* deliberate metaphor, since, when annotating metaphor in text, annotators mostly do not have access to the actual intention of the author nor do they know under which circumstances a metaphor is produced. The central question for a metaphor-related word to be considered potentially deliberate according to DMIP is whether "the source domain of the MRW is part of the referential meaning of the utterance". To answer this question, we particularly looked at the subtypes of potentially deliberate metaphors outlined by Reijnierse et al. (2018). These are *direct* metaphor, *novel* or *unconventionalized* metaphor and *extended* metaphor.

We define *direct metaphor* as already illustrated in section 3. In line with Reijnierse et al. (2018) and Egg and Kordoni (2022), we consider a metaphor to be *novel* if its contextual meaning is not represented in the dictionary. This is the case in (3), where *human capital* in its contextual meaning is intended to refer to being passionate about faith. In the LDOCE, a corresponding sense description for this is not available, as the only sense refers to "people and their skills as a factor of production", thus *human capital* is a novel metaphor. The cross-domain mapping from the domain of business to the domain of work is moreover signalled by the use of quotation marks. Due to this missing conventionalized meaning, the recipient needs to have the source domain of the metaphor present and link this to the target, namely being passionate about faith, which would render the metaphor potentially deliberate according to DMIP.

(3) Just keep developing that spiritual "human capital".

A metaphor is *extended* according to Reijnierse et al. (2018), if multiple MRWs within a text span express the same mapping between two semantic domains. Example (4), from a forum post where someone expresses frustration with laws related to charity work, illustrates this. Here the domain of agriculture is mapped onto the domain of charity. However not only the act of charity itself is expressed through the metaphor *plant the seed*, but also possible legal considerations beforehand as *judge the soil*. In order to process this metaphor, the source domain AGRICULTURE needs to be kept in mind and preparations before planting seeds need to be connected to legal preparations before organizing charity, making the metaphor potentially deliberate according to DMIP.

(4) Not my job to judge the soil, just plant the seed

In addition, we considered it a marker of potential deliberateness if a metaphor underwent metaphor shifting processes as outlined in Cameron (2008). Metaphor shifting means that a metaphor from a previous post or external source is either repeated, relexicalized, explicated or challenged in another post. We argue that in these cases, attention needs to be drawn to the source domain. The examples

(5) and (6), from two different posts, illustrate this: (5) introduces the *Father* metaphor, which is further discussed in (6). The explanation why the user likes the *Father* metaphor for God draws attention to the source domain FATHER/PARENT, which makes it fulfill the criteria of DMIP.

(5) How do you feel about calling God “Father”?

(6) I like to use Father because it shows a position like a parent , one who cares for me and loves me as a parent would.

In the DMIP annotation phase, all previously identified MRWs were annotated as *potentially deliberate* or *non-deliberate*. If the former label was chosen, we provided at least one reason from the indicators of potential deliberateness outlined above. The annotators were also able to select “other” and provide an explanation if they found examples which fulfilled the central criteria of Reijnierse et al. (2018) but which did not fit any of the existing descriptions. We also did not consider these labels to be mutually exclusive. For example, an MRW with a meaning not present in the dictionary and which occurs together with other MRWs expressing the same domain-mapping may be both extended and novel. Figure 1 shows how common these types of potentially deliberate metaphor are. We see that extended metaphor is very prominent in our data, which is in line with the findings of Egg and Kordoni (2022) for sermons.

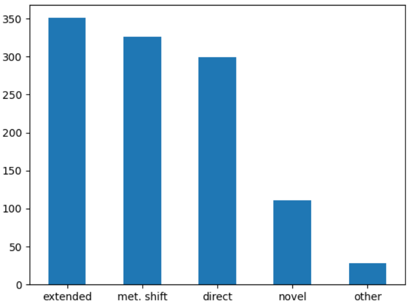

Figure 1: Distribution of markers of potential deliberateness across all potential deliberate MRWs

The DMIP annotation was carried out by two of the annotators, and disagreements were resolved via discussion. The DMIP annotation yielded a substantial agreement of $\kappa = 0.63$ between the two annotators before adjudication, only slightly below the agreement of $\kappa = 0.70$ reported in Reijnierse et al. (2018).

4. Metaphor Detection

4.1. Setup

In order to assess the quality of the transfer from the genres represented in the VUAMC to religious subreddits for the task of metaphor detection, we selected two state-of-the-art models for automatic metaphor detection. The first is DeepMet (Su et al., 2020), the best-performing model in the 2020 metaphor detection shared task. DeepMet uses the transformer architecture by simulating a reading comprehension task with two transformer layers, one to encode the entire sentence together with the word in question and POS features, and one to encode only the local context of a word, plus the word itself and POS features. The outputs are transformed to a feature vector via average pooling and given to a final layer that outputs labels via a softmax function.

Second, we use MelBERT (Choi et al., 2021), which achieved competitive performance with DeepMet. The architecture of MelBERT is inspired by the linguistic theories of MIP (Metaphor Identification Procedure) (Pragglejaz Group, 2007), the predecessor of MIPVU, and the Selectional Preference Violation model SPV (Wilks, 1975), which is concerned with how well a word fits its context. It uses two RoBERTa encoders, one with the entire sentence as input and one for the word in isolation. MIP is simulated by a layer that compares the embedding of a word in context with the embedding of the same word in isolation, whereas the SPV layer compares the contextual embedding of a word with the embedding of the entire sentence in the [CLS] token. We use the same hyperparameters as in the original papers with the exception of setting the maximum sentence length to 256, to account for longer sentences in our data. For better interpretability, we used the versions of DeepMet and MelBERT that did not use ensemble learning or bagging techniques.

Dataset	VUA20		Reddit	
	Train	Test	Train	Test
Tokens	72,611	22,196	1,559	14,981
% MRWs	29%	17,94%	22.64%	21.16%

Table 2: Training/Test splits for our data

We considered it preferable to train on datasets that actually annotated all MRWs in a sentence over just single source-target pairs per sentence like Mohler et al. (2016). We thus fine-tuned the models on three different dataset combinations: the VUA20 training data (Leong et al., 2020), a small in-genre dataset of two annotated threads from the religious subreddits, and a combination of

VUA20 and this small Reddit dataset. All models were tested on the remaining, large portion of the annotated Reddit data described in section 3. The detailed train-test-splits are provided in Table 2.

4.2. Results

	MelBERT			DeepMet		
	P	R	F1	P	R	F1
VUA20	67.75	60.44	63.89	71.53	56.03	62.83
Reddit	28.56	70.47	40.65	64.52	38.96	48.58
both	67.88	68.86	68.37	73.47	61.51	66.96

Table 3: Precision, recall and F1-score for the metaphor class when using different training sets and evaluating on our Reddit data.

Model	P	R	F1
DeepMet	73.9	73.2	73.5
MelBERT	76.4	68.6	72.3

Table 4: Within-corpus precision, recall and F1-score for DeepMet and MelBERT without bagging and ensemble learning reported in Su et al. (2020) and Choi et al. (2021), fine-tuned and tested on the VUA20.

Table 3 shows the results of our experiments when fine-tuning on either VUA20, the in-genre data or a combination of both. Both MelBERT and DeepMet achieve similar results when fine-tuned on the VUA20 training data and tested on our Reddit data. Compared to an evaluation on the VUA20 test data (shown in Table 4), the results are around 10 points lower, suggesting that the model was to some extent able to generalize to the new textual genre but both overgeneralized for some metaphors and failed to recognize others in the Reddit test set. Moreover, for both models, recall seems to be the larger issue compared to precision.

For all models, using the small Reddit training set alone led to an overall worse performance compared to fine-tuning on VUA20, which may be expected, given its 70 times larger size. However, it is noteworthy that the two models behaved differently. For MelBERT, recall increased notably in this scenario while precision dropped sharply, whereas DeepMet showed only slight performance drops across all metrics. This may suggest that DeepMet was better able to generalize from the small dataset but missed some genre-specific metaphors. Adding the small Reddit data to VUA20 for fine-tuning, however, led to an increase in all metrics, especially in recall, for both models.

False Negatives				False Positives			
MelBERT		DeepMet		MelBERT		DeepMet	
father	45	father	42	spirit	45	spirit	24
children	17	children	23	context	19	context	19
mother	15	kingdom	19	lot	18	lot	18

Table 5: Three most frequent false negatives (left) and false positives (right) for the models when fine-tuning only on VUA20, with their frequencies.

4.3. Error Analysis and Discussion

The low recall, especially when fine-tuning without the Reddit data, may hint at systematic issues related to genre differences when identifying metaphors. Cross-genre recall is also notably lower compared to in-genre fine-tuning for both models. To investigate this hypothesis, we look on the one hand at the most frequent false negatives (= unrecognized MRWs) when fine-tuned on VUA20 only, and on the other hand at the most frequent false positives (= non-metaphoric words tagged as MRW). Table 5 shows results which may indeed be indicative of cross-genre generalization problems. The false negatives are dominated by family related terms, with *kingdom* being the only exception. Moreover, the 15 metaphoric instances of *mother* unrecognized by MelBERT were also not recognized by DeepMet. The parent terms are mostly used as a metaphor for God and *children* as a metaphor for believers. The word *father* occurs in the VUA20 training data 27 times, in 26 of these cases it is used literally. The only exception is example (7), which fundamentally differs from the *father* metaphor to describe God, because the family domain is here mapped onto the domain of inventions. The situation for *mother* seems to be similar as it occurs 27 times in the VUA20 training data with only three metaphoric uses.

(7) His mentor, Enrico Fermi, later to be called 'the father of the bomb', and Emilio Segre (who died in California earlier this year).

When fine-tuning MelBERT on the small Reddit dataset exclusively, the previously mentioned family-related metaphors seem to pose less of a problem as only five metaphoric instances of *father* and *children* were not recognized and only six metaphoric instances of *mother* remained unrecognized. The picture is a bit less clear for the combined training data as *father* is still the most frequently missed MRW, however with only half of the instances compared to fine-tuning on VUA20 alone, showing at least some learning effect from the additional in-genre data. Similar observations can be made for *mother* (missed in only eight cases) and *children* (missed seven times).

DeepMet fine-tuned on the Reddit data only struggled slightly more with the family metaphors as it did not recognize *children* in 19 cases, *mother* in nine cases and *father* in eight cases. When fine-tuned on the combination of VUA20 and Reddit, 23 cases of *father*, ten cases of *mother* and eight cases of *children* were not found. However, this still represents an improvement over fine-tuning on the VUA20 training data only.

The false positives in Table 5 also point at struggles with genre-specific vocabulary. The term *spirit*, which was overwhelmingly mistagged as MRW, occured in our Reddit data mostly as a part in the composite noun *Holy Spirit*, where it would not be considered metaphor-related following the guidelines of MIPVU. It did not appear at all in the VUA20 data with this sense and, in general, *spirit* only occurred four times there, all labeled as MRW. This may explain the strong tendency to label *spirit* in *Holy Spirit* as MRW by the model. Here, only the combination of VUA20 and Reddit data in fine-tuning had a slight effect, as only 24 instances of *spirit* were mistakenly considered to be MRW by the models.

The cases of *lot* and *context* are not related to genre differences but still present another case where generalization between datasets was problematic. They represent a conflict between our annotations and the annotation work of Steen et al. (2010). For *context*, the following definitions are provided in the LDOCE:

1. the situation, events, or information that are related to something and that help you to understand it
2. the words that come just before and after a word or sentence and that help you understand its meaning
3. take/quote something out of context

Steen et al. (2010) appears to have labeled *context* in the first sense mostly as metaphoric. However, it is questionable whether any of these meanings can be considered more "basic" (i.e. more concrete, specific or human-oriented (Steen et al., 2010)). Hence, we decided against this condition to be fulfilled for such cases of *context*.

The case of *lot* is even more complex since *lot* in our data as well as in the VUA20 data is tagged as a noun in the phrase *a lot*. In contrast, the dictionaries list *a lot* as an adverb. MIPVU advises to look at the entry for the same part-of-speech as the actual word in question. For their annotation, Steen et al. (2010) indeed must have considered the entry for the noun when annotating the phrase *a lot*, as most instances in the VUA20 training data are tagged as MRW. We, on the other hand, mostly focused on the entry of *a lot* as an adverb and see the noun POS tag for *lot* as an error, and thus did not annotate it as an MRW. Moreover, we also doubt whether the condition "related by similarity" between *a lot* and any of the senses of *lot* as a noun in the dictionary is fulfilled.

Finally, we also noticed inconsistencies for the annotations of *context* and *lot* in the VUA20 dataset. The examples (8) and (9) were not marked as MRW. We argue that in (8), the contextual meaning of *context* would be equal to sense #1 in the LDOCE, which in other instances was labeled as metaphorical. A similar case can be made for (9). Given these inconsistencies, we stick to our reasoning and the decision to not consider these usages of *lot* and *context* to be metaphor-related.

(8) The theory and practice of international agreements, viewed in a game theory context;

(9) you ca n't get a lot with a fiver.

Both *context* and *lot* did not present any problems at all when fine-tuning on our Reddit data only, confirming our suspicion that this issue may be due to inconsistencies in the VUA20 annotation. Here, adding the small Reddit data to the VUA20 corpus had only minor effects.

The previous inconsistencies regarding *lot* and *context* as well as the inter-annotator agreement presented in Section 3.2 show that metaphor annotation is still a relatively challenging task for humans, despite the strict guidelines provided by Steen et al. (2010). In particular, some MRWs are harder identify than others for human annotators. At the suggestion of a reviewer, we investigated whether MRWs which are hard to identify for the automatic systems correspond to those with initial disagreement between human annotators.

Tables 6 and 7 set initial agreement by the annotators in relation to correct classification as MRW by the models with the most beneficial training data combination. It can be seen that a notably higher share of MRWs where annotators initially disagreed was unrecognized, compared to MRWs where the annotators originally agreed. Moreover, 638 MRWs for which initial disagreement was reported were not recognized by either model. This suggests that the more difficult cases for human annotators may also be problematic for computational models.

	Agreement	Disagreement
Correct	1201	982
Incorrect	301	686

Table 6: Comparison of the correct classification as MRW and initial agreement among annotators for MelBERT trained on both VUA20 and Reddit data

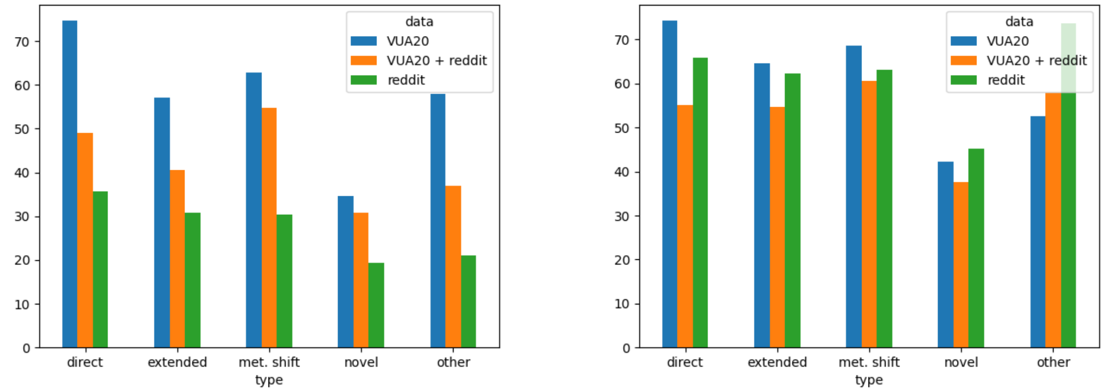

Figure 2: Unrecognized potentially deliberate MRWs (in %) by MelBERT (left) and DeepMet (right)

	Agreement	Disagreement
Correct	1088	862
Incorrect	414	806

Table 7: Comparison of the correct classification as MRW and initial agreement among annotators for DeepMet trained on both VUA20 and Reddit data

4.4. The Role of Deliberate Metaphors

Figure 2 shows that the majority of potentially deliberate metaphors was not marked as metaphoric at all by the models when fine-tuned on VUA20 only. The types of potentially deliberate metaphor that appeared to be more difficult than others were also the most common ones in our Reddit data: More than half of all direct and extended metaphors as well as metaphors that have undergone some sort of metaphor shifting were not recognized as MRWs, with direct metaphors being particularly problematic.

(10) Promiscuity is like going to a birthday party and only eating the sweet part of the cake - the frosting. Then walking out. You 're missing out on getting to know the others at the party which requires patience and time and unknowns which are scary.

The high number of unrecognized MRWs that were subject to metaphor shifting may be partially explained by the fact that the parent metaphors for God were the explicit subject of discussion in one of the threads. Thus, many family metaphors were manually annotated as potentially deliberate MRWs that underwent metaphor shifting. For direct metaphors, we found that among those unrecognized direct MRWs, a considerable number contributed to very long and elaborate metaphoric comparisons, sometimes even spanning over multiple sentences such as the one shown in example 10. Following the conventions of MIPVU, the content words would all receive the label MRW as they form a direct metaphor. However, the vast majority of MRWs here was not classified as such by the models. Words such as *birthday*, *party*, *cake* and *frosting* appear in example 10 in the same context as they would appear in literal contexts outside of metaphorical comparisons. Both models consider the representation of the word and the context in which it occurs. For example 10, it is likely that no contextual clashes can be identified, which would be in line with the findings of Del Tredici et al. (2019), who report similar issues with extended metaphors in the context of semantic change detection via word2vec embeddings.

Moreover, the sentence is the largest unit of context considered by our models. An example like (10) would be split up into three different sentences that are processed individually. While the first sentence would still contain one word from the target domain (*promiscuity*) as well as a lexical marker that introduces a metaphoric comparison (*is like*), the second and third sentence consist entirely of vocabulary from the source domain PARTY without any indications of it being figurative. No model was able to identify all MRWs in example (10) and only MelBERT fine-tuned on our Reddit data managed to identify any MRWs in the last two sentences. For MelBERT fine-tuned on the Reddit data only, we also report by far the lowest share of unrecognized potentially deliberate MRWs as Figure 2 shows. However, given the general tendency of MelBERT fine-tuned on the Reddit training set to overgeneralize, it may be doubted whether this is because it actually learned the properties of such elaborate metaphors. The Reddit data also had a slightly positive effect on MelBERT for the recognition of potentially deliberate metaphors in general when added to the VUA20 data. The impact for DeepMet on the other hand was relatively low.

5. Conclusion and Future Work

We annotated threads from two religious subreddits via both MIPVU, for a binary distinction between metaphoric and literal use, and DMIP, to obtain information on how deliberately these metaphors were used. We used a large section of this annotated data to evaluate the performance of two metaphor detection systems on texts from religious online forums when fine-tuned on data from other genres, a small training set consisting of the rest of our Reddit data or a combination of both. We found that while some aspects of metaphorical language may be transferred regardless of the genre, performance drops when testing and fine-tuning on different genres. However, a small amount (around 1500 tokens) of additional, genre-specific data may already have a beneficial impact. Moreover, metaphors on which annotators already disagreed may have caused problems for the models. Finally, several subtypes of potentially deliberate metaphors were indeed harder to detect than other metaphors.

We hypothesize that the latter issue has direct implications for the worse performance of metaphor detection on texts with religious language. Extended metaphors and very elaborate direct metaphors are phenomena that appear to be characteristic for texts dealing with religion. However, current metaphor detection systems seem not completely fit for larger stretches of metaphoric text from the same domain. Another reason for the worse performance on the forum data was that the models, when fine-tuned on data from different genres, failed to recognize conceptual metaphors that are common in one genre but rare in the other, like the GOD IS PARENT metaphor that dominated our Reddit data. This effect may be due to a lexical bias of metaphor detection systems towards metaphors that have been seen in fine-tuning, and a lack of generalization. Our results show that this issue may be partially counteracted by adding data from the target genre.

Our data echoes some results of Egg and Kordoni (2022) in showing that religious communication is particularly rich in deliberate metaphor. We consequently plan to further annotate data from religious online forums to obtain a suitable large-scale training set for in-genre automatic metaphor detection. More generally, we also call for the inclusion of a broad range of genres when constructing training sets for automatic metaphor detection as some conceptual metaphors may be very genre-specific. We additionally suggest to further investigate the relationship between disagreement in annotation and wrong classification by the models and, finally, in future research on automatic metaphor detection we propose a stronger focus on the problematic cases of extended and direct metaphor by suggesting methods for automatic metaphor detection that take larger contexts into consideration.

6. Acknowledgements

The authors would like to thank Simon Kreutz, Hey Chiu Pang and Lina Rodenhausen for collaborating on the annotations, Frederik Elwert for additional advice and the anonymous reviewers for their valuable comments.

We moreover would like to thank the PC^2 (Paderborn Center for Parallel Computing) for granting us compute time on the Noctua 1 cluster.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), SFB 1475, Project ID 441126958.

7. Ethics Statement

In the annotation process, we manually checked all our Reddit data and made sure that it does not contain any information that would reveal the users' identity or link it to their religious beliefs. Moreover, the dataset does not contain usernames.

The student research assistants conducted all their annotation work within a fixed work contract and were paid according to public pay scales. We also made sure that the annotators were not exposed to any content in the forums that would be potentially offensive or harmful to them.

8. 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.

Nikita Artemov, Elsa Küppers, Sebastian Reimann, Lina Rodenhausen, and Alexandra Wiemann. forthcoming. Taking mipvu further around the world — and through the ages: Mipvu for religion-related texts: Challenges and benefits. *Metaphor Papers*.

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).

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 86–91, New Orleans, Louisiana. 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.

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, Trento, Italy. Association for Computational Linguistics.

Lynne Cameron. 2008. Metaphor shifting in the dynamics of talk. *Confronting metaphor in use*, pages 45–62.

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.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. 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 2069–2075, Minneapolis, Minnesota. 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.

Markus Egg and Valia Kordoni. 2022. Metaphor annotation for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2556–2562, Marseille, France. European Language Resources Association.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: from identification, interpretation, generation to application. *Artificial Intelligence Review*, pages 1–67.

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.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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).

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.

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

Thomas Proisl. 2018. SoMeWeTa: A part-of-speech tagger for German social media and web texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 665–670, Miyazaki. European Language Resources Association ELRA.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2018. DMIP: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147.

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. OCLC: ocn557407891 tex.ids: steen_method_2010.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *The Second Workshop on Figurative Language Processing*, Seattle, WA.

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