OPINIONS ARE BUILDINGS: Metaphors in Secondary Education EFL Essays

Anna Hülsing[1], Andrea Horbach[2,3,4]
[1]University of Hildesheim, Germany
[2]Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[3]University of Kiel, Germany, [4]FernUniversität in Hagen, Germany
huelsing@uni-hildesheim.de
horbach@leibniz-ipn.de

Abstract

Automatic metaphor detection has been an active field of research for years. Yet, it was rarely investigated how automatic metaphor detection can aid language learning. We therefore present MEWSMET, a corpus of argumentative essays (MEWS[1]) written by English as Foreign Language (EFL) learners annotated for metaphors. We differentiate between two kinds of metaphors: metaphors that are comprehensible to native speakers, even though they themselves would not use them (comprehensible metaphors, CMs) and metaphors that native speakers would use (target language metaphors, TLMs). We use MEWSMET in two ways: Firstly, we analyze our annotations and find out that there is a positive linear correlation between essay score and the number of TLMs, while no correlation is found between essay score and the number of CMs. Secondly, we explore how metaphor detection models perform on MEWSMET. We find that metaphor detection is a hard task given our noisy learner data, and that metaphor detection models tend to be better at identifying all metaphors (TLMs+CMs) instead of just TLMs, even though only TLMs can be used as a feature for automatic essay-scoring.

1 Introduction

Conceptual Metaphor Theory claims that metaphorical linguistic expressions manifest our way of thinking. One of the most well-known examples for a metaphorical linguistic expression is *to spend time*. Here, the conceptual domain TIME is described by means of the conceptual domain MONEY. The metaphorical linguistic expression thus shows that time is considered a limited and valuable resource (Lakoff and Johnson, 1980b). Metaphorical linguistic expressions are therefore not merely ornamental, but omnipresent in our everyday life (Lakoff and Johnson, 1980a, Shutova and Teufel, 2010).

Detecting metaphorical linguistic expressions automatically is beneficial for a range of natural language processing applications, such as emotion detection (Dankers et al., 2019, Li et al., 2022), identification of mental health problems (Zhang et al., 2021, Gutiérrez et al., 2017), or propaganda detection (Baleato Rodríguez et al., 2023). Even though metaphors play an important role in education (Niebert and Gropengiesser, 2012, Mouraz et al., 2013, Oxford et al., 1998), it is only rarely investigated how metaphor detection (MD) can be employed to facilitate language learning.

Beigman Klebanov et al. (2018) have presented a corpus annotated for metaphors that is based on the ETS Corpus of Non-Native Written English[2] – a collection of argumentative essays provided by TOEFL test takers. They show that the use of argumentation-relevant metaphors provides information about a writer's English language proficiency. We build on and extend this work in several ways as detailed in the following.

First, our study addresses whether the same relation between metaphoric language use and language proficiency also holds for younger writers. Although mean age of the writers in the study by Beigman Klebanov et al. (2018) is not given (Blanchard et al., 2013), we assume that – as

[1]Measuring Writing at Secondary Level (see Keller, 2016 and Keller et al., 2020)

[2]https://catalog.ldc.upenn.edu/LDC2014T06

Anna Huelsing and Andrea Horbach. Opinions Are Buildings: Metaphors in Secondary Education Foreign Language Learning. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 78–95.

[Children] are likely to take over *(adopt)* the opinion of the people [...] that are around them.
Young children should live their lifes and should not have to build (*form*) their own opinion about something.
This often brings *(puts)* parents in difficult situations.

Table 1: Example sentences with metaphorically used verbs (underlined) taken from MEWS data (Keller et al., 2020). They are comprehensible in English, even though L1 English speakers would probably use different expressions such as the ones given in brackets.

TOEFL tests are often taken by students who want to study at a university where English is the language of instruction – most writers are in their last year of high-school or have recently graduated from high-school. In contrast, our study is based on the MEWS dataset by Keller et al. (2020), which addresses German-speaking EFL learners in earlier years of their education, while also using TOEFL writing prompts[3]. We assume that the general proficiency level will be lower in our dataset than in the one by Beigman Klebanov et al. (2018). In addition, our dataset comprises essays of all proficiency levels, while the one by Beigman Klebanov et al. (2018) only consists of medium- and high-proficiency essays.

Secondly, we investigate the relationship between proficiency level and metaphors that English L1 speakers comprehend, even though they themselves would not actively use them; examples are shown in Table 1. Samaniego Fernández et al. (2005) demonstrate that professional translators introduce new expressions and conceptual structures in a target culture when transferring metaphors that are non-novel in the source language to a novel metaphor in the target language. The translated expressions "seem to have been understood correctly, and this proves their [i.e. the metaphors'] transparency: they can be interpreted precisely because they appeal to our recognition of underlying symbolism." In our dataset, students also use metaphors that seem anomalous in the target language in the sense that L1 speakers would not use them. Yet, the metaphorical expressions are perfectly comprehensible for target language speakers because they create new (and sometimes even appealing) conceptual mappings (e.g. *to build an opinion*: an opinion is – or should be – hard work just as building a house)[4]. We will call these metaphors **comprehensible metaphors** (CMs), as opposed to metaphors which target language speakers would actively use (**target language metaphors**, TLMs). We will examine the scores human raters gave to essays containing CMs in order to find out whether they rather occur in low- or high-proficiency essays.

Next, we investigate how well metaphor detection models perform on more noisy data from such younger, and partly less-proficient writers in detecting metaphors – both CMs and TLMs. To do so, we leverage the best-performing model from the 2020 Shared Task on Metaphor Detection (Leong et al., 2020), namely DeepMet (Su et al., 2020). Our study will focus on verbs only for several reasons. First, Cameron (2003) report that about half of all metaphors in educational discourse are found in verbs. Second, other parts of speech, especially prepositions, are often not seen as being metaphorical by laypeople (cf. Beigman Klebanov and Flor, 2013), which would pose an additional difficulty during the annotation process. Third, many metaphor detection datasets that potentially serve as training data, have been annotated just for verbs.

Our study makes the following contributions: **1)** We present the MEWSMET corpus. Here, an additional layer is added to the MEWS-dataset (Keller et al., 2020), where we annotated metaphors that are perfectly acceptable in the target language English (TLMs) as well as metaphors which are comprehensible but which native speakers would not use (CMs). **2)** We describe the relationship between TLMs and the scores human raters attributed to the student essays. We do so to confirm the trend Beigman Klebanov et al. (2018) have observed for high-school graduates also for younger and less-proficient students, namely that the use of metaphors provides insights into a learner's proficiency level. **3)** We describe the relationship between CMs and students' proficiency levels. **4)** We provide insights into the behaviour of metaphor detection models on noisy learner data for both TLMs and CMs.

For code and data see `https://github.com/AnHu2410/MEWSMET_code`.

[3]The prompts are different from those used in the ETS Corpus of Non-Native Written English, i.e. also different from the TOEFL dataset by Beigman Klebanov et al. (2018).

[4]The expression *to build an opinion* is based on a false friend, as the German equivalent to *to form an opinion* is *(sich) eine Meinung bilden*, where the word *bilden* is phonologically similar to the English verb *to build*.

2 Related Work

In this section we provide the scientific background to the three main fields of this study: metaphor annotation, metaphor detection and automatic essay scoring.

2.1 Metaphor Annotation

A widely applied example of a metaphor annotation guideline is the Metaphor Identification Procedure (MIP; Pragglejaz Group, 2007) and its extension, MIPVU (Steen et al., 2010). The underlying idea is that a token is used metaphorically if its meaning in a certain context deviates from a more "basic" meaning of this word, as defined by a contemporary dictionary. For example, the basic (i.e. first) meaning of the verb *to build* according to the online version of the Longman Dictionary of Contemporary English[5] is *to make something, especially a building or something large*, with examples ranging from houses and bridges to birds' nests. In the expression *to build an opinion*, clearly this concrete basic meaning is not applicable.

We follow the annotation guideline from Mohammad et al. (2016). It is based on MIP (Pragglejaz Group, 2007), but condensed and enriched with examples (see Appendix A.1.1), which we deemed suitable for our annotators who had no prior experience in the identification of metaphors. While MIP and MIPVU were originally designed for annotating metaphors in English, there have been attempts to use the guidelines for other languages such as Spanish (Sanchez-Bayona and Agerri, 2022).

Beigman Klebanov et al. (2018) annotate argumentation-relevant metaphors, i.e. metaphors that help the writer of an argumentative essay to advance an argument. In stark contrast to Pragglejaz Group (2007) and Steen et al. (2010), they did not provide "formal definitions of what a literal sense is in order to not interfere with intuitive judgments of metaphoricity" (Beigman Klebanov and Flor, 2013). This line of thought also emerges in other annotation studies, such as Tsvetkov et al. (2014) and Piccirilli and Schulte im Walde (2022), as they rely on intuitive definitions of metaphoricity.

The distinction between – and annotation of – novel and conventionalized metaphors is an increasingly active research topic (Parde and Nielsen, 2018, Do Dinh et al., 2018, Egg and Kordoni, 2022, Reimann and Scheffler, 2024a). This distinction is also relevant for our dataset, and has been been annotated for future use. Another distinction which is highly relevant for our study is given by Reijnierse et al. (2018), who in their annotation protocol differentiate between deliberately and non-deliberately used metaphors. After all, deliberately used metaphors cannot simply be learnt from a textbook and could therefore hint at a higher language competency. We have not annotated whether or not metaphors in our dataset are used deliberately, but leave this to future work.

2.2 Metaphor Detection

An early approach to automatic metaphor detection was developed by Birke and Sarkar (2006), who used a word-sense disambiguation approach to classify literal and non-literal usages of verbs. Conceptual Metaphor Theory (Lakoff and Johnson, 1980b) claims that metaphors transfer knowledge from a concrete, familiar domain to a more abstract domain. Therefore, Turney et al. (2011) used abstractness scores of context words as features for their logistic regression classifier. The idea of "conceptual features" also inspired Tsvetkov et al. (2014) and Köper and Schulte im Walde (2016), who – in addition to abstractness and other scores – used semantic supersenses from WordNet (Miller, 1994) and scores representing distributional fit, respectively.

Early neural models, such as Do Dinh and Gurevych (2016) (a multilayer perceptron with word embeddings), showed a performance comparable to non-neural classifiers; however, they became popular because they did not require feature engineering. Later neural models clearly outperformed the non-neural classifiers: Dankers et al. (2019) used several multi-task learning models and reached state-of-the-art results in 2019 for both metaphor and valence/arousal/dominance (VAD) prediction. During the 2020 Metaphor Detection Shared Task (Leong et al., 2020), DeepMet overall performed best (Su et al., 2020); the authors transformed metaphor detection into a reading comprehension task and observed state-of-the-art results. We use this model in our study to compare its performance on the corpus by Beigman Klebanov et al. (2018) and our corpus.

[5] https://www.ldoceonline.com. We use this corpus-based dictionary for our annotation since it was also used by Steen et al. (2010).

Ma et al. (2021) fine-tuned BERT for MD. To perform word-based metaphor classification, they copied the input sentence and masked the target word. The original sentence and the masked copy were used as input for a sequence classification task. Uduehi and Bunescu (2024) also mask the target word, and compute the expectation of a literal meaning in the given context. Then, they compute the estimation of the realized meaning of the target word in order to predict whether the target word violates the expectation of a literal word. Li et al. (2023) exploited the fact that many datasets are based on the Metaphor Identification Process (MIP; Pragglejaz Group, 2007), where a word is annotated as metaphorical if its contextual meaning is dissimilar to its "more basic meaning" (among further criteria). While prior models (such as MelBERT by Choi et al. 2021) grounded on MIP use decontextualized representations of the target word, Li et al. (2023) successfully gathered the representation of the target word from sentences where it was used literally. While research on metaphors in English has received a lot of attention, and also metaphor detection in other low- to high-resource languages is turning into an active field (Aghazadeh et al., 2022, Lai et al., 2023, Schuster and Markert, 2023), research on metaphors in texts of English learners is rare. Stemle and Onysko (2018) used a bidirectional RNN and fastText embeddings to detect metaphors for the 2018 Shared Task on Metaphor Detection (Leong et al., 2018). As training data for their embeddings they use TOEFL tests (Blanchard et al., 2013) of different proficiency levels (among others); in contrast to our study (and that of Beigman Klebanov et al., 2018), they use these texts to detect metaphors in standard language and not in learner language.

2.3 Metaphors in Automatic Essay Scoring

In automatic essay scoring, the task is to predict the quality of an essay either on a holistic scale or for specific aspects of an essay such as language or structure. For holistic scoring, both linguistic form and content are usually taken into consideration and the correct usage of metaphoric expressions can be seen as one aspect of linguistic proficiency. Yet to the best of our knowledge, metaphors have so far not been integrated into automated essay scoring systems, as – until some years ago – it has been claimed that the automatic metaphor detection for non-conventionalized metaphors would not work reliably enough (Graesser and McNamara, 2012). For essay scoring in Chinese, Yang et al. (2019) used a number of features, including the number of metaphors. Given their examples, though, their notion of metaphors rather corresponds to a simile with specific lexical items marking their occurrence. However, there have been recent successes integrating features on the related topic of concreteness of multi-word expressions into essay scoring (Wilkens et al., 2022) highlighting the importance to consider complex linguistic phenomena.

3 Annotation Study: Metaphor Annotation in Learner Essays

The goal of the following annotation study is twofold: First, we aim at investigating the relationship between essay scores and the use of metaphors. Second, our annotation results are used as train and test sets in the subsequent experimental study on automatic metaphor detection in learner texts.

3.1 Annotation Data

The dataset used in our study is a subset of the MEWS dataset by Keller et al. (2020), a collection of argumentative essays written by German and Swiss EFL learners. The essays are written on the basis of four different TOEFL prompts, two of them being independent prompts (the students are given a prompt only) and two of them being source-based, i.e. the prompt refers to a reading text. We focus on the independently-written essays only as source-based essays might mainly contain metaphors in standard language adopted from the text. The following two prompts were used. The students were asked whether they "agree or disagree with the following statement":

- *Television advertising directed toward young children (aged two to five) should not be allowed.* (Prompt "TV-Ads")
- *A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught.* (Prompt "Teacher")

For each essay, expert raters' scores are available. Two raters scored the essays on a scale between one and five (with five being the best score). If the two ratings were only one point apart (e.g. rater A: 3; rater B: 4), the average was taken as the

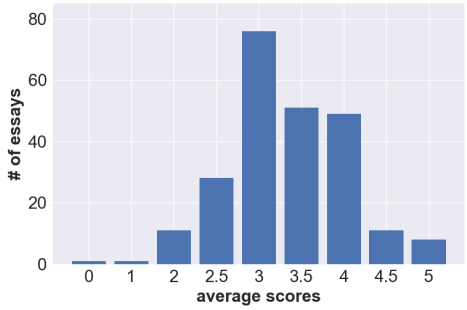

Figure 1: Number of essays per score in MEWSMET.

overall score. Otherwise, a third adjudicator rated the essay in order to obtain the overall score. The essays were written in the penultimate year before graduation; half of them at the beginning of the school year (T1) and half at the end (T2).

We randomly selected 236 essays (120 with prompt "TV-Ads", 116 with prompt "Teacher") from Swiss students; Figure 1 shows the number of essays per score. As can be seen, all proficiency levels are taken into account. These essays contain 8025 target verbs (excluding stop words, see Appendix B) which were automatically detected with the off-the-shelf NLTK POS-tagger which utilizes the Penn Treebank tagset (Bird and Loper, 2004).

3.2 Annotation Guidelines

Our goal is the annotation of verbal metaphors in learner essays. Our guidelines were adopted from Mohammad et al. (2016), who provide specific definitions for metaphorical and non-metaphorical usages compared to guidelines that rely on intuition (cf. Beigman Klebanov and Flor, 2013). We deemed this kind of guidance helpful for this structurally difficult task. In contrast to Beigman Klebanov et al. (2018), who only focus on argumentation-relevant metaphors, we chose to annotate all verbal metaphors in order to have more comprehensive material for analysis. In addition to a binary decision for metaphorical vs. literal usage, annotators had to label each target verb with one of the following four labels (examples taken from MEWSMET):

- **non-metaphorical**: for literal usages, e.g. *children learn in their smal age to consume and to spend money*
- **conventional metaphor**: for frequent metaphorical usages the annotator has seen before, e.g. *Sometimes you spend even more time with a particular teacher than your parents.*
- **creative metaphor**: when the annotator felt that the verb was metaphorical but rarely used in this context, e.g. *[TV-]channels [are] flooded with tons of ads.*.
- **uncommon translation of a German conventionalized metaphor**: a metaphor that has a German conventionalized metaphor as basis, but the English translation used here is uncommon, e.g. *This often brings parents in difficult situations.* (literal translation of the following German sentence: *Das bringt Eltern oft in schwierige Situationen.*).

The guidelines can be found in Appendix A.1.1.

3.3 Annotation Procedure

The annotation procedure was conducted in three stages using the annotation platform INCEpTION (Klie et al., 2018) as detailed in the following.

Phase 1 – Sample Annotation by Experts: Annotating metaphors is generally considered a difficult task with rather low inter-annotator agreement. For example, Reimann and Scheffler (2024a) report a Cohen's $\kappa = 0.60$ for annotating metaphors in religious online forums after discussing disagreements and adjudication, and Beigman Klebanov et al. (2018) report a Cohen's $\kappa = 0.56$ after a first round of annotation and a Cohen's $\kappa = 0.62$ after showing the annotators their partner's annotations for essays with high disagreement values, and asking them to reconsider their original annotations. Annotating not standard language, but learner language adds an extra layer of difficulty. To check the feasibility of the task and the quality of our annotation guidelines, we first asked two Swiss-German researchers in the field of English didactics to annotate a small subset (5 essays) sampled from MEWS that is not part of the subset described above (Section 3.1). The annotators were given the main annotation guideline as presented in Appendix A.1.1.

In this first annotation round, inter-annotator agreement was low (Cohen's $\kappa = 0.22$). Therefore, we discussed unclear cases and extended the main guideline (see Appendix A.1.2) to improve their clarity. Based on these improved guidelines, the experts annotated a second sample of 5 essays; as Cohen's κ increased to a value of 0.37, we con-

sidered the guideline additions to be useful. Of course, the inter-annotator agreement still was not even moderate; however, given the difficulty of the task, we considered it to be sufficient for a first round of annotations.

Phase 2 – Main Annotation Study: Next, 236 essays taken from the MEWS corpus (see Section 3.1) were annotated by two annotators, who are pursuing their master's degrees to become English teachers in Germany. For the purpose of training, they first annotated a MEWS-based toy corpus on the basis of our revised guidelines and discussed the results. Both annotators independently annotated the actual data. For adjudication after the first round of annotations, Annotator A was given the information whether her annotation differed from the annotations of Annotator B. The difference can be one of the following:

- A: metaphor, B: no / uncommon metaphor
- A: uncommon metaphor, B: metaphor / no metaphor
- A: no metaphor, B: uncommon metaphor/ metaphor

I.e., the nature of the difference was not disclosed to the annotator and the difference between creative and conventional metaphor was not taken into account at all.

Annotator A was asked to check these cases and correct them if she made an obvious mistake. After that, Annotator B did the same for the remaining disagreements. Finally, the first author of this paper manually checked all annotations and discussed cases which possibly contradicted the annotation guidelines with Annotators A and B.

Phase 3 – Check by Native Speakers: Two English L1 speakers (one American English, one British English speaker) were asked to check whether the metaphors found in Phase 2 were a) expressions that a L1 English speaker might use, b) that an L1 English speaker would not use but which are comprehensible, and c) that are incomprehensible. To avoid bias by language errors surrounding the metaphorical expression, the sentences were corrected and only the relevant part of the sentence was shown to the annotators.

3.4 Annotation analysis

3.4.1 Inter-Annotator Agreement

After the first round of the main metaphor annotation study, agreement for the binary decision between metaphorical and non-metaphorical was moderate with Cohen's $\kappa = 0.42$. As mentioned before, metaphor annotation generally is a field with rather low inter-annotator agreement, and using learner essays from all proficiency levels poses an additional difficulty. Therefore, the low level of agreement after the first round was to be expected. After the final round, Cohen's κ reached a high value of 0.88. The confusion matrix for the binary decision is shown in Table 2; even though for 89 target verbs the annotators did not agree, for the vast majority they agreed in their annotations. For 362 target verbs they agreed that they are used metaphorically.

		Ann B	
		met	non
Ann A	met	362	57
	non	32	7574

Table 2: Confusion matrix illustrating the inter-annotator agreement for the binary metaphor annotation task (*metaphorical* vs. *non-metaphorical*).

Agreement for the 4-way-task ("conventionalized", "creative", "uncommon", "no metaphor") was lower with Cohen's $\kappa = 0.74$, and the annotations are represented in the confusion matrix shown in Table 3. While agreement on non-metaphorical expressions is very high and they mostly agreed on metaphors that are based on German conventionalized expressions that are uncommon in English, disagreement was high for whether a metaphor is creative or conventional. The distinction between creative and conventional is hard even for native speakers (compare Parde and Nielsen, 2018); as our annotators are not native speakers, the distinction is even harder, because they are not as familiar with certain conventionalized expressions as native speakers are.

		Ann B			
		conv	creat	unc	non
Ann A	conv	183	6	6	28
	creat	94	34	4	27
	unc	3	3	29	2
	non	23	4	5	7574

Table 3: Confusion matrix illustrating the inter-annotator agreement for the metaphor annotation (4-way-annotation: *conventional*, *creative*, *uncommon*, *non-metaphorical*).

		Ann 2		
		incompr	compr	L1
Ann 1	incompr	3	6	1
	compr	6	20	1
	L1	7	91	227

Table 4: Confusion matrix illustrating the inter-annotator agreement for the check by native speakers (*incomprehensible metaphor*, *comprehensible metaphor*, *L1 metaphor*).

For the native speaker check, i.e., the 3-way annotation whether a metaphor was L1-like, comprehensible or incomprehensible, Cohen's κ reached a value of 0.24. This rather low value is mostly caused by the fact that more metaphors were considered L1-metaphors for Annotator 1 than for Annotator 2 (see Table 4). Annotator 1 was more tolerant towards metaphors such as *to fall into a down* (meaning: *to become depressed*) that could be seen as a creative invention of the writers. This may be due to Annotator 1's Bachelor's degree in English Language and Creative Writing (Annotator 2 had no background relevant for the task).

3.4.2 Quantitative Analysis

We collected annotations for a total of 8025 target words in 236 essays. We only counted those target verbs as being used metaphorically in the subsequent analyses where both annotators agreed that the verb was metaphorical, i.e. where they chose one of the following labels: conventional metaphor, creative metaphor, or uncommon translation of a German conventionalized metaphor. This was the case for 362 verbs. These 362 verb tokens consisted of 149 types. We did not perform adjudication for the individual labels (e.g., conventional metaphor), as we only take into account the binary label (metaphorical, non-metaphorical) in this study.

The 362 verbs that both German annotators annotated as being metaphorical were shown to the English native speakers. For their check, we decided to err on the side of caution and use the least optimistic label, i.e. if one annotator decided that an expression is incomprehensible while the other decided it was comprehensible, we chose the label "incomprehensible". 23 target verbs were annotated as being incomprehensible by at least one annotator. These target verbs were counted as being non-metaphorical, even though the writer might have intended to use a metaphor here. 112 target verbs were annotated as being comprehensible (CMs). For 227 metaphorical expressions both annotators declared that they could have been uttered by an English L1 speaker (TLMs).

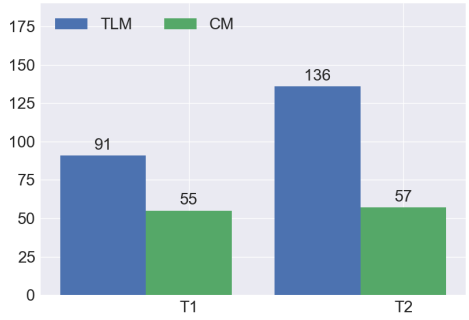

Figure 2: Number of CMs and TLMs found at beginning (T1) and end of school year (T2).

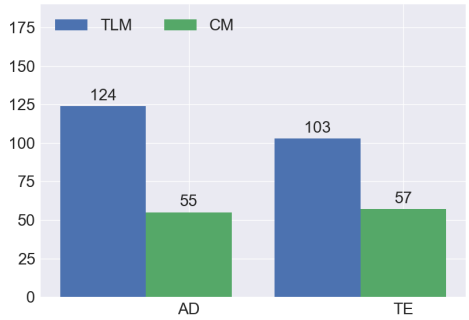

Figure 3: Number of CMs and TLMs found for the prompt "TV ads" (AD) and for the prompt "Teacher" (TE).

Figure 2 shows the amount of metaphors (TLMs and CMs) found at T1 and T2, respectively. As can be seen, the amount of CMs stays nearly the same for T1 and T2, but the amount of TLMs rises by 50%. This indicates that the learners' proficiency improves within one year, and that TLMs could be a useful feature in essay scoring, whereas CMs might not be.

Figure 3 shows the amount of metaphors (TLMs and CMs) found for each prompt. While slightly more TLMs occur in the essays on TV-Ads than on Teachers, the number of CMs is roughly the same for both prompts. The balance between both prompts is important, since we split the entire MEWSMET-corpus into two parts (MEWS_Ads and MEWS_Teacher), and use them as training and testing data.

3.4.3 Relationship between Metaphors and Essay Quality Scores

In order to investigate the relationship between the number of metaphors per essay and the essay's

holistic score, we counted the number of TLMs and CMs in each essay, and normalized the number of TLMs and CMs by the number of each essay's characters in order to control for essay length. Then we obtained the score (1 to 5) attributed to each essay by expert raters. The correlations between the number of TLMs, CMs as well as all metaphorical expressions (TLMs plus CMs), each divided by the number of characters, and the respective essay scores in terms of Pearson's ρ are presented in Table 5.

The results show that there is a weak, yet significant positive linear correlation (p-value < 0.05) between essay score and the number of metaphors that English L1 speakers would use (TLMs). No correlation between the essay scores and the number of comprehensible metaphors (CMs) was observed. The combined correlation between essay score and all metaphorical expressions (both TLMs and CMs) is weak, and this correlation is not significant.

	Pearson's ρ	**p-value**
TLM/score	0.143	0.028
CM/score	-0.011	0.863
both/score	0.118	0.070

Table 5: Pearson's correlation between target language metaphors / comprehensible metaphors / both types of metaphors combined (controlled for essay length) and essay score.

4 Experimental Study: Automated Metaphor Identification in Learner Essays

After having manually identified metaphors in the previous section, we now turn to the question of how well existing metaphor detection algorithms perform on MEWS learner data using our annotations as gold standard.

4.1 Experimental Setup

4.1.1 Classifier

We use DeepMet (Su et al., 2020) to detect metaphors in learner text. DeepMet transforms metaphor detection into a reading comprehension task, i.e. the model is trained to answer questions based on a given sentence. Their model takes the global context (i.e. the whole sentence), local context (i.e. the words before and after the target word that are enclosed by punctuation such as commas) and two types of part-of-speech as features, which are represented via BERT embeddings. These embeddings are fed into a siamese architecture based on two Transformer encoder layers. Their output is reduced to one feature vector by average pooling, which is the input to a metaphor discrimination layer. We chose this model as it showed the best performance in the 2020 metaphor detection Shared Task (Leong et al., 2020).

4.1.2 Evaluation Procedure

We use the evaluation procedure presented in Su et al. (2020), where stratified 10-fold cross-validation is performed. In each fold, a model is trained based on a subset (90%) of the training data and used to make predictions on the entire set of the test data. The predictions for all training folds are summed up (leading to a number between 0 and 10 for each test instance i). This sum is divided by the number of folds (in our case 10). A metaphor preference parameter α (determined in previous experiments) indicates which prediction is the final prediction for each test instance. The default value is 0.2, so if at least two models predicted instance i to be metaphorical, the final prediction is metaphorical; else, the final prediction is non-metaphorical.

4.1.3 Training Data

As mentioned before, we use two splits of MEWSMET, namely MEWS_Teacher and MEWS_Ads. We train and test in both directions, i.e. we train on MEWS_Teacher and test on MEWS_Ads and vice versa. In addition to the data we annotated ourselves, we use two other datasets: firstly, a very large corpus annotated for metaphors, namely the VU Amsterdam Metaphor Corpus (VUA) by Steen et al. (2010). This corpus is sampled from the British National Corpus (BNC) and covers academic texts, conversation, fiction, and news texts, which means that it contains standard English. The data was annotated under the MIPVU protocol (Steen et al., 2010). Secondly, we use the TOEFL corpus which, as mentioned before, is sampled from the ETS Corpus of Non-Native Written English, and contains argumentative essays written by EFL learners shortly before or after graduating from secondary education. Even though this corpus is not as big as the VUA corpus, it contains learner language similar to MEWSMET. Only argumentation-relevant

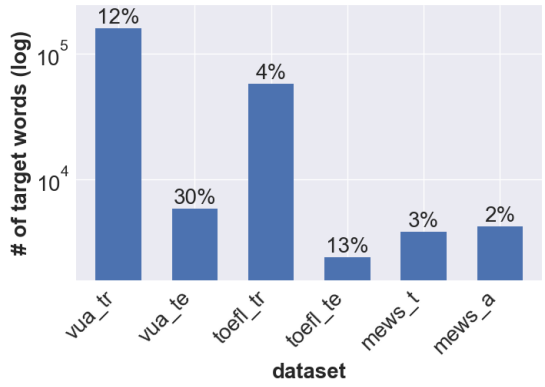

Figure 4: Amount of training (*tr*) and testing data (*te*) for VUA and TOEFL compared to size of MEWS_Teacher and MEWS_Ads on a logarithmic scale. Percentage of metaphorically used target verbs given on top of each column.

metaphors were annotated here (Beigman Klebanov et al., 2018). For both VUA and TOEFL, we used metaphor annotations for all parts of speech for training, and we evaluated on the datasets where only verbs are annotated for metaphoricity (as done by Su et al., 2020). The stark differences in the amounts of training and testing instances for the two additional corpora, compared to our dataset, are illustrated in Figure 4.

4.1.4 Computing Hours and Infrastructure

It took about 30 hours to train the VUA model, 12 hours to train the TOEFL model and 2 hours to train the MEWS models. Experiments were performed on an AMD EPYC 74F3 24-Core Processor and NVIDIA RTX A6000 GPUs.

4.2 Performance of Metaphor Detection Method on MEWSMET

If we want to use metaphors as features for automatic essay scoring, they have to be detected automatically and reliably. Therefore we investigate how well metaphor detection models perform on noisy student data such as MEWS.

4.2.1 Experiment 1: Metaphor detection performance across different datasets

To assess how hard the task of metaphor detection is on our dataset compared to existing metaphor datasets, we compare performance across datasets when training and testing on data from the same dataset. Results for training and evaluating the DeepMet model on VUA and TOEFL data are reported in Su et al. (2020); to ensure comparability with our results on MEWS data we repeated the experiments on our own GPU machines.[6].

To compare the performance on VUA and TOEFL to our data, we first used the MEWS_Teacher split of our data for training and MEWS_Ads for testing, and secondly MEWS_Ads for training and MEWS_Teacher for testing. In both datasets, we only considered TLMs, since we assumed that this metaphor type is closer to the metaphors annotated in VUA and TOEFL. The results are shown in Table 6. The hyperparameters were taken from the paper by Su et al. (2020) with seed = 12.

	Precision	**Recall**	**F1**
VUA	70.9	81.9	76.0
TOEFL	64.1	82.8	72.3
MEWS_T	35.8	19.4	25.1
MEWS_AD	56.3	8.7	15.1

Table 6: Results for training on VUA / TOEFL / MEWS data and testing on the corresponding test data. For MEWS, the training data is mentioned in the table (e.g. MEWS_T refers to training on MEWS_Teacher and evaluating on MEWS_Ads). Here, only TLMs were taken into account. The results are determined with the preference parameter $\alpha = 0.2$.

In this evaluation setup, DeepMet performs best on the VUA data, closely followed by the TOEFL data. On MEWS_Ads and MEWS_Teacher it performs worst by a large margin.

In the course of the evaluation we observed that for the two MEWS test datasets the results also varied greatly across different training folds, while this was not the case for VUA and TOEFL data. Table 7 shows the mean and standard deviation (SD) of precision, recall and F1 across all folds for the 4 models. Here, precision, recall and F1 are calculated for each fold without using the preference parameter α.

While the F1 standard deviation for VUA and TOEFL is lower than 2 F1-points, for MEWSMET (trained on MEWS_Teacher) it is 10.7 points and 6.5 (trained on MEWS_Ads). During cross-validation, the test data stays the same, and as 90% of the training data are used for each fold, the difference between the individual folds does not vary largely either. The best guess is that the extreme

[6]Su et al. (2020) report $F1 = 80.4$ for VUA-verb and $F1 = 74.9$ for TOEFL-verb. We attribute differences to our results to slightly different GPU settings.

	Precision	Recall	F1
VUA	77.6 ± 2.1	69.6 ± 3.9	73.3 ± 1.6
TOEFL	72.0 ± 4.8	64.9 ± 6.5	67.9 ± 1.8
MEWS_T	34.3 ±33.6	8.2 ± 11.1	10.3 ±10.7
MEWS_AD	52.0 ± 32.4	4.2 ± 4.3	7.1 ± 6.5

Table 7: Mean and SD scores across precision, recall and F1 for test data on each training fold without using the preference parameter α. The folds are identical with the ones used for Table 6.

differences in training data size account for this behaviour. In order to see whether the training dataset is indeed too small for the model to learn properly, we shrunk the VUA training dataset to a size comparable to MEWS_Teacher; the Mini-VUA consists of 3600 training instances, of which 100 are tagged as being metaphorical. The result for training on Mini-VUA and testing on the VUA test dataset is a precision of 26.6 ± 36.8, a recall of 0.2 ± 0.5, and an F1-score of 0.4 ± 0.1. These numbers show that the model does not learn at all from Mini-VUA. We therefore expect our models to perform better with a larger amount of training data, too.

4.2.2 Experiment 2: Model Performance for Different Training Datasets

As discussed above, larger amounts of training data are needed for DeepMet to perform well on MEWSMET. Therefore, we investigated which training data is most suitable for our task of detecting metaphors in learner language – a very large corpus based on standard English (VUA), or a medium-sized corpus based on EFL data (TOEFL). The evaluation described above was also applied here; again we used the hyperparameters from the paper (Su et al., 2020) with seed = 12. Whereas for the previous experiment we focused on TLMs only, here we present the results for TLMs only versus all metaphorical expressions (TLMs plus CMs) in Tables 8 and Table 9.

In terms of F1, the best performance for both test datasets (MEWS_Teacher and MEWS_Ads) and for both TLMs and TLMs+CMs was seen for the model trained on TOEFL. Across both prompts as well as across TLMs and TLMs+CMs, precision is higher than recall when training on MEWSMET. The results are generally higher for TLMs+CMs than for TLMs only.

4.2.3 Experiment 3: Combining TOEFL with Target Data

As shown in Section 4.2.2, large amounts of training data alone do not lead to better results on MEWSMETS; in-domain training data seems to be necessary.[7] As our dataset is too small for the model to learn, we next use a combination of our data (MEWS_Teacher) in combination with the larger TOEFL corpus as training data. We are mainly interested in detecting TLMs, so for MEWS_Teacher we only considered TLMs as metaphors. The results are reported in Table 10. In terms of F1, DeepMet trained on both TOEFL and MEWS_Teacher achieves the best results of all models for both TLMs and TLMs+CMs.

4.3 Discussion

The results of our experiments yield five main insights. Firstly, large amounts of training data are vital for DeepMet to perform well. As is shown in Table 6, DeepMet performs much better when training and testing on VUA or TOEFL data than on MEWSMET. Here, the training datasets for the VUA and TOEFL experiments are much larger than for MEWSMET experiment (see Figure 4). When reducing the amount of VUA training data to match the size of the MEWSMET corpora, DeepMet fails at the classification task for the VUA test set (F1 = 0.4).

The second insight is that in-domain training data is needed. When we increased the training data by using VUA and TOEFL (see Tables 8 and 9), and tested on MEWSMET, the model trained on TOEFL-data outperformed the model trained on VUA-data. This behaviour was seen across prompts and for both TLMs and TLMs+CMs. This shows that large amounts of training data are needed only to an extent; after a certain threshold (that has to be determined in future work), in-domain data becomes more important than more training data. The importance of in-domain data was also highlighted by the fact that the best performance overall was seen when training on TOEFL and MEWS_Teacher, and testing on MEWS_Ads.

Thirdly, it became clear that the results for detecting TLMs+CMs are generally higher than for detecting TLMs only (see Tables 8, 9 and 10). This means that the models are better at detect-

[7] By in-domain we mean language that EFL learners used in argumentative essays for various prompts.

	TLMs only			TLMs + CMs		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
TOEFL	12.0	80.7	**20.8**	17.2	80.5	**28.4**
VUA	8.8	92.7	16.1	12.8	93.3	22.5

Table 8: Performance of DeepMet fine-tuned on TOEFL and VUA, and evaluated on the split of our dataset that is based on the prompt ***TV-Ads***.

	TLMs only			TLMs + CMs		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
TOEFL	14.6	86.4	**24.9**	23.1	88.1	**36.6**
VUA	10.5	90.3	18.7	16.5	91.9	28.0

Table 9: Performance of DeepMet fine-tuned on TOEFL and VUA, and evaluated on the split of our dataset that is based on the prompt ***Teacher***.

	TLMs only			TLMs + CMs		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
TOEFL+MEWS_T	18.7	63.7	28.9	27.4	64.8	38.5

Table 10: Performance of DeepMet fine-tuned on TOEFL-data plus MEWS_Teacher, and evaluated on the split of our dataset that is based on the prompt ***TV-Ads***.

ing metaphors that are comprehensible, but that a native speaker would not use. This, however, is problematic, since TLMs can be an indicator of language proficiency, while CMs apparently cannot. If metaphors were to be used as features in automatic essay scoring, an additional module would be needed that extracts TLMs.

Our fourth insight is that the model tends to overidentify metaphors, which can be seen by the high recall and low precision across all experiments that were carried out with a sufficient amount of training data. One explanation for this behaviour is that the percentage of metaphorical expressions in MEWSMET is lower than in VUA and TOEFL training data (MEWSMET: 2% and 3%, VUA: 30%, TOEFL: 4%, see Figure 4). Also, the preference parameter α, originally designed to improve recall, has to be fine-tuned to MEWSMET data (we used a value of 0.2 as suggested by Su et al., 2020).

Lastly, it should be mentioned that a more reliable metaphor detection method has to be found, as our best model (trained on TOEFL and MEWS_Teacher, see Table 10) shows a rather weak F1-score of 28.9 for detecting TLMs only.[8]

5 Error Analysis

In order to get a clearer picture on why DeepMet performs rather poorly on MEWSMET data, we performed an error analysis. For this we used the best-performing model – DeepMet trained on TOEFL-data plus MEWS_Teacher – and looked at the predictions it made for MEWS_Ads, taking into account both TLMs and CMs. The first thing we noticed is that many differences between the annotations and the predictions concerned verbs where the concrete meaning is not the basic meaning (anymore). These verbs include *to direct*, *to confront*, *to support*, *to create*, *to target*, or *to manipulate*. For instance, the four example sentences given for the first listed meaning of *to direct* in the Longman Dictionary[9] are as follows:

(1) The machine directs an X-ray beam at the patient's body.

[8] In addition to the results presented above, we used the metaphor detection model by Ma et al. (2021), because it is in theory able to make reliable predictions with as little as 200 training instances, as has been shown by Hülsing and Schulte Im Walde (2024) in a multilingual setup. However, the results we received for our MEWS-data were very poor (F1 $<$ 14.4), which indicates that the model works well for standard language, but not for learner language.

[9] https://www.ldoceonline.com/dictionary/direct, date of access: 15.08.2024

(2) The new route directs lorries away from the town centre.

(3) I'd like to direct your attention to paragraph four.

(4) I want to direct my efforts more towards my own projects.

As none of these meanings entails sensual perception, the basic meaning is abstract, even though there might be instances where the word is used in a concrete way, e.g. *to direct the fire extinguisher at something*. In our guidelines based on MIP (Pragglejaz Group, 2007) we state that for metaphorical usage the meaning of a word in context "tend(s) to differ from the basic meaning", and we ask the annotators to compare the meaning in a given context to the basic meaning, i.e. the first meaning mentioned in the Longman dictionary (see guidelines in Appendix A.1.1). Therefore, the meaning of *to direct* in a context such as *advertising directed toward young children*[10] does not "differ from the basic meaning" and is labelled as being literal, even though our model labels it as being metaphorical. This might be due to the fact that the majority of data used for fine-tuning stems from the TOEFL-data where the annotation is not based on MIP (Pragglejaz Group, 2007), but rather based on the annotators' intuitions (Beigman Klebanov et al., 2018). The following sentences in the dataset by (Beigman Klebanov et al., 2018) contain the verb *to direct*, and two out of three labels are metaphorical[11]:

(5) At a first sight, it can be inferred that young people [...] seem to have become more ego-directed, in order to prevent themselves from the duties that a society is asking them. → literal

(6) it is the nature of the humen, but this intrest need to be directed in the right way but unfortunetlly the same can be directed by some people whom not civilized. → metaphorical (both)

This indicates that the different guidelines account for differences in classification.

A second source of differences between annotations and predictions are personifications. In line with conceptual metaphor theory (Lakoff and Johnson, 1980b), we explicitly consider personifications as metaphors (cf. Appendix A.1.2). Therefore, all of the following expressions in our MEWS data were annotated as being used metaphorically:

(7) [...] parents think that advertise threatens their child [...]

(8) If a advertise is made well it teaches the child something [...]

(9) [...] I saw an advertisement, which was directly telling children that they should go to a certain water park [...]

However, the model predicted them not to be metaphors, which is probably again due to the different annotation guidelines used for the training and the testing data. As the guidelines by Beigman Klebanov et al. (2018) are based on intuition, personifications are not specifically mentioned, so it can be assumed that the annotators did not consider them metaphors. The fact that the verb *entertains* was labelled as being used literally in following sentence from the TOEFL-data confirms this assumption:

(10) [...] the computer graphic which entertains many people in films or TVs can not invented without computer.

Thirdly, highly conventionalized expressions, such as *to raise a question* or *to come to the conclusion* were annotated as being used metaphorically and predicted as being used literally. Even though neither of these expression could be found in the training data by Beigman Klebanov et al. (2018), the following sentence was found where the word *to raise* is used similarly:

(11) And even though their usage has raised certain environmental concerns [...]

Again, *raised* is not annotated as being used metaphorically in the TOEFL-data, probably because it is too conventionalized and did not "help the author advance her argument" (Beigman Klebanov et al., 2018).

These three reasons for misclassifications hint at the need for training data that was created with the help of comparable annotation guidelines.

[10] It should be noted that the word *directed* is used in the prompt "TV-Ads" and should therefore be excluded when analyzing the correlation between proficiency level and the number of metaphors per essay.

[11] Only the two instances labelled as metaphorical are true verbs, the other one being a deverbal adjective.

6 Conclusion

In our study we set out to investigate the relationship between metaphors and essay scores. We found that EFL learners create new conceptual mappings, which are perfectly comprehensible for native speakers in spite of being uncommon (comprehensible metaphors, CMs). However, this strategy – which is absolutely serviceable in everyday life – does not give us any insights into the proficiency level of a learner, as our results suggest. Rather, language proficiency seems to correlate only with the use of metaphors that a native speaker would use (target language metaphors, TLMs).

If we want to use the number of metaphors in an essay as a feature for automatic essay scoring, we need to detect metaphors automatically. Previous studies have shown that metaphor detection methods such as DeepMet (Su et al., 2020) perform well on EFL learner data by Beigman Klebanov et al. (2018). However, such methods had not been extensively validated for younger and less proficient learners as present in our data. We showed that large amounts of training data are necessary to train a model that learns to detect metaphors in MEWSMET, however, standard English data is not useful, but new in-domain data is needed to achieve decent model performance. Here, training and testing data should ideally be annotated under the same annotation guidelines, as our error analysis revealed.

We also showed that DeepMet tends to be better at classifying CLs than TLMs. This poses a challenge, since only the number of TLMs per essay positively correlates with language proficiency. What is needed, therefore, is a method that reliably differentiates between TLMs and CMs, if we want to use the number of metaphors as features for essay scoring.

7 Outlook

Our MEWSMET dataset allows further analyses: First of all, differentiating between TLMs and CMs is vital. Pedinotti et al. (2021) use a dataset consisting of conventional metaphors and creative metaphors. They matched each of these metaphors with their literal counterpart and a nonsensical expression of the same syntactic structure. They used the pseudo-log-likelihood score (PLL) by Wang and Cho (2019) to measure the degree of plausibility that BERT attributes to a sentence. In doing so, they show that BERT is able to discern creative metaphors from nonsense expressions. As future work, we will apply this score to see whether it can also discern TLMs from CMs.

What has not been taken into account yet is the degree of conventionalization. Although our annotators assigned the labels "creative" and "conventional" to all metaphorical instances that they believed to be acceptable English (for all others they assigned the label "uncommon translation of a German conventionalized metaphor" or "non-metaphorical"), these labels should be confirmed by native speakers, or checked against a corpus-based dictionary as is commonly done to detect creative metaphors (Reimann and Scheffler, 2024b)[12]. Scores indicating novelty could weigh the metaphorical labels; after all, metaphors that the learner has frequently heard or even learnt from a textbook should be treated differently than creative metaphors that learners form themselves, when looking at the correlation with essay scores.

Also, the proximity to German metaphors should be taken into account when using metaphors as features for essay scoring. In this study, we annotated metaphors that are uncommon because they are translated from a German conventional metaphor (e.g. *to build an opinion*). We did not carry out further analyses on these uncommon translations due to their small number; only 29 expressions were annotated as being uncommon by both annotators, and two were considered incomprehensible during the check by the native speakers. However, there are probably many metaphors that originate from a parallel between the source and the target language, some that are incomprehensible but certainly others which are CMs or even TLMs. One example is *das bringt uns zum nächsten Punkt*, which can be translated word for word into *this brings us to the next point*. A learner's proficiency can be more clearly predicted when they use metaphors that do not run in parallel to German metaphors, for example, when *eine Meinung bilden* is translated into *to form an opinion* instead of *to build an opinion*.

8 Acknowledgements

We thank Stefan Keller and Flavio Lötscher (PH Zürich) for their help in refining the annotation guidelines. We are also grateful for the annotations provided by Rachel Castleberg, Rosalind

[12] We use the terms novel and creative interchangeably.

Isaacs, Jordana Plagge, and Charlotte Ventura. Anna Hülsing is supported by the German Federal Ministry of Education and Research (grant no. FKZ 01JA23S03C). Andrea Horbach's work was partially conducted at "CATALPA - Center of Advanced Technology for Assisted Learning and Predictive Analytics" of the FernUniversität in Hagen, Germany.

9 Ethics Statement

Our annotators were paid according to German minimum wage regulations.

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.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489, Dubrovnik, Croatia. Association for Computational Linguistics.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

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.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. 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.

Daniel Blanchard, Joel R. Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013:15.

Lynne Cameron. 2003. *Metaphor in educational discourse*. Advances in Applied Linguistics. Continuum, London, UK.

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.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, California. Association for Computational Linguistics.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

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.

Arthur Graesser and Danielle McNamara. 2012. Automated analysis of essays and open-ended verbal responses. In H. Cooper, editor, *APA handbook of research methods in psychology*, volume 1, pages 307–325. American Psychological Association.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930,

Copenhagen, Denmark. Association for Computational Linguistics.

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.

Stefan Keller. 2016. Measuring Writing at Secondary Level (MEWS). Eine binationale Studie. *Babylonia*.

Stefan D. Keller, Johanna Fleckenstein, Maleika Krüger, Olaf Köller, and André A. Rupp. 2020. English writing skills of students in upper secondary education: Results from an empirical study in switzerland and germany. *Journal of Second Language Writing*, 48:100700.

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. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

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.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff. 1987. *The death of dead metaphor*, volume 2 of *Metaphor and Symbolic Activity*. De Gruyter Mouton, Berlin, Boston. 2010.

George Lakoff and Mark Johnson. 1980a. Conceptual metaphor in everyday language. *Journal of Philosophy*, 77(8):453–486.

George Lakoff and Mark Johnson. 1980b. *Metaphors we Live by*. University of Chicago Press, Chicago.

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.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2022. The secret of metaphor on expressing stronger emotion. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 39–43, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2021. Improvements and extensions on metaphor detection. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 33–42, Online. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Ana Mouraz, Ana Vale, and Raquel Rodrigues. 2013. The use of metaphors in the processes of teaching and learning in higher education. *International Online Journal of Educational Sciences*, 5:99–110.

Kai Niebert and Harald Gropengiesser. 2012. Understanding and communicating climate change in metaphors. *Environmental Education Research - ENVIRON EDUC RES*, 19:1–21.

Rebecca Oxford, Stephen Tomlinson, Ana Barcelos, Cassandra Harrington, Roberta Lavine, Amany Saleh, and Ana Longhini. 1998. Clashing metaphors about classroom teachers: Toward a systematic typology for the language teaching field. *System*, 26:3–50.

Natalie Parde and Rodney Nielsen. 2018. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? Testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prisca Piccirilli and Sabine Schulte im Walde. 2022. Features of perceived metaphoricity on the discourse level: Abstractness and emotionality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5261–5273, Marseille, France. European Language Resources Association.

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

Sebastian Reimann and Tatjana Scheffler. 2024a. 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. 2024b. When is a metaphor actually novel? annotating metaphor novelty in the context of automatic metaphor detection. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 87–97, St. Julians, Malta. Association for Computational Linguistics.

Eva Samaniego Fernández, María Sol Velasco Sacristán, and Pedro Antonio Fuertes Olivera. 2005. *Translations we live by: The impact of metaphor translation on target systems*, pages 61–82. Secretariado de Publicaciones, Valladolid.

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.

Jakob Schuster and Katja Markert. 2023. Nutcracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna A Kaal, Tina Krennmayr, Tryntje Pasma, et al. 2010. *A method for linguistic metaphor identification*. John Benjamins Publishing Company Amsterdam.

Egon Stemle and Alexander Onysko. 2018. Using language learner data for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 133–138, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Oseremen Uduehi and Razvan Bunescu. 2024. An expectation-realization model for metaphor detection. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 79–84, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Rodrigo Wilkens, Daiane Seibert, Xiaoou Wang, and Thomas François. 2022. MWE for essay scoring English as a foreign language. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 62–69, Marseille, France. European Language Resources Association.

Yiqin Yang, Li Xia, and Qianchuan Zhao. 2019. An automated grader for chinese essay combining shallow and deep semantic attributes. *IEEE Access*, 7:176306–176316.

Dongyu Zhang, Nan Shi, Ciyuan Peng, Abdul Aziz, Wenhong Zhao, and Feng Xia. 2021. MAM: A metaphor-based approach for mental illness detection. In *Computational Science – ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part III*, page 570–583, Berlin, Heidelberg. Springer-Verlag.

A Appendix

A.1 Annotation Guideline

A.1.1 Main Guideline

Look at each essay individually. For each essay perform the following steps:

1. Read each sentence and pay attention to the target verbs, which are already tagged.

2. The label "Metaphorical Usage" should be given to a target verb if you believe that this word is used metaphorically. Add the label "Metaphorical Usage" where missing. The following label descriptions (taken from Mohammad et al., 2016) should help you:

   - Literal usages tend to be more basic, and have a more straightforward meaning; they are more physical and more closely tied to our senses (vision, hearing, touching, tasting).
     Example 1: The enemy shot down our aircraft.
     → non-metaphorical verb usage, no labelling necessary
   - Metaphorical usages tend to differ from the basic meaning and tend to be more complex and more distant from our senses. They often are more abstract, vague, and surprising. Also, they tend to bring in imagery from a different domain.
     Example 2: He shot down the student's proposal.
     → label: "metaphorical verb usage"

   At the end of step 2, all metaphorically used verbs should have two labels ("Target Verb" and "Metaphorical Usage").

3. Assign one of the following labels to each target verb that you labelled as being metaphorical:

   - Label "Conventionalized Metaphor": If, in your opinion, the verb represents a conventionalized metaphor, you recognize it to often be used together with one or more of the given context words.
     Example: Susan often spends her time at the swimming pool.
     → The word *spend* is often used together with the word *time*.
   - Label "Creative Metaphor": If, in your opinion, the verb represents a creative metaphor, you do not recognize the verb being usually used together with one or more of the given context words.
     Example: The present sews together the past and the future.
     → The word *sew* is usually not used together with words such as *present* or *past*.
   - Label "Uncommen Translation Conventionalized": If the verb represents an uncommon translation of a German conventionalized metaphor, you recognize a German conventionalized metaphor as the basis for the translation, but you think that the English translation is not common.
     Example: *eine Meinung bilden*, student translation: *to build an opinion*.
     NB: This label should only be given if you believe that the underlying German expression contains a conventionalized metaphor **and** if the resulting English phrase is uncommon or unidiomatic. It should not be given if the English phrase is unidiomatic or uncommon, but no German metaphor is the source for the error. This label should only be given in clear cases such as the afore mentioned phrase *to build an opinion*.

   At the end of step 3, all metaphorically used verbs should have three labels ("Target Verb", "Metaphorical Usage" and one of the following labels: "Conventionalized Metaphor", "Creative Metaphor", "Uncommon Translation Conventionalized").

A.1.2 Additional Notes

As we are dealing with authentic, and therefore noisy text, there will be expressions where the metaphoricity of a verb is unclear. In order to

clarify which words should be tagged as being metaphorical and which should not, the following examples are given as anchors for the annotation.

- Words such as *to direct* or *to confront* should not be tagged as being metaphorical. These words can have a straightforward, more physical meaning (for example *to direct the extinguisher at the fire*), but this is currently not the basic meaning, as these words are in the vast majority of occurences used in an abstract way. Therefore, here the abstract meaning is the basic meaning.

- Very frequent verbs such as *have/be/do/make...* have not been tagged as "Target Verbs" in the annotation documents, because they are mostly used as auxiliary verbs. In cases where these words occur as full verbs (e.g. *have a conversation*), the metaphorical meaning is determined mainly by the following noun, while the verb carries no or little meaning (cf. light verb phrases). As we are establishing the metaphoricity of the verbs, it is fair to say that these verbs carry no metaphorical meaning, and are therefore excluded.

- Target verbs in expressions such as *to spend time* or *to cover topics* should be tagged as being metaphorical. The expressions are highly conventionalized, but – as opposed to light verb phrases such as *have a conversation* – the meaning of the expression does not only rest on the noun, and therefore the verb carries some of the metaphorical weight.

- Idioms can be metaphors, too. For example, the verb *break* is used metaphorically in the expression *to break the ice* and should be tagged as being a metaphor. However, there are many idioms which do not have a metaphorical origin (*break a leg*, *talk to Huey on the big white telephone*) or where the origin is unclear (*it's raining cats and dogs*). These should not obtain the label "Metaphorical Usage".

- Phrasal verbs should be tagged as being metaphorical only if the basic meaning of the entire phrasal verb usually is more straightforward/physical/... (see example above: *to shoot down*). They should not be tagged as being metaphorical if only the base verb usually is more straightforward/physical/.... Example: *to miss out* should not be tagged as being metaphorical, even though the basic meaning of the base verb (*to miss*) might be more straightforward/physical/....

- Personifications should be annotated as metaphors, too. Example: *Money rules the world.*

- If a verb is used as part of an extended metaphor, it should be tagged as being used metaphorically. Example: *His head was a dovecote, most thoughts flew out, only some stayed inside.* Here, the target words should be marked as being used metaphorically.

- If you are unsure what the basic meaning of a verb is, consult the online version of the Longman Dictionary: `https://www.ldoceonline.com/dictionary/`. Be aware that there are homonyms, so there might be more than one basic meaning of a verb (for example: *to lie* can refer to the position of a person or to a person not telling the truth).

- Dead metaphors, i.e. metaphors that do not exist anymore because the mapping from source to target domain can no longer be understood without historical knowledge (compare Lakoff, 1987), should not be tagged as being metaphorical. Examples: *footage, pedigree*.

B Stop Words

In addition to the commonly used stop words (*be*, *do*, *should*, *can*, *have*, *would*) we also excluded the word *make*, because it is very often used by students as a placeholder for a verb they do not know, for example: *because school makes our future* or *make good grades*.