Harder than Finding the Lost Sheep? Towards Automatically Suggesting Deliberate Metaphor Annotations in German Sermons

Ronja Laarmann-Quante and Stefanie Dipper
Ruhr University Bochum, Germany
Faculty of Philology
Linguistics Department
{ronja.laarmann-quante|stefanie.dipper}@rub.de

Abstract

Automatic metaphor detection so far has largely focused on English data annotated for all kinds of metaphors including ubiquitous conventionalized ones. In this paper, we focus on deliberate metaphors in German sermons, i.e., metaphors that are used with a specific communicative goal. This task is harder because there is less training data available, and deliberate metaphors are very rare. Our goal is to support human annotators with automatically generated suggestions, so we strive above all for high recall. Using multilingual transfer learning based on various metaphor datasets and different transformer models, the highest recall we achieve is .70 (precision .10). Our results suggest that larger context windows beyond the sentence level are not helpful and that adding in-domain data even when annotated with different guidelines and in a different language is beneficial.

1 Introduction

According to Lakoff and Johnson (1980), metaphors are a cognitive mechanism that we use to think and understand the world by making complex, unfamiliar concepts understandable through familiar, concrete concepts. Visible signs of this fundamental mechanism are the numerous conventionalized metaphors in language, where a word is used in a context other than its usual one, thereby transferring its meaning. For example, the word *kill* is used in the phrase *kill the process* to mean *end the process*. The semantic domain of the usual context of the word *kill* (death) is called the source domain, while the domain to which the transfer takes place is called the target domain (computing) (Lakoff and Johnson, 1980).

A fundamental distinction in metaphors concerns their communicative purpose (Steen, 2023). The majority of metaphors that occur in everyday language are conventionalized, i.e., the transferred meaning is usually listed in the dictionary, alongside the usual meaning. The use of such metaphors is usually completely unobtrusive and does not involve any particular communicative goals. However, when a speaker uses a metaphor in an utterance to make something understandable with the help of some other concept, this is referred to as a **deliberate metaphor**. A characteristic of such metaphors is that the source domain plays a role in the statement, thereby creating a different perspective. Steen (2011) provides the example of football coach Rinus Michels, who said *football is war*, thereby inviting his audience to view a football match from the perspective of a war event.[1]

The work reported here was carried out in the context of the Collaborative Research Center (CRC) "**Metaphors of Religion**".[2] The CRC's starting assumption is that religion is an area in which metaphors necessarily play a central role, since the ultimate subject of religion—the transcendent—is not directly accessible and cannot be referred to literally, but can only be talked about through metaphors (Krech et al., 2024). Religious meaning-making therefore occurs primarily through metaphors, as evidenced by central Christian terms and concepts such as salvation or the Last Judgment.

All members of the CRC examine and analyze metaphors by annotating metaphorically used expressions in texts. In other words, we are not so much interested in conceptual, abstract metaphors in the sense of Lakoff and Johnson (1980) (such as LIFE IS A JOURNEY) than in linguistic instances of metaphors, which do not necessarily conform to a common conceptual metaphor. This paper presents

[1] Note that conventionalized metaphors can therefore also be deliberate: when the source domain plays a communicative role; for an example, see Sec. 2. Novel metaphors, i.e. newly created, non-conventional metaphors, are a subset of deliberate metaphors.

[2] https://sfb1475.ruhr-uni-bochum.de/

Proceedings of SIGHUM 2026, pages 164–175

a study on the automatic detection of metaphorically used expressions. Our goal is to support the annotation work by developing a system that pre-annotates candidates for metaphors, which are then reviewed by human annotators.

Religious discourse is not only a domain rich in deliberate metaphors (see also Egg and Kordoni, 2022) but metaphors also often stretch over larger parts of a text, called **extended metaphors**, potentially posing an additional challenge for the automatic detection, which has typically been carried out sentence-wise (Reimann and Scheffler, 2024).

Related work There are only few corpora annotated for metaphors, especially for languages other than English (see e.g. Lu and Wang, 2017 for Chinese; Sanchez-Bayona and Agerri, 2022 for Spanish; Dipper et al., 2024, Dipper, 2025 for German). The standard metaphor corpus is the English VU Amsterdam Metaphor Corpus (VUAMC) by Steen et al. (2010) with texts from four registers: academic texts, news texts, fiction, and conversation. The texts have been annotated word by word according to the MIPVU guidelines (Steen et al., 2010), which mark all "metaphor-related words" (MRWs) and do not distinguish between conventionalized and deliberate metaphors. Most work on automatic metaphor detection has focused on this corpus and F1 scores on the metaphor class of $> .75$ have been achieved on the VUAMC (see e.g. the shared tasks in Leong et al., 2018, 2020).

To date, there are only a few studies that explicitly target the detection of deliberate metaphors. Neidlein et al. (2020) showed that performance drops considerably when detecting non-conventionalized metaphors compared to conventionalized ones in the VUAMC. Reimann and Scheffler (2024) analyzed a corpus of online religious communication (see also Sec. 2) and found that several subtypes of deliberate metaphors were harder to detect than other metaphors.

Contribution and research questions Our paper tackles two research gaps, namely (1) optimizing the detection of deliberate metaphors and (2) focusing on a non-English dataset, i.e., a German corpus of sermons. Our specific research questions are as follows:

RQ1: Given the sparsity of annotated training data, how beneficial is it to add further data that does not precisely meet our objectives because it originates from other domains or languages or has been annotated according to different guidelines.

RQ2: Is a larger context size beyond the sentence level beneficial for identifying metaphors, given the observation of Reimann and Scheffler (2024) and Egg and Kordoni (2022) that religious discourse contains many extended metaphors?

We follow the approach of Berger et al. (2024) using multilingual transfer learning. We compare different multilingual transformer models and different transformer architectures including sentence transformers and a longformer model that captures more context. Our experiments focus on optimizing the recall of the metaphor class. The idea is that a high-recall system can be used in an assisted annotation scenario, as sketched above. The system would pre-annotate the data which is then checked by human annotators. Their primary task would be reduced to dismissing false positives, thereby speeding up the overall annotation process for generating new training data.

Our code and data is available under `https://gitlab.ruhr-uni-bochum.de/vamos-cl/latech-clfl_2026_metaphors`.

2 Data

We use four publicly available datasets described in the following. Our goal is to predict deliberate metaphors in the Sermon dataset while the other datasets serve as additional training data for fine-tuning the transformer models. In all datasets, metaphors are annotated on a token basis, i.e., every word used metaphorically is marked as such. None of the datasets provides (complete) information about which metaphorically used words belong to the same metaphorical image, so this information is not available to our system. The number of documents, sentences, tokens and annotated metaphors (i.e. metaphorically used words) per dataset is given in Table 1.

Sermon (de) (Dipper, 2025) is our primary dataset, which is annotated for deliberate metaphors according to the guidelines by Dipper et al. (2025). The guidelines provide for various labels for metaphor-related words. For example, expressions that are central to the metaphorical image are annotated with the label "center", while less central expressions are given the underspecified label "MRW" (for "metaphor-related word"). Conventionalized metaphors that are deliberately used in the present context are labeled "revitalized". The label "anchor" is used for literal expressions that are the target of the metaphorical transfer.

	Relieved	*from*	*the*	***shackles***	*of*	*guilt*	*and*	*debt,*	***breathing***	***freely***,	*not*	*being*	***caught***	*in*	*an*	*overwhelming*	*past.*
Original	m	–	–	c	–	a	–	a	m	c	–	–	rvt	–	–	–	a
Binary	m	l	l	m	l	l	l	l	m	m	l	l	m	l	l	l	l

Figure 1: Translated and adapted example from the Sermon set with original and binary labels. Original labels are: c: center; rvt: revitalized; m: MRW; a: (literal) anchor (see Dipper et al. (2025) for full documentation of these labels).

The following examples are taken from the sermon dataset but translated to English for demonstration purposes. Fig. 1 shows an annotated example; metaphorically used words are printed in bold. The phrase *caught in the past* is rather conventional (in the German original), but in combination with *relieved from the shackles*, the original image of being caught becomes alive again and "revitalized".

Ex. (1) shows a construction that is typical for sermons: different but often related images are used for the same subject matter (here: for the journey of life). Ex. (2) shows a metaphor that is typographically highlighted by quotes.

(1) Perhaps it is not out of the question that in the coming weeks we will remember here and there the basic **lines** of our lives, our **pilgrimage** between birth and death, our being on the **road** in search of ourselves and of God.

(2) Have we overcome all problems and are we already, so to speak, "at our **destination**"?

For our system, we map these labels to binary labels *m* (metaphorical) and *l* (literal), as shown in Fig. 1. We randomly split the documents in two sets, Sermon A and B, to perform two-fold cross-validation.

TEDx (de) (Dipper et al., 2024) contains German TEDx talks, i.e. non-religious texts, annotated for deliberate metaphors according to the same guidelines as the sermons. We again map the annotations to binary labels.

Reddit (en) (Reimann and Scheffler, 2024) contains threads from Christian subreddits in English. It was first annotated following the MIPVU guidelines (see Sec. 1) and secondly, deliberate metaphors were identified using the DMIP procedure (Reijnierse et al., 2018). We extracted the latter annotations (binary labels).

VUA (en) (Reijnierse et al., 2019) is a version of the VUAMC (see Sec. 1) which has been enriched with deliberate metaphor annotations according to DMIP. We use these annotations (binary labels).

Dataset	#docs	#sents	#toks	#met	%met
Sermon A	8	889	16,720	270	1.61
Sermon B	7	826	13,126	319	2.43
TEDx	10	1,260	19,794	259	1.31
Reddit	301	1,815	37,171	979	2.63
VUA	118	15,440	189,981	1,109	0.58

Table 1: Overview of deliberate metaphor datasets. Deliberate metaphors are a rare phenomenon, accounting for 0.58–2.63% of all tokens in the different datasets.

3 Method

We implement the metaphor detection as a binary sequence labeling task, where each token is classified as metaphorical or not. We compare several experimental settings: (1) **different pre-trained multilingual transformer models**:[3] (i) mBERT (Devlin et al., 2019), (ii) XLM-RoBERTa (Conneau et al., 2020), (iii) two sentence transformers, namely Cross English & German RoBERTa for Sentence Embeddings by Philip May and (iv) multilingual SBERT (Reimers and Gurevych, 2019, 2020), and (v) the longformer model XLM-Long for sequence lengths up to 4096 tokens; (2) **different training data**: all possible combinations of the datasets in Sec. 2 for finetuning the transformer models; (3) **different context sizes** fed into the transformer: sentences, windows of 50 and 100 tokens,[4] and whole documents.

Our code is based on the one by Wachowiak et al. (2022). All experiments were run on a Linux workstation with a single Nvidia RTX 4000 Ada GPU with 20 GB of memory. No model needed more than one hour to be finetuned. We used the default AdamW optimizer with no weight decay, a training batch size of 8, a learning rate of 2e-5 and finetuned each model for 8 epochs. The best model

[3] All models are taken from Huggingface, the exact links are given in Table 8 in Appendix E.

[4] Windows do not overlap but it is made sure that sentences are not cropped, i.e., the number of 50 or 100 tokens per window is typically exceeded to accommodate the rest of the last sentence.

		all	with met		%met
sentences	Sermon A	889	150	(17%)	7.04
	Sermon B	826	180	(22%)	9.77
	TEDx	1,260	118	(9%)	12.21
	Reddit	1,815	457	(25%)	8.38
	VUA	15,440	610	(4%)	7.53
windows 50	Sermon A	266	102	(38%)	4.04
	Sermon B	214	112	(52%)	4.63
	TEDx	332	77	(23%)	5.66
	Reddit	704	321	(46%)	5.30
	VUA	3,157	506	(16%)	3.45
windows 100	Sermon A	151	76	(50%)	3.05
	Sermon B	119	83	(70%)	3.47
	TEDx	185	61	(33%)	3.95
	Reddit	479	261	(54%)	4.33
	VUA	1,765	435	(25%)	2.30
documents	Sermon A	8	8	(100%)	1.61
	Sermon B	7	7	(100%)	2.43
	TEDx	10	10	(100%)	1.31
	Reddit	301	185	(61%)	3.54
	VUA	118	93	(79%)	0.67

Table 2: Number and percentage of units (sentences, token windows, documents) with metaphors (*with met*) out of all units (*all*). *%met* shows the percentage of metaphorical tokens in the units.

found during finetuning (based on a validation set of 10% of training data) was saved. Since we aim for a high-recall system to assist manual annotation (see Sec. 1), we set the metric to be improved to recall of the metaphor class. All models, except XLM-Long, have a maximum input length of 512 tokens. In order not to lose information when this limit is exceeded, we use a sliding window approach with a 32-token overlap between windows. At inference time, only the first prediction for each token occurring in an overlapping sequence is kept.

One problem in the detection of deliberate metaphors is the high class imbalance (see Table 1). To mitigate this, another experimental setting of ours is reducing each dataset to the units that contain metaphorical tokens (henceforth called **condensed** datasets). Table 2 shows the number of units (sentences, token windows, or documents) with metaphors for each dataset.[5]

4 Results

We evaluate the predictions with precision, recall and F1 score on the metaphor class using scikit-learn (Pedregosa et al., 2011). In general, the models achieve considerably higher precision with the original datasets than with the condensed datasets, which in turn yield clearly higher recall scores. This is not unexpected, as the condensed datasets simulate a higher proportion of metaphors. Since our goal is to achieve the highest possible recall scores, we focus on the condensed setting in our discussion.

The main results based on the condensed dataset are shown in Table 3.[6] For completeness, the full results on the original dataset are given in Appendix B.

Since our primary dataset of sermons is small, we randomly divided it into two parts, Sermons A and B, to make the results as generalizable as possible. We always report the mean of the evaluations on Sermons A and B. For settings that use the sermons for fine-tuning, we perform a two-fold cross-validation.[7]

In general, the highest recall scores (marked in red) and F1 scores (blue) tend to be achieved with larger data sets (bottom of the table), while high precision (green) is also achieved with smaller training data sets located in the upper part of the table. However, the most significant differences arise on the basis of the various context windows. Single sentences as context windows clearly yield the best results with regard to recall. The larger the context window, the worse gets the recall. Precision, however, benefits from larger contexts, but the differences are much smaller than with recall. This pattern contradicts the hypothesis that religious discourse requires larger context sizes because it contains many extended metaphors (**RQ2**).

One reason might be that in the condensed datasets, the density of metaphors is highest on the sentence level (see Table 2). For the uncondensed datasets, the pattern of decreasing performance with increasing context size is not as clear.

When comparing the models with sentential context windows with regard to recall, the two sentence

[5]Since at inference time it is unknown which units will contain metaphors, the test set is not condensed but always used as is.

[6]mBERT shows somewhat similar patterns to XLM-RoBERTa. While it has a slightly higher peak performance in recall (.61) compared to XLM-RoBERTa (.59), its results are overall slightly worse, so we report the full results only in Table 4 in Appendix A.

[7]The individual results of Sermon A and B for the condensed datasets are shown in Appendix C and D. In general, performance is better on test set B. The higher performance may depend on the higher proportion of metaphors in this set, which may benefit models that have been trained on a condensed data set and/or are designed to achieve the highest possible recall.

	XLM-RoBERTa												ST en-de			ST multiling			XLM-Long		
	sentence			w50			w100			document			sentence			sentence			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
s(ermon)	.22	.17	.33	.02	.14	.01	.00	.00	.00	.00	.00	.00	.22	**.16**	.42	.20	**.15**	.35	.00	.00	.00
t(edx)	.16	.12	.24	.02	**.28**	.01	.00	.00	.00	.00	.00	.00	.16	.10	.49	.15	.10	.32	.00	.00	.00
r(eddit)	.21	.14	.44	.22	.18	.29	.20	.18	.24	.18	.22	**.16**	.18	.10	**.64**	.18	.11	.61	.17	.18	.17
v(ua)	.20	.16	.27	.13	.18	.10	.08	.19	.05	.01	**.38**	.01	.17	.11	.42	.15	.09	.51	.00	.00	.00
r+s	.24	.16	.50	.25	.24	.26	.23	.26	.21	.12	.29	.09	.21	.13	.56	.22	.14	.56	.19	.33	.14
t+s	.23	.16	.43	.23	.24	.23	.14	**.44**	.12	.00	.00	.00	.22	.14	.54	.22	.14	.54	.00	.00	.00
v+s	.25	.17	.46	.27	**.28**	.25	.23	.27	.21	.11	.28	.07	.21	.13	.59	**.23**	.14	.56	.11	.28	.06
r+t	.23	.15	.52	.22	.21	.24	.21	.23	.19	.18	.33	.13	.20	.13	.53	.20	.12	.60	.15	.26	.11
v+t	.23	.15	.44	.26	.21	.33	.22	.26	.19	.06	.23	.03	.18	.10	.55	.20	.12	.54	.13	.28	.08
v+r	.23	.16	.43	.27	.23	.35	.25	.28	.22	.16	.20	.14	.19	.11	**.64**	.17	.10	**.70**	.13	.20	.10
r+t+s	.25	.16	**.59**	.28	.21	**.39**	.25	.27	.24	.16	.27	.11	**.23**	.14	.61	.22	.14	.56	.17	.28	.13
v+t+s	**.26**	**.19**	.45	.28	.25	.32	.28	.29	.28	.09	.31	.05	.22	.14	.56	.22	.14	.56	.14	**.34**	.09
v+r+s	.25	.17	.51	**.30**	**.28**	.33	**.31**	.30	**.32**	**.20**	.29	.15	.21	.13	.61	.22	.14	.55	**.23**	.28	**.20**
v+r+t	.24	.15	.55	.29	.23	.38	.28	.27	.30	.11	.24	.07	.19	.11	.60	.20	.12	.63	.14	.25	.10
v+r+t+s	.25	.16	.55	.29	.27	.31	.30	.32	.28	.17	.29	.13	.21	.13	.60	**.23**	.14	.57	.22	.32	.17

Table 3: Results based on the condensed datasets for models XLM-RoBERTa (model (ii) from Sec. 3), Cross English & German RoBERTa (ST en-de, (iii)), multilingual SBERT (ST multiling, (iv)), and XLM-Long (v); ST = sentence transformer. The highest values for F1 score (f1), precision (p), recall (r) across training datasets per setting are boldfaced and colored. Columns *sentence*, *w50*, *w100* (windows of 50/100 tokens) and *document* refer to the context window fed into the transformer.

transformers perform better than XLM-RoBERTa and also appear to require less training data (one or two datasets vs. three datasets). The best recall scores are achieved with datasets that include Reddit. The top score of .70 recall (with a precision of .10 and an F1 score of .17) is achieved by multilingual SBERT on a training set combining Reddit and VUA.

We can tentatively conclude that adding in-domain training data (Reddit), even when annotated with different guidelines, is beneficial (**RQ1**), which would be in line with the observation of Reimann and Scheffler (2024). The language difference does not seem to play a role for the multilingual models; perhaps the larger size of the English-language Reddit data compared to the German-language sermons is the decisive factor.

An error analysis reveals some typical sources of false positives: Words marked as literal quotations by quotation marks are often classified as metaphors. Quotation marks are in fact often an indication of a metaphor (see Dipper et al., 2025), as in Ex. (2). In addition, coordination and enumerations appear to be structures that can trigger false positives, compare the enumeration-like structures with multiple metaphorical words in Ex. (1) and Fig 1. Finally, according to the guidelines in Dipper et al. (2025), passages that a sermon cites from the Bible should not be annotated at all, even if they contain metaphors (e.g. *vine* in Jesus' famous saying *I am the true vine*). Since the models were not trained to detect Bible quotations, they often classify words within such quotations as metaphors, which does not match the gold standard here but would be correct in other, non-biblical contexts.

5 Conclusion & Future Work

We presented experiments on the automatic detection of deliberate metaphors in German sermons using multilingual transfer learning. The model achieving the highest recall (.70) was multilingual SBERT finetuned on an English dataset of online religious communication, highlighting the benefits of adding in-domain data. Sentence-based classification performed best, contradicting our hypothesis that larger contexts are beneficial for detecting metaphors. This is potentially due to the fact that the density of metaphors was highest at this level. Our next steps will be to use the model to assist human annotators in creating more in-domain training data. Currently, a limitation is that the precision scores are low, requiring the annotators to dismiss many false positives. With more training data available, our goal is to build more balanced models.

Acknowledgments

We are very grateful to the anonymous reviewers for their helpful comments. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – 441126958.

References

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefanie Dipper. 2025. Metaphorical heads and literal dependents: Syntactic properties of metaphors in German. In *Proceedings of the 23th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, page 81–90, Ljubljana, Slovenia.

Stefanie Dipper, Adam Roussel, Alexandra Wiemann, Won Kim, and Tra-my Nguyen. 2024. Guidelines for the annotation of deliberate linguistic metaphor. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 53–58, NAACL, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Stefanie Dipper, Alexandra Wiemann, and Adam Roussel. 2025. Guidelines zur Annotation von deliberaten Metaphern. *Metaphor Papers*, 27.

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.

Volkhard Krech, Tim Karis, and Frederik Elwert. 2024. Metaphors of religion: A conceptual framework. *Metaphor Papers*, 1.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press, Chicago, London.

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.

Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation*, 51(3):663–694.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2019. Metaphor in communication: the distribution of potentially deliberate metaphor across register and word class. *Corpora*, 14(3):301–326.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 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. 2011. From three dimensions to five steps: the value of deliberate metaphor. *metaphorik.de*, 21.

Gerard J. Steen. 2023. Thinking by metaphor, fast and slow: Deliberate Metaphor Theory offers a new model for metaphor and its comprehension. *Frontiers in Psychology*, 14.

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

Lennart Wachowiak, Dagmar Gromann, and Chao Xu. 2022. Drum up SUPPORT: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 44–53, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Results of mBERT on Condensed Data

	mBERT											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.22	.14	.49	.21	**.22**	.22	.17	.25	.15	.00	.00	.00
tedx	.18	.12	.44	.20	.17	.26	.13	.15	.12	.00	.00	.00
reddit	.18	.11	.48	.18	.14	.23	.18	.13	.27	.16	.15	.16
vua	.21	**.15**	.36	.17	.21	.15	.12	.21	.08	.02	.03	.02
reddit+sermon	.22	.14	.54	.23	.20	.29	.24	.22	.28	.19	.22	.17
tedx+sermon	.22	.14	.53	.24	**.22**	.31	.24	.24	.25	.11	.24	.08
vua+sermon	.22	.14	.48	**.26**	**.22**	.33	.23	.24	.22	.11	.27	.08
reddit+tedx	.20	.12	.56	.22	.16	.33	.21	.20	.22	.09	.21	.05
vua+tedx	.18	.10	**.61**	.24	.20	.31	.24	.22	.28	.09	.25	.06
vua+reddit	.21	.13	.58	.21	.17	.30	.18	.19	.18	.18	.17	**.20**
reddit+tedx+sermon	.21	.13	.59	.25	.21	.33	.22	.20	.25	.18	.26	.14
vua+tedx+sermon	.22	.14	.54	.26	.21	.35	.24	**.26**	.23	.18	**.29**	.13
vua+reddit+sermon	**.23**	.14	.54	**.26**	.21	**.36**	**.25**	.25	.25	**.22**	.27	.19
vua+reddit+tedx	.20	.12	.56	.23	.18	.33	.24	.20	**.29**	.13	.22	.09
vua+reddit+tedx+sermon	.22	.14	.55	.25	**.22**	.29	**.25**	.24	.27	.15	.21	.13

Table 4: Results of mBERT (mean of Sermon set A and B) based on the condensed datasets.

B Results on Original (Non-Condensed) Data

	XLM-RoBERTa											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.22	.31	.17	.05	.46	.03	.00	.00	.00	.00	.00	.00
tedx	.11	.30	.07	.05	.30	.03	.00	.00	.00	.00	.00	.00
reddit	.21	.23	.20	.20	.25	.16	.21	.26	.17	.15	.21	.12
vua	.08	.38	.04	.10	.39	.05	.05	.22	.03	.00	.00	.00
reddit+sermon	.16	.27	.12	.24	.35	.19	.24	.33	.19	.23	.33	.18
tedx+sermon	.17	.35	.11	.14	.28	.10	.17	.32	.11	.00	.00	.00
vua+sermon	.18	.30	.13	.16	.33	.10	.22	.36	.16	.09	.30	.05
reddit+tedx	.21	.29	.16	.15	.33	.10	.20	.39	.14	.10	.31	.06
vua+tedx	.18	.27	.13	.12	.30	.07	.11	.35	.06	.10	.30	.06
vua+reddit	.19	.25	.15	.19	.24	.15	.24	.31	.19	.22	.23	.20
reddit+tedx+sermon	.19	.35	.13	.25	.35	.19	.22	.33	.17	.20	.28	.15
vua+tedx+sermon	.17	.33	.12	.17	.37	.11	.25	.38	.19	.15	.31	.10
vua+reddit+sermon	.15	.31	.10	.23	.34	.18	.22	.31	.16	.19	.28	.15
vua+reddit+tedx	.10	.23	.07	.19	.33	.13	.21	.34	.15	.20	.28	.15
vua+reddit+tedx+sermon	.22	.32	.17	.25	.36	.19	.23	.35	.17	.21	.31	.16

	mBERT											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.17	.31	.14	.18	.30	.14	.12	.24	.08	.00	.00	.00
tedx	.05	.18	.03	.06	.27	.03	.05	.19	.03	.00	.00	.00
reddit	.11	.14	.09	.14	.15	.13	.17	.18	.16	.07	.13	.05
vua	.05	.20	.03	.09	.20	.06	.09	.27	.05	.04	.20	.02
reddit+sermon	.18	.28	.14	.27	.28	.25	.25	.26	.24	.16	.22	.14
tedx+sermon	.20	.29	.15	.20	.33	.15	.20	.29	.16	.11	.24	.08
vua+sermon	.17	.31	.12	.20	.29	.17	.20	.31	.15	.20	.27	.16
reddit+tedx	.14	.18	.11	.13	.26	.08	.12	.26	.07	.16	.23	.13
vua+tedx	.17	.21	.15	.19	.24	.15	.13	.31	.08	.09	.28	.05
vua+reddit	.09	.19	.06	.23	.20	.26	.19	.21	.18	.09	.20	.06
reddit+tedx+sermon	.19	.27	.15	.24	.28	.21	.22	.29	.19	.17	.21	.15
vua+tedx+sermon	.14	.27	.10	.22	.31	.17	.21	.31	.16	.15	.26	.10
vua+reddit+sermon	.19	.35	.14	.24	.31	.20	.25	.31	.22	.22	.27	.18
vua+reddit+tedx	.09	.22	.06	.17	.26	.12	.16	.26	.12	.17	.23	.13
vua+reddit+tedx+sermon	.19	.29	.15	.25	.29	.22	.22	.33	.17	.16	.23	.15

	ST de-en			ST multilingual			XLM-Long		
	sentence			sentence			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.17	.34	.14	.21	.22	.20	.00	.00	.00
tedx	.07	.32	.04	.11	.19	.08	.00	.00	.00
reddit	.22	.17	.32	.18	.13	.33	.17	.21	.15
vua	.07	.37	.04	.11	.18	.08	.00	.13	.00
reddit+sermon	.22	.30	.18	.23	.28	.20	.23	.31	.19
tedx+sermon	.21	.34	.16	.22	.23	.21	.00	.00	.00
vua+sermon	.18	.31	.12	.22	.27	.19	.10	.31	.06
reddit+tedx	.25	.24	.26	.18	.23	.15	.20	.25	.17
vua+tedx	.19	.27	.15	.19	.20	.18	.12	.27	.07
vua+reddit	.14	.24	.10	.19	.15	.28	.23	.28	.19
reddit+tedx+sermon	.21	.30	.17	.23	.25	.21	.23	.27	.20
vua+tedx+sermon	.20	.34	.14	.25	.29	.22	.16	.27	.11
vua+reddit+sermon	.23	.35	.20	.20	.27	.16	.24	.31	.19
vua+reddit+tedx	.19	.24	.16	.18	.25	.15	.21	.31	.16
vua+reddit+tedx+sermon	.23	.33	.18	.24	.29	.21	.24	.34	.19

Table 5: Results based on the original, not condensed, dataset (mean of Sermon sets A and B).

C Results on Sermon Set A (Condensed)

	XLM-RoBERTa											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.21	.15	.34	.04	.27	.02	.00	.00	.00	.00	.00	.00
tedx	.14	.10	.23	.02	.33	.01	.00	.00	.00	.00	.00	.00
reddit	.17	.11	.39	.18	.13	.27	.17	.14	.22	.16	.18	.14
vua	.17	.14	.23	.13	.17	.10	.07	.16	.05	.01	.25	.00
reddit+sermon	.20	.12	.49	.21	.20	.23	.22	.23	.22	.19	.28	.15
tedx+sermon	.18	.12	.36	.25	.23	.28	.25	.29	.22	.00	.00	.00
vua+sermon	.20	.13	.47	.25	.25	.24	.25	.26	.25	.10	.24	.06
reddit+tedx	.18	.12	.45	.21	.18	.25	.20	.20	.19	**.20**	**.35**	.14
vua+tedx	.21	.14	.42	.28	.22	.38	.23	.25	.21	.03	.17	.02
vua+reddit	.21	.14	.40	.24	.20	.31	.22	.25	.20	.15	.17	.14
reddit+tedx+sermon	.19	.11	**.55**	.25	.19	.36	.25	.26	.24	.15	.26	.10
vua+tedx+sermon	**.22**	**.15**	.39	.28	.23	.36	**.32**	**.31**	**.34**	.12	.33	.07
vua+reddit+sermon	.20	.14	.41	.27	.24	.33	.28	.27	.28	.19	.23	**.16**
vua+reddit+tedx	.20	.13	.49	**.29**	.23	**.40**	.29	.27	.31	.11	.24	.07
vua+reddit+tedx+sermon	.21	.14	.46	.26	.25	.26	.28	.28	.27	.18	.28	.14

	mBERT											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.17	.10	.44	.20	.18	.23	.24	**.25**	.23	.00	.00	.00
tedx	.14	.09	.40	.20	.15	.27	.17	.19	.16	.00	.00	.00
reddit	.14	.08	.43	.15	.11	.22	.15	.11	.27	.14	.12	.16
vua	**.19**	**.13**	.33	.16	.18	.14	.13	.24	.09	.03	.03	.03
reddit+sermon	.16	.09	.48	.19	.14	.29	.21	.16	**.31**	.15	.18	.13
tedx+sermon	.17	.10	.51	.23	.17	.35	**.25**	.21	**.31**	.16	.25	.12
vua+sermon	.18	.12	.39	**.25**	**.20**	.33	**.25**	.23	.27	.05	.23	.03
reddit+tedx	.16	.09	.49	.19	.13	.31	.21	.20	.23	.11	**.26**	.07
vua+tedx	.14	.08	**.56**	.24	.19	.33	.24	.20	.29	.10	.25	.06
vua+reddit	.17	.10	.53	.18	.13	.27	.16	.16	.16	.18	.16	**.21**
reddit+tedx+sermon	.15	.09	.55	.23	.17	.36	.19	.16	.25	.18	.23	.15
vua+tedx+sermon	.18	.11	.53	.24	.18	.36	**.25**	.24	.26	.18	.25	.14
vua+reddit+sermon	.17	.10	.46	.23	.17	**.37**	.22	.21	.24	**.22**	.23	**.21**
vua+reddit+tedx	.17	.10	.53	.22	.16	.33	.23	.19	**.31**	.15	.25	.11
vua+reddit+tedx+sermon	.17	.10	.51	.23	**.20**	.27	.23	.22	.23	.20	.20	.20

	ST de-en			ST multilingual			XLM-Long		
	sentence			sentence			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	**.18**	**.11**	.44	.16	.10	.38	.00	.00	.00
tedx	.12	.07	.43	.12	.08	.29	.00	.00	.00
reddit	.13	.08	.56	.14	.08	.56	.16	.15	.17
vua	.13	.08	.37	.13	.07	.51	.00	.00	.00
reddit+sermon	.15	.09	.51	.17	.10	.50	.19	.26	.15
tedx+sermon	.17	.10	.51	.17	.10	.47	.00	.00	.00
vua+sermon	.17	.10	.54	**.18**	**.11**	.50	.08	.24	.05
reddit+tedx	.15	.09	.45	.17	.10	.55	.14	.23	.10
vua+tedx	.13	.08	.47	.16	.10	.51	.12	.26	.08
vua+reddit	.14	.08	.56	.14	.08	**.66**	.12	.17	.09
reddit+tedx+sermon	.16	.10	**.57**	**.18**	**.11**	.54	.20	.24	.17
vua+tedx+sermon	**.18**	**.11**	.50	**.18**	**.11**	.53	.16	.30	.11
vua+reddit+sermon	.17	.10	.50	**.18**	**.11**	.48	.24	.25	**.23**
vua+reddit+tedx	.13	.08	.48	.17	.10	.61	.15	.26	.10
vua+reddit+tedx+sermon	.17	.10	.50	**.18**	**.11**	.55	**.25**	**.33**	.20

Table 6: Results on **Sermon set A** (condensed setting).

D Results on Sermon Set B (Condensed)

	XLM-RoBERTa											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.24	.19	.33	.00	.00	.00	.00	.00	.00	.00	.00	.00
tedx	.18	.14	.25	.02	.23	.01	.00	.00	.00	.00	.00	.00
reddit	.26	.17	.50	.26	.23	.31	.24	.22	.26	**.21**	.27	**.17**
vua	.22	.18	.30	.12	.18	.09	.09	.21	.06	.01	**.50**	.01
reddit+sermon	.28	.20	.51	.29	.28	.29	.24	.29	.21	.05	.30	.03
tedx+sermon	.28	.20	.50	.21	.26	.18	.04	**.60**	.02	.00	.00	.00
vua+sermon	.29	.21	.45	.29	.31	.27	.21	.29	.17	.12	.32	.08
reddit+tedx	.27	.18	.59	.23	.23	.23	.22	.26	.20	.16	.32	.11
vua+tedx	.25	.17	.47	.23	.19	.29	.22	.27	.18	.08	.30	.05
vua+reddit	.26	.18	.47	.31	.26	.38	.27	.30	.25	.18	.23	.14
reddit+tedx+sermon	**.31**	.20	.64	.31	.24	**.43**	.26	.28	.24	.18	.28	.13
vua+tedx+sermon	**.31**	**.22**	.51	.27	.27	.28	.24	.28	.21	.06	.29	.03
vua+reddit+sermon	.30	.20	.61	**.33**	**.32**	.34	**.34**	.32	**.36**	.20	.35	.14
vua+reddit+tedx	.27	.17	.61	.29	.24	.37	.28	.27	.29	.10	.25	.06
vua+reddit+tedx+sermon	.28	.18	**.65**	.32	.29	.35	.32	.36	.29	.17	.29	.12

	mBERT											
	sentence			w50			w100			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.28	**.19**	.55	.23	.25	.20	.11	.25	.07	.00	.00	.00
tedx	.22	.14	.48	.21	.18	.25	.09	.11	.08	.00	.00	.00
reddit	.21	.13	.53	.20	.18	.24	.20	.16	.28	.18	.18	.17
vua	.23	.17	.39	.19	.24	.16	.11	.18	.08	.02	.03	.02
reddit+sermon	.27	.18	.60	.27	.25	.29	.26	.29	.24	**.23**	.27	**.21**
tedx+sermon	.26	.17	.54	.26	**.26**	.26	.23	.27	.19	.06	.23	.03
vua+sermon	.27	.17	.57	.28	.24	.33	.21	.26	.18	.17	.30	.12
reddit+tedx	.25	.15	.62	.25	.20	.35	.20	.19	.21	.06	.16	.04
vua+tedx	.21	.12	**.66**	.24	.21	.29	.25	.23	.26	.09	.24	.05
vua+reddit	.24	.15	.64	.25	.20	.33	.21	.23	.19	.19	.19	.19
reddit+tedx+sermon	.26	.16	.63	.27	.25	.29	.25	.25	.25	.18	.28	.13
vua+tedx+sermon	.26	.17	.56	.29	.24	.35	.24	.29	.20	.18	**.32**	.13
vua+reddit+sermon	**.29**	**.19**	.62	**.30**	.25	**.36**	**.28**	**.30**	.26	.21	**.32**	.16
vua+reddit+tedx	.23	.14	.59	.25	.21	.32	.24	.20	.28	.11	.19	.08
vua+reddit+tedx+sermon	.28	.18	.59	.27	.24	.30	**.28**	.27	**.30**	.10	.22	.07

	ST de-en			ST multilingual			XLM-Long		
	sentence			sentence			document		
train	**f1**	**p**	**r**	**f1**	**p**	**r**	**f1**	**p**	**r**
sermon	.27	**.20**	.39	.25	**.20**	.33	.00	.00	.00
tedx	.20	.12	.54	.19	.13	.35	.00	.00	.00
reddit	.22	.13	.71	.21	.13	.66	.19	.20	**.18**
vua	.21	.13	.48	.18	.11	.51	.00	.00	.00
reddit+sermon	.27	.17	.60	.27	.17	.63	.19	**.39**	.12
tedx+sermon	.27	.18	.56	.27	.17	.61	.00	.00	.00
vua+sermon	.25	.16	.64	**.28**	.18	.62	.13	.33	.08
reddit+tedx	.25	.16	.60	.24	.15	.64	.16	.28	.12
vua+tedx	.22	.13	.64	.23	.15	.58	.13	.30	.08
vua+reddit	.23	.14	.72	.20	.12	**.73**	.15	.22	.11
reddit+tedx+sermon	**.29**	.19	.65	.26	.17	.58	.15	.32	.09
vua+tedx+sermon	.27	.17	.63	.27	.17	.60	.11	.38	.07
vua+reddit+sermon	.25	.15	**.73**	.26	.16	.61	**.23**	.32	**.18**
vua+reddit+tedx	.24	.15	.71	.24	.14	.66	.13	.25	.09
vua+reddit+tedx+sermon	.26	.16	.71	.27	.17	.60	.18	.31	.13

Table 7: Results based on **Sermon set B** (condensed setting).

E Model References

mBERT	https://huggingface.co/google-bert/bert-base-multilingual-cased
XLM-RoBERTa	https://huggingface.co/FacebookAI/xlm-roberta-base
Cross EN & DE RoBERTa	https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer
multilingual SBERT	https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
XLM-Long	https://huggingface.co/markussagen/xlm-roberta-longformer-base-4096

Table 8: Links to used models, last accessed on December 23, 2025.