Cross-Lingual and Cross-Cultural Transfer of Talk Move Classification to German Science Classrooms

Christian Wartena[1], Christian Schumburg[2], Andreas Nehring[2], Marcel Ebert[3], Friederike Korneck[3], David Schmitt[4], Marie Irmer[4], Birgit Neuhaus[4]

[1]Hochschule Hannover - Institute for Applied Data Science
[2]Leibniz Universität Hannover - Institit für Didaktik der Naturwissenschaften
[3]Goethe Universität Frankfurt - Institut für Didaktik der Physik
[4]Ludwig-Maximilians-Universität München - Lehrstuhl für Didaktik der Biologie
christian.wartena@hs-hannover.de, nehring@idn.uni-hannover.de,
korneck@em.uni-frankfurt.de, birgit.neuhaus@lrz.uni-muenchen.de

Abstract

Talk moves are discourse categories used to analyse classroom conversations. They help to understand the types of interaction between teachers and students, and can be used to provide information on teaching quality and give teachers feedback. Automatic classification of talk moves is therefore useful and has recently been studied by some research groups, almost all of which use English data only. We have built a small corpus of German science classroom transcripts and investigated whether it is possible to train a model with a very limited amount of data or to train a multilingual model with English data. Our results show that both is possible but that training with a small German training corpus yields better results than training on a large English corpus. Slightly better results are obtained when both English and German data are used for training.

Keywords: Educational NLP, Cross-Lingual Transfer, Classroom Discourse Analysis, Talk Move Detection

1. Introduction

Codings (classifications of utterances based upon education science frameworks, similar to 'annotations' in the field of computer science) of spoken language in classrooms are a useful approaches in research on instructional quality. The data based upon codings are either used to provide teachers with feedback on their own teaching, or to do quantitative research on what criteria of instructional quality are more or less effective in school – and contribute to learning outcomes such as knowledge, competencies or interest and motivation (Dorfner et al., 2017).

Recent developments in artificial intelligence have led to a focus in research projects on using large language models (LLMs) to code classroom discourse based on the talk moves framework (e.g. Suresh et al., 2021; Jacobs et al., 2022, 2025). Automatic detection and classification of talk moves has primarily been conducted on English data in the fields of mathematics and literacy (Ganti et al., 2025). Consequently, training data to develop classifiers are only available for English and the aforementioned fields. Given the costs of manual annotation, we can either train with a limited amount of German training data or use the English data to train a multilingual model. It is unclear (1) if English data can be used to train a multilingual model and analyze German data and which specific challenges arise when transferring the model from one language to another, (2) if a model that is build up in mathematics can be transferred to natural sciences (biology, chemistry, physics) and if there are there subject-specific differences between biology, chemistry and physics, (3) if classroom cultures in the U.S. and Germany are similar enough for LLMs to be transferred from one country to another.

In the present paper we introduce a small corpus of German transcripts from science classes (biology, chemistry, physics) and investigate whether this corpus can be annotated according to the annotation scheme from Suresh et al. (2021) developed for mathematics classes in English, and whether we can train a multilingual model on the English data and apply it to our German corpus. We find that transfer is possible and that the classifiers give reasonable results. However, classifiers trained on a limited amount of German training data, clearly outperform the multilingual models. Using both English and German training data is possible as well: we first finetune a multilingual model on the large English dataset and subsequently on the small German dataset. This results in classifiers giving slightly but not convincingly better results than the classifiers trained on the German data only.

The remainder of this paper is organized as follows: Section 2 looks at talk moves and their importance for educational research. In Section 3 we review existing approaches to classify talk moves automatically. Section 4 introduces the English data used for training and describes the new German data, as well as the challenges that arise when annotating these data according to an annotation scheme developed for English. In Section 5 we describe the classifiers that we trained, the results of which finally are given in Section 6.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1989–1997
11–16 May 2026.

Table 1: Examples of Talk Moves. All examples are translations of utterances from our German corpus.

Talk Move	Example
	Teacher
0 Cannot be assigned to any Talk Move	
1 Keeping everyone together	T: Have a quick chat with your partner about this.
2 Getting student to relate	T: Are there any counter-suggestions to this question from Lilli?
3 Restating	S: I think they both fall to the ground at roughly the same speed. T: They both fall to the ground at the same speed.
4 Revoicing	S: Charged particles. T: Charged particles that we have defined as ions.
5 Press for accuracy	T: Describe what else you can see around the hair!
6 Press for reasoning	T: Explains what function the sebaceous glands could have?
	Student
0 Cannot be assigned to any Talk Move	
1 Relating to another student	S1: I think if it was a cold day before, then 15 °C is already very warm. S2: No, 15 °C is cold
2 Asking for more information	S: Does an electron have little or no mass?
3 Making a claim	S: The noble gas configuration states that each atom strives for a fully occupied outer shell with 8 outer electrons.
4 Providing evidence	S: That's why it has to go a little to the left and I believe that this weight always wants to go down.

2. Talk Moves

Talk moves are instructional strategies used by teachers to facilitate rich and productive classroom discussions (O'Connor et al., 2015). These strategies shift the focus of conversations from teacher-centered questioning toward student-driven dialogue, aiming at promoting deeper understanding and engagement and not just replication of knowledge. Rather than following the traditional pattern of teacher initiation, student response, and teacher evaluation, talk moves encourage collaborative understanding by prompting students to share, justify, and build on each other's ideas. This approach fosters true discussions where knowledge is co-constructed, rather than simply transferred. Phrases of students and teachers can be assigned to different talk moves.

Key talk moves include urging to keep everyone together (prompting students to be active listeners and orienting students to each other), getting students to relate to another's ideas, pressing for accuracy (prompting students to make a contribution or use content knowledge), or pressing for reasoning (prompting students to explain, provide evidence, share their thinking behind a decision, or connect ideas or representations). An overview of the talk moves used in the present study along with some examples is given in Table 1. These discourse moves aim at supporting students in developing their reasoning skills and promote a supportive learning community.

Even though Talk Moves is a generic framework, the corresponding classifications must be interpreted in a discipline-specific way. For example, in mathematics, proofs can play a particular role in the category 'pressing for reasoning', whereas in the natural sciences, assertions can hardly be 'proven' (deductively), but must be verified or falsified by connecting data and theories through argumentation.

Country comparative analyses of filmed lessons have existed since the nineteen-nineties (e.g. Hiebert et al., 1999; Klieme and Schweig, 2020). The implications of these studies often point out that lessons within a country follow cultural pattern of teaching. This might impact classroom discourse. As a simple example: classrooms that are more teacher-centered might entail larger amounts of teachers talk or questions whether or student-centered lessons might lead to increased student talk moves including claims and evidences.

3. Related Work

Various recent papers investigate the automatic detection and classification of talk moves in audio transcripts. A number of papers use the large data set of transcribed math lessons from Suresh et al. (2021, 2022a). Suresh et al. (2021) show that trans-

former based classifiers outperform LSTM based classifiers for these data. Their best performing model reaches a micro average F1-score for the retrieval of talk moves of 0,79. The input for the classifier is the sentence to be classified. The immediately preceding sentence is added as context to a teacher sentence if that sentence was spoken by a student and vice versa for student sentences. In a follow up paper Suresh et al. (2022b) experiment with adding more context to the sentences that has to be classified and show that more context can help to improve the classification performance. However, none of the models give results exceeding the previously reported 0,79-F1 score.

Kupor et al. (2023) classify 5 different talk moves using a similar approach but using a separate fine tuned classifier (each trained with different parameters and using different types of context) for each talk move. The weighted average F1 score is 0.58 and 0.63 for RoBERTA and GPT3, respectively[1].

More recently zero and one shot classifiers using prompting techniques with instructional large language models have come into focus. Moreau-Pernet et al. (2024) compare fine tuning an encoder only model (Roberta) with a classification layer, a classifier using sentence embeddings from a model that was not fine tuned and prompting a fine tuned version of GPT 3.5. The prompt includes a description of the talk moves, a few examples and the transcript of the whole lesson. The model is asked to reproduce the whole transcript and add labels to each teacher utterance. This last model has the best performance with a weighted F1 score of 0.93. This result cannot be compared with the result from Suresh et al. (2021) since (1) they use different data sets and (2) Moreau-Pernet et al. (2024) include the class 'No talk move' in their average, while Suresh et al. (2021) excludes this class. The weighted (not micro!) average of the fine tuned prompting approach seems to be roughly 0.81.

Wang et al. (2025) also compare fine tuning and prompting approaches but use the dataset from Suresh et al. (2021). While they do not get good results from prompting, their best results are obtained by using parameter efficient fine tuning with LoRa techniques of large pretrained models. They report a F1-score of 0.86 using Llama-3-2-3B. Again, this score cannot be compared directly to the scores from Suresh et al. (2021) since here, like in the work of Moreau-Pernet et al. (2024) macro-average is used and the none-class is included in the average (Wang, 2025).

Ganti et al. (2025) annotate four datasets from different educational settings and form different domains. They find that most models can be transferred quite well from one dataset to another, i.e. a classifier trained on one dataset can be used to classify sentences from one of the other datasets. They found that the performance declines when the distance between the datasets increases. Furthermore, they found that the fined tuned models they tested are less generalizable than the zero and few shot approaches. Besides this, this study confirms the finding form other studies that adding more context in general improves the performance.

The goals of the last study are very similar to our goals. However, the distance between the training and test data is much larger than that between the datasets used by Ganti et al. (2025): we move not only to another discipline, but also to another country, another language and to codes assigned by a different team of annotators.

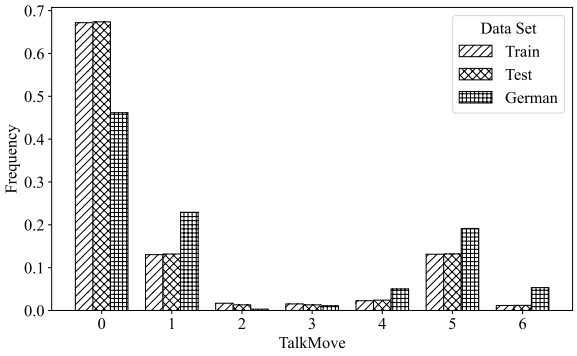

Figure 1: Relative frequencies of teacher talk moves in the original train and test data and in the new German dataset.

4. Data

For our study we built and annotated a small corpus of German transcripts. In addition we use English data to train a multilingual model.

4.1. English Data

For training and for validation of our implementation we use the test and training data from Suresh et al. (2021), that are available on GitHub[2]. These data consist of 501 transcribed and annotated math lessons, resulting in 235.926 annotated sentences (174.567 teacher sentences and 60.359 student sentences), carefully divided in a training and a test set. The training set consists of 151,264 teacher and 53,079 student sentences. The test set has 23,303 teacher and 7,262 student sentences. The distribution of the different talk moves in test and training data is shown in Figures 1 and 2.

[1]Average scores are not reported in the paper but computed by us.

[2]https://github.com/SumnerLab/TalkMoves

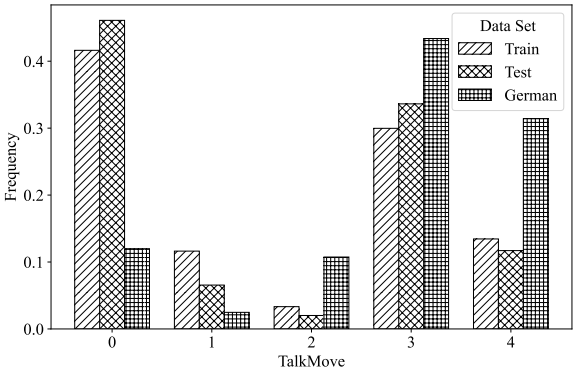

Figure 2: Relative frequencies of student talk moves in the original train and test data and in the new German dataset.

4.2. Collecting and Annotating German Data

In order to annotate German classroom transcripts from physics, chemistry and biology the coding manual that was originally designed in English for the subject of mathematics, had to be translated into German and adapted to the new subjects. This process required careful consideration to ensure appropriate transfer across languages. The transfer to the subject areas of biology, chemistry and physics posed a further hindrance, for example the evidence in the natural science subjects are missing. In addition, the scientific inquiry, which plays a central role in the natural science subjects, is not taken into account in the original coding manual.

For these reasons, the coding manual was iteratively concretized until an acceptable agreement (Fleiss'-$\kappa > 0.6$) was reached between three annotators. In five cycles, sentences from all three subjects were coded by all three annotators. In case of discrepancies, the coding manual was adjusted accordingly. Finally, the Fleiss' kappa was determined for 150 sentences between three annotators per subject and can be seen in Table 2.

Classroom transcripts from previous research projects were used as part of the analysis. The data basis includes both real (chemistry, physics) and scripted (biology) teaching units that represent different school contexts. The transcripts were selected on the premise of broad variation in terms of content and didactics, as different sub-sequences were selected. A total of approximately 900 sentences, 300 per subject, were used to build a test dataset. Of these, 150 sentences each came from the triple coding.

Remarkably, the distribution over the classes in the German test dataset differs substantially from the distribution in the English data (see Figures 1 and 2). For the teacher talk moves we see that the 2 (Getting student to relate) and 3 (restating) are extremely rare in the German data. For the student talk moves class 1 is in German less frequent than in English. The most striking difference however is, that the class 0 (not a talk move) makes up 42% of the English data but just 12% of the German data. These differences might reflect a completely different classroom situations in the recorded sessions, but partially also can be caused by the small size of the German dataset, or it can reflect a different understanding of the talk moves between the annotators from both datasets.

Table 2: Agreement between the three annotators on the three German subset using Fleiss-κ.

	Teacher	Student
Physics	0.75	0.92
Chemistry	0.80	0.73
Biology	0.80	0.75

In addition to this balanced test set of 900 sentences, 3,526 sentences were annotated from physics classroom transcripts and 1,188 and for chemistry classroom transcripts. We will use this data as training data, although it is approximately a factor of 50 smaller than that for English. The exact numbers for each subset is give in Figure 3.

5. Method

For training the models we used the methods described in Suresh et al. (2021), since the results of this approach are the best reported on this dataset up to now.

In this approach we use separate models to classify teacher and student sentences. The classifier gets sentence pairs as input. The second sentence is the sentence that has to be classified. For the teacher model the first sentence is the last student sentence if that immediately preceded the teacher sentence and an empty string otherwise. For the student model the first sentence is the last teacher sentence or the empty string if the last teacher sentence was not immediately preceding.

We did not use the implementation provided by the authors of the original paper on GitHub but used our own implementation using some parts of the original implementation, especially the evaluation functions. We decided to do less preprocessing and basically giving the original text to the transformer model without changing case and removing punctuation and non-ASCII characters (Actually, Suresh et al. (2022b) already suggest that removing these preprocessing steps might improve the results). We see below that this gives slightly better results.

Since we have two possible datasets to train on (both for the student and teacher utterances), we

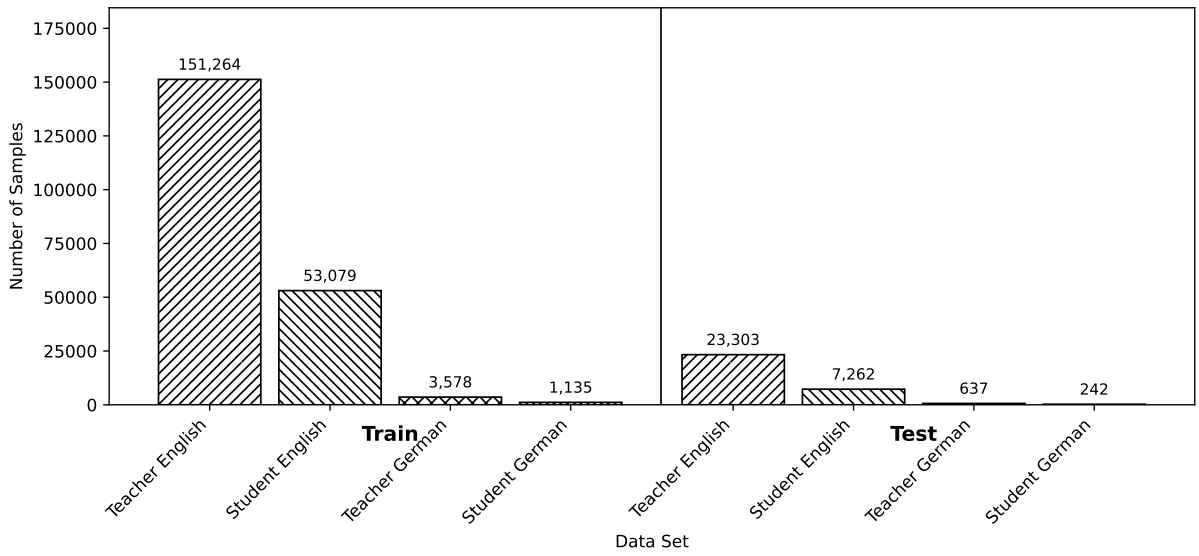

Figure 3: Data sizes by split, language, and role.

consider three different training possibilities: 1) finetuning a model on English data only, 2) finetuning a model on German data only and 3) finetuning a model on German data that was finetuned before on English data.

We used three pretrained multilingual language models: BERT multilingual base model (cased) (Devlin et al., 2018), (multilingual) DeBERTaV3 (He et al., 2021b,a), and XLM-RoBERTa (Conneau et al., 2020). We also used the monolingual English Bert base model and the GBERT base and large models (Chan et al., 2020) to see the difference of performance between the mono- and multilingual model. We used only base models (except for GBERT large, that has the same size as the multilingual base models) as we did not have the computational resources to use the large variants.

We used the standard models for training sequence classification from Hugging Face. Given the small amount of German training and test data (see Figure 3) we decided not to use a part of the data as validation data for parameter optimization and early stopping. Instead we used the same parameter settings for all models and data sets. All parameter settings used are the same as those used by (Suresh et al., 2021). When training with one dataset we train 4 epochs (as Suresh et al. (2021) did). In the cases where we continue training with a second smaller dataset we use 5 epochs. We also tried using class weights: since we are only interested in the classification of talk moves and the majority of the utterances is not a talk move at all, we gave weight 1 to all talk moves and weight 0.2 to the class of sentences that do no represent a talk move. We also experimented with LoRA finetuning (Hu et al., 2021), which updates a small number of low-rank adapter weights instead of the full model parameters. Usually LoRa finetuning is used for saving computation time, but e.g. Wang et al. (2025) also reports better performance for classification tasks. We use the implementation of LoRA from Huggingface (`https://huggingface.co/PEFT` with standard parameters (rank of matrices is set to 8, alpha is set to 16 and dropout to 0.1). This reduces the models from about 279 million trainable parameters to 0,89 million parameters.

6. Results

For the evaluation we follow Suresh et al. (2021) as well and report micro-averaged F1 score and Mathews Correlation Coefficient (MCC) for the classes representing a talk move and excluding the class 0 in all averages.

Table 3 gives the result for the teacher models. We see a few remarkable things. In the first place our reproduction of the results of Suresh et al. (2021) is better than the results reported in their paper. The table gives only the best result they reported, but we also get better results if we use the same pretrained model, e.g. XLM Roberta Base. The results on the same data by Moreau-Pernet et al. (2024) and Wang et al. (2025) probably are better than the results reported here, but as mentioned before, they use a different evaluation method and do not give MCC or micro averaged F1 for the talk moves classes. However, our goal is not to improve state of the art results, but to reproduce these results and transfer the models to German.

We see that the models trained on English data clearly are able to classify German teacher utterances as well, although performance on the German data remains below that on the English data. Interestingly, adding class weights and LoRA fine

Table 3: Results for the Teacher Model on the original English dataset and the new German dataset. The first row gives the results reported by Suresh et al. (2021) using Roberta Large. The next two rows give our reproduction of their results using various language models and our own implementation of the method described by Suresh et al. (2021).

	English		German	
Model	F1	MCC	F1	MCC
Roberta Large (reported)	0.7933	0.7779		
BERT Base Cased	0.8099	0.8075		
BERT ML Base Cased	0.8067	0.8056	0.4973	0.4787
mDeBERTa	**0.8173**	0.8149	0.5626	0.5420
mDeBERTa (weighted)	0.8169	**0.8153**	0.5597	0.5382
XLM Roberta Base	0.8098	0.8102	0.5616	0.5437
XLM Roberta Base (weighted)	0.8133	0.8111	0.5760	0.5489
XLM Roberta Base / LoRA	0.7529	0.7554	0.5954	0.5734
XLM Roberta Base / LoRA (weight.)	0.7511	0.7511	**0.6151**	**0.5894**

Table 4: Results for the Teacher Model on the new German dataset trained only on the small German training set.

Model	F1	MCC
GBERT base	0.5528	0.5130
GBERT base (weighted)	0.6056	0.5605
GBERT large	0.6708	0.6238
GBERT large (weighted)	0.6748	0.6322
mDeBERTa	**0.6846**	**0.6469**
mDeBERTa (weighted)	0.6790	0.6262
XLM Rob.	0.6402	0.6144
XLM Rob. (weighted)	0.6574	0.6070
XLM Rob. / LoRA	0.0780	0.0288
XLM Rob. / LoRA (weighted)	0.4327	0.2349

Table 5: Results for the Teacher Model on the new German dataset trained on the English training data and small German training set.

Model	F1	MCC
mDeBERTa	0.6624	0.6289
mDeBERTa (weighted)	0.6468	0.6018
XLM Rob.	0.6868	0.6524
XLM Rob. (weighted)	**0.6910**	**0.6624**
XLM Rob. / LoRA	0.5902	0.5672
XLM Rob. / LoRA (weighted)	0.6259	0.6027

tuning does have little effect or a negative effect on the English data. LoRa fine tuning led to extremely bad results in the case of DeBERTa and is not reported here. In the case of XLM Roberta they cause an improvement for the German data. This suggests that extensive fine tuning with English data diminishes the models multilingual capabilities.

Table 4 gives the results for the models trained on the German training data only. Remarkably, the multilingual models perform better than German models. Furthermore, we see that the results are clearly better than those obtained by training on the English data. When using both datasets the results are slightly better (see Table 5), but the differences are very small.

Tables 6 and 7 give the results form the best performing model for the three subsets. We see that the results for physics and chemistry are better than those for biology. Since the German training data did not contain data form biology classes, we might be tempted to use that as an explanation. However, when only english data from math classes are used, biology has the lowest scores as well. In fact, especially chemistry and biology get a boost from the additional finetuning with German data. The difference might just be caused by the small size of the datasets.

Table 8 gives precision, recall and F1 for each talk move class for the teacher sentences using the best performing model. We see that class 0 has a high recall and low precision while most other classes have a higher precision and lower recall. This reflects the tendency to place any doubtful case in class 0, the majority class.

As already mentioned, the majority of the errors are cases in which a sentence is not recognized as a talk move at all, i.e. it the sentence is falsely assigned to class 0. Remarkably, for the teacher sentences this error type includes many imperatives. German imperatives are very rare in written language and even in most spoken registers we do not find many imperatives, but they are very frequent in the classroom context, like in the following

example:

(1) Und denkt dran, Kräfte einzuzeichnen.
And remember to draw in forces.

Morphological the imperative plural is identical to the third person singular of the indicative. Thus it is likely that the LLMs do not understand these sentences correctly. This type of error occurs in both the models trained only on English data and those trained on German data.

Table 6: Results for the Teacher Model on the three German subsets using the XLM Roberta Base model fine with LoRA and class weights on English data only.

Data	F1	MCC
Physics	0.6448	0.6390
Chemistry	0.5668	0.4934
Biology	0.5166	0.5015

Table 7: Results for the Teacher Model on the three German subsets using the XLM Roberta Base model fine tuned with class weights on English and German data.

Data	F1	MCC
Physics	0.6831	0.6765
Chemistry	0.7568	0.6918
Biology	0.6000	0.5930

Table 8: Classification report for the Teacher Model on the German test data using the XLM Roberta Base model fine tuned with weights on English and German data.

Class	Precision	Recall	F1
0	0.79	0.93	0.80
1	0.79	0.69	0.73
2	0.00	0.00	0.00
3	0.67	0.57	0.62
4	0.85	0.53	0.65
5	0.75	0.67	0.71
6	0.70	0.41	0.52

Suresh et al. (2021) do not give results for the student model. In Tables 9 and 10 we see that these are much behind the results of the teacher model, despite the fact that there are less classes for the student sentences. Probably, the student sentences are harder because recording (and thus transcription) quality of the student speech might be worse, or because the students tend to use shorter and vaguer sentences. However, when German training data are used the gap seems to become smaller.

7. Discussion

In the present paper we used an AI-based coding of spoken language in classroom developed for math classes in the US and transferred it to spoken language in science classrooms in Germany. The results suggest that the models trained on the data in (1) English for (2) math classes in (3) the U.S.A. can indeed be applied to German transcripts of classes on different subjects. Thus the concept of talk moves seems to be present in the same way in both data sets. However, we see that using a much smaller German training dataset gives already better results. Apparently, the language transfer is hard for the models and a few data in the target language helps more than a large amount of data in another language. We would expect that models trained on both datasets profit from the large amount of the English data and the better fit of the German data. Indeed the results of this configuration have the highest correlation to the ground truth, but the differences with the models finetuned on German data only is very small.

We did not find any evidence that transferring the LLM from mathematics to natural sciences causes major problems, neither in biology, nor in chemistry or physics.

Looking into the classification errors we see that models seem to have problems analyzing German plural imperatives, that are extremely rare in written language and even in most spoken registers. Probably the LLMs see imperative forms as third person singular (that are morphologically identical) and as a consequence do not recognize the correct discourse function of these verbs.

In summary, this study provides only initial results on how to transfer English-language based LLMs to German-based models in the field of classroom analysis. Moreover, the annotations presented here (Talk Moves), are based on a low-inference coding system, that relies heavily on language. However, empirical research in the field of instructional quality of the last decades, has shown that especially annotations based on high-inference coding systems are much better indicators for instructional quality. It therefore remains as an open question, if the results presented here, are transferable to high-inference coding systems. Nevertheless, in future, the approach presented here appears to be useful for providing immediate feedback to teachers using LLMs – an aspect that was not feasible with human-driven coding due to its time consuming nature.

Though our results are encouraging, their quality might not yet be high enough to build an application

Table 9: Results for the Student Model on the original English dataset and the new German dataset.

Model	English F1	English MCC	German F1	German MCC
BERT Base Cased	0.7225	0.6659		
BERT ML Base Cased	0.7187	0.6618	0.4703	0.2701
mDeBERTa	0.7275	0.6708	0.5401	0.3625
mDeBERTa (weighted)	0.7257	0.6636	**0.5714**	**0.3864**
XLM Roberta Base	**0.7343**	**0.6801**	0.5489	0.3792
XLM Roberta Base (weighted)	0.7261	0.6610	0.5293	0.3490
XLM Roberta Base / LoRA	0.7020	0.6326	0.4957	0.3296
XLM Roberta Base / LoRA (weighted)	0.6609	0.5587	0.5278	0.3384

Table 10: Results for the Student Model on the new German dataset trained on the English training data and the small German training set.

Model	F1	MCC
XLM Rob.	0.6415	0.4896
XLM Rob. (weighted)	**0.6745**	**0.5341**
XLM Rob. / LoRA	0.5637	0.3740
XLM Rob. / LoRA (weighted)	0.5885	0.3934

that provides teachers with automatic feedback. In order to improve the results we can try to improve the classification model itself by giving more context or by using other pretrained models that have less problems with spoken German. Probably, annotating more data will give the most improvement.

Limitations

All models were trained and evaluated on a small Linux server with only one GPU. Thus the number of configurations that could be tested was limited. Using different class weights, and other models could have lead to slightly better results. However, we see that the results of all models are quite close to each other, thus we do not expect completely different results from untested configurations.

For the same reason we tested only base models. Larger models might give better results. However, in future applications, uploading the transcripts to external servers will be unwanted or even forbidden. Thus the use of base models also reflects a realistic application scenario.

Finally, all classroom recordings and transcripts were made under strict data protection rules with consent form teachers, students and their parents. These data protection rules do not allow us to distribute the transcripts (not even in anonymized form).

Ethics Statement

For this study, we used German classroom transcripts collected at schools in Bavaria, Hesse, and Lower Saxony. All transcripts were created under the current data protection rules and were anonymised.

Future recording and analysis of classroom conversations could affect teachers' working conditions. However, projects focusing on talk moves use automatic analysis with low computational resources. This would enable teachers to record and analyse their own classrooms without the need for other people to listen to the recordings or upload them to foreign servers.

8. Bibliographical References

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Tobias Dorfner, Christian Förtsch, and Birgit J. Neuhaus. 2017. Die methodische und inhaltliche Ausrichtung quantitativer Videostudien zur Unterrichtsqualität im mathematisch-naturwissenschaftlichen Unterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, 23(1):261–285.

Achyutarama R. Ganti, Steven R. Wilson, and Geoffrey Louie Wing-Yue. 2025. Cross domain classification of education talk turns. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6897–6917, Abu Dhabi, UAE. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

James Hiebert, James W. Stigler, and Alfred B. Manaster. 1999. Mathematical features of lessons in the timss video study. *ZDM*, 31(6):196–201.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Lai, and Tamara Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, 112:103631.

Jennifer Jacobs, Abhijit Suresh, Brandon M Booth, Tamara Sumner, Jeffrey Bush, Chelsea Brown, and Sidney K D'Mello. 2025. Automating feedback from recorded instructional observations: using ai to detect and support dialogic teaching. In *Research Handbook on Classroom Observation*, pages 341–365. Edward Elgar Publishing.

Eckhard Klieme and Jonathan Schweig. 2020. Opportunities to learn. Technical report, OECD, Paris.

Ashlee Kupor, Candice Morgan, and Dorottya Demszky. 2023. Measuring five accountable talk moves to improve instruction at scale.

Baptiste Moreau-Pernet, Yu Tian, Sandra Sawaya, Peter Foltz, Jie Cao, Brent Milne, and Thomas Christie. 2024. Classifying tutor discursive moves at scale in mathematics classrooms with large language models. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 361–365, New York, NY, USA. Association for Computing Machinery.

Catherine O'Connor, Sarah Michaels, and Suzanne Chapin. 2015. "Scaling Down" to explore the role of talk in learning: From district intervention to controlled classroom study. In Lauren Resnick, Christa Asterhan, and Sherice Clarke, editors, *Socializing Intelligence through Talk and Dialogue*, pages 111–126. American Educational Research Association.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022a. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Abhijit Suresh, Jennifer Jacobs, Vivian Lai, Chenhao Tan, Wayne Ward, James H. Martin, and Tamara Sumner. 2021. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022b. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

Deliang Wang. 2025. Private communication. E-Mail.

Deliang Wang, Yaqian Zheng, Jinjiang Li, and Gaowei Chen. 2025. Parameter-efficiently fine-tuning large language models for classroom dialogue analysis. *IEEE Transactions on Learning Technologies*, pages 1–15.