Lucie Flekova, Tahir Sousa, Margot Mieskes, Iryna Gurevych

Document-level school lesson quality classification based on German transcripts

Analyzing large-bodies of audiovisual information with respect to discourse-pragmatic categories is a time-consuming, manual activity, yet of growing importance in a wide variety of domains. Given the transcription of the audiovisual recordings, we propose to model the task of assigning discourse-pragmatic categories as supervised machine learning task. By analyzing the effects of a wide variety of feature classes, we can trace back the discourse-pragmatic ratings to low-level language phenomena and better understand their dependency. The major contribution of this article is thus a rich feature set to analyze the relationship between the language and the discourse-pragmatic categories assigned to an analyzed audiovisual unit. As one particular application of our methodology, we focus on modelling the quality of lessons according to a set of discourse-pragmatic dimensions. We examine multiple lesson quality dimensions relevant for educational researchers, e.g. to which extent teachers provide objective feedback, encourage cooperation and pursue thinking pathways of students. Using the transcripts of real classroom interactions recorded in Germany and Switzerland, we identify a wide range of lexical, stylistic and discourse-pragmatic phenomena, which affect the perception of lesson quality, and we interpret our findings together with the educational experts. Our results show that especially features focusing on discourse and cognitive processes are beneficial for this novel classification task, and that this task has a high potential for automated assistance.

1 Introduction

Analyzing large-bodies of audiovisual information with respect to discourse-pragmatic categories is a time-consuming, manual activity. This task is of great importance in a wide variety of domains, including education, psychology, criminal forensics, sociology, and human resources management, since the volume of audiovisual information is steadily growing. At the same time, the tools for automatic speech recognition are getting ever more mature to be applied under realistic conditions.

Given the transcription of the audiovisual recordings, we propose to model the task of assigning discourse-pragmatic categories as supervised machine learning task. The major contribution of this article is thus a rich feature set to analyze the relationship between the language and the discourse-pragmatic categories assigned to an analyzed audiovisual unit. By analyzing the effects of a wide variety of feature classes, we can

trace back the discourse-pragmatic ratings to low-level language phenomena and better understand their dependency.

As one particular application of our methodology, we focus on modelling the quality of lessons according to a set of discourse-pragmatic dimensions. Educational researchers extensively analyze the interaction between teachers and students in all age groups in order to find components of teaching effectiveness. A commonly employed method is based on videography, in which acoustical and visual elements of lessons are recorded. Researchers analyze these recordings and assess the quality of interaction both between the teacher and the students, and among the students, on a variety of levels, in order to design teacher trainings and educational interventions. This assessment is a very tedious process, as multiple trained experts have to evaluate the quality independently. During this process, each annotator has to go over the complete material several times in order to assign quality scores on the various dimensions and levels of depth relevant for educational researchers. For details on the procedure see e.g. Rakoczy (2006). Currently, this task is carried out completely manually, which does not allow for scaling such studies to a wider scope. Supporting it with automatic methods would help to reduce this impediment.

We show that our feature set can reveal important insights into the nature of language associated with the studies of quality dimensions of the lessons. These findings can be utilized in a number of ways: i) to automate the task of analyzing the quality dimensions manually, and thus scale the technology to monitor huge amounts of data, and ii) to provide feedback to educational specialists about the lessons and help to conceptualize the appropriate interventions. The feature set itself can be re-used in the future for similar tasks.

In close cooperation with educational researchers, we processed a data set of transcripts of German mathematics lessons and created multiple machine learning models to classify the lessons on a range of quality dimensions. Thanks to the manual expert annotations, provided by educational researchers, we were able to train and test each model on a highly reliable gold standard. We make both the data[1] and the software[2] available to the research community.

Our key contribution is two-fold: Firstly, we determine which aspects of the verbal behaviour expressed in the transcripts (such as sentiment polarity or discourse structure) are predictive for the ratings of lesson quality dimensions (such as constructive feedback, student cooperation and reasoning path development), and we offer interpretation of our findings. Secondly, by modeling the problem as a text classification task with NLP features, we also demonstrate that this educational research task has a high potential for automation. Our initial results show that especially features related to e.g. discourse, sentiment and cognitive processes allow for good lesson classification within the quality dimensions studied here. The data set of transcripts is publicly available. To our

[1] `https://www.ukp.tu-darmstadt.de/index.php?id=11716`

[2] `https://github.com/UKPLab/pythagoras` - Kindly keep in mind that the aim of publishing the experimental code is mainly to provide a reference for feature implementation details rather than to distribute a fully documented and tested piece of software.

knowledge, we are the first to use this type of data for an automated natural language analysis with this focus, especially in German.

The paper is organized as follows: Section 2 presents related research, mainly in the area of educational NLP, and Section 3 provides a description of the data set we used in general, while in Section 4 we describe and motivate the subset we used in the current study. In Section 5, we describe our text classification approach. Section 6 presents the results along with the suggested interpretation of our findings and their discussion. Section 7 concludes our work and addresses future research directions.

2 Related Work

As the task we present here has not yet been covered, there is little previous work in direct relation. Our work fits best into the growing field of *Educational NLP*. The series of workshops on *Innovative Uses of NLP for Building Educational Applications* give an overview over the current trends. Previously addressed tasks are largely varied, including, but not limited to, the assessment of student tests (spoken and written) (Cheng et al., 2014; Kharkwal and Muresan, 2014; Loukina et al., 2014, 2015; Farra et al., 2015; Napoles and Callison-Burch, 2015), computer-assisted translation (Ahrenberg and Tarvi, 2014) and vocabulary-constrained natural language generation (Swanson et al., 2014).

Group Cooperation as expressed in the interaction among the students is an important phenomenon in learning. Machine-learning techniques were used to automatically classify elements of group interaction in written German conversations (Rosé et al., 2008), based on manual annotation of the available data. The authors gathered the data using a chat system. Students who participated in the study had to type their conversations. The resulting dataset contained 250 conversations and a total of 1,250 German text segments. It was annotated for argumentative knowledge construction based on an elaborate coding scheme. The authors then derived a variety of features such as punctuation, token uni- and bigrams, POS bigrams and presence of non-stop words and rare words. The results indicate that phenomena of group interaction can be reliably detected based on textual information. This is important for our work, as we also only rely on textual information (see Section 5.2). Gweon et al. (2009) automatically predict student activity levels in group meetings using only average student talk time and overlap. With these basic features, they can already reliably differentiate between the students taking lead in conversation and the ones back-channeling. Others, such as Kersey et al. (2009) focus on computationally measuring the shifts of initiative as a predictor of knowledge co-construction, a high-level concept explaining the effectiveness of peer learning (Hausmann et al., 2004). In one of the tasks, the authors found a significant correlation between the post-test score and the number of shifts in dialog initiative between speakers.

Tutor/Teacher-Student Interaction and Feedback has been studied extensively, as an important mechanism in teaching (see for example Hattie and Timperley (2007) and Hattie (2009)). Studies on the impact of feedback on mistake vs. feedback

when correct found that feedback types were not predictive of post-test results (Di Eugenio et al., 2005). Other studies (Chen et al., 2011) on the correlation between dialog acts and learning gain in informatics found "several dialog act sequences that significantly correlate with learning gain"(Chen et al., 2011, p.73). Examples are a prompt followed by an instruction and feedback. However, the effective dialogue act sequences varied per topic studied in the class, not being conclusive for a generic application. Additionally, it was observed that there is a "tendency of dialog partners to adjust various features of their speech to be more similar to one another" (Ward and Litman, 2007, p.57). The authors hypothesize that the convergence towards the tutor might be associated with learning, and show that lexical overlap in consecutive utterances can discriminate well between a tutoring dialog and randomly ordered text, which we translated into features to use in our work (see Section 5.2).

Student Emotion Detection has been the focus of multiple studies aiming at automatic emotion classification under various conditions. Researchers studied the emotion of students in human-human tutoring dialogues as opposed to human-computer ones (Litman and Forbes-Riley, 2006). The authors compare results on positive, negative and neutral utterances both based on lexical and surface features from transcripts and on acoustic-prosodic features. Their findings indicate that, based on the transcripts alone, it is possible to achieve comparable emotion prediction results to using transcripts *and* recordings. This is important for our research, as the speaker emotion detection plays a role in several lesson quality dimensions we study. Other researchers demonstrated that student uncertainty negatively correlates with learning success (Forbes-Riley and Litman, 2011). In another work, the authors additionally found (Forbes-Riley et al., 2012) that student disengagement negatively affects learning success. Reasons for disengagement were found to be: Presenting a problem for too long and presenting a too hard problem. Additionally, a short time interval between the question of an automated tutoring agent and the answer of a student was a strong predictor for student disengagement.

The following two sections (3 and 4) present the data and the ratings created in the study by educational researchers. A subset of these is then used in our NLP analysis as described from section 5 onwards.

3 The Pythagoras Data set

The data used in this paper originates from a bi-national (Germany and Switzerland) study presented by Lipowsky et al. (2009), in which 40 classes of 8^{th} (Germany) and 9^{th} (Switzerland) grade students were video-taped during five of their mathematics lessons. During three of the lessons each class was introduced to the Pythagorean Theorem (*Theory*) and during two lessons each class dealt with mathematical problem solving (*ProbSol*) in general (i.e. not related to the Pythagorean Theorem). The whole study thus contains 200 videos, each lesson being 40-50 minutes long.

3.1 Manual Transcription

The videographic studies are very sensitive to privacy issues, therefore the recordings are available only with a well-founded request and under tight restrictions. To facilitate the studies and the accessibility of the material, 193 videos were manually transcribed and anonymized. These anonymized transcripts are now available to research communities of various fields. Table 1 shows a snippet of the transcript.

The transcripts include elements such as laughter, coughing and door slams. Pauses were not marked specifically, beyond using "..." for short pauses and splitting a segment[3] into two segments for the same speaker if he/she paused longer. Dialectal elements were translated to Standard German and the utterances were anonymized (e.g. Schueler #F).

The recordings were carried out in the actual classrooms, using only one head-mounted microphone for the teacher and a microphone attached to the camera. Therefore, the recording quality is occasionally quite low, which also accounts for inaudible passages. Additionally, the transcribers were asked to not transcribe a range of phenomena such as hesitations, in order to keep the transcription efforts low. Only the beginning time of each utterance was given, therefore any rhetoric pauses cannot be determined.

In addition to the transcripts, manual summaries for each lesson and fine-grained lesson segment annotations of the situations observed in the classroom were produced and made available. These represent in detail, for example, whether homework is being discussed or a proof is being shown.

Time	Speaker	Dialog
00:12:41:01	S	Wenn es um den Pythagoras geht, dann ist ja klar, dass das ()! *If this is about Pythagoras, then it is obvious that ()!*
00:12:49:28	SN	Ja, doch! *Yes, of course!*
00:12:51:00	T	Klar, SCHUELER#F., wie lautet der denn? *Sure, student#F., what is it then?*

Table 1: Snapshot of a part of a transcript. Time stamps and speakers (Teacher (T), Student (S) or New Student (SN)) are marked. The transcribed parentheses () in the first utterance indicate that part of the conversation was not audible to the transcriber and was therefore, not transcribed. Anonymized student names are indicated by "SCHUELER#F". The translations to English below each segment are our own.

3.2 Lesson Quality Assessment

In order to rate the interaction between teachers and students and among students, an annotation scheme was developed by Rakoczy and Pauli (2006). In this scheme each aspect of the interaction (we further refer to these aspects as 'dimensions') is described as a basic idea and a list of indicators such as "Students do not mock each other". For each of the dimensions defined, each lesson was rated on a 4-point Likert scale by

[3] A *segment* in our wording is a stretch of speech uttered uninterrupted by a single speaker, as shown in Table 1. Other words used in the literature for this phenomenon are *turns* or *utterances*.

2-3 expert annotators. The average score of all annotators is taken as a final lesson score in each of the dimensions. These scores are also called "high-inference ratings" in educational research literature. The annotators were encouraged to consider: frequency or duration of the specific behavior during a lesson, intensity of this behavior and distribution across students. The general impression was based on all three parameters. For all dimensions, the inter-rater reliability was assessed to ensure the quality of the annotations. The inter-rater agreement was measured using the generalization coefficient, as detailed in Clausen et al. (2003). This coefficient "expresses the relative amount of true variance in the variance observed." (Lipowsky et al., 2009, p. 532), hence accounting for chance agreement. The reliability threshold recommended by educational researchers is 0.66.

Based on these quality ratings, educational researchers performed further analysis. For example, Rakoczy (2006) looked into the correlation between classroom conditions and motivational and cognitive development of the students, but found no significant correlation. Lipowsky et al. (2009) analyzed the student performance based on a post-test. Their findings include, but are not limited to, that disturbances in the classroom correlate with the performance, whereas the feedback type does not affect it. In our work we use a reliable selection of these ratings, as explained below.

4 Dimensions Analyzed in Our Study

The lesson transcripts and the high-inference ratings presented above are the basis for our computational study. Out of the 193 only 187 transcripts (115 Theory and 72 ProbSol) could be used. The data in the remaining transcripts was corrupted beyond repair, i.e., the transcription did not follow the recommended format and could not be correctly segmented by any automated methods.

Our final data set thus contains 78,242 transcribed conversation segments from a total of 140 hours of recordings, as detailed in Table 3. Each conversation segment, further referred to as "utterance", is a set of sentences which has its own speaker and time annotation, as previously shown in Table 1. Further analysis of the data set, such as rating correlations, is provided in Section 5, in order to facilitate direct comparison with the results.

As the transcribers for the manual transcriptions also transcribed unspoken elements, such as laughter (see Section 3.1 above), we could take these phenomena into account.

But, they also had to translate dialectal elements (for example from the Swiss German dialect) to Standard German. Hence, we could not take into account the effect of dialectal phenomena on the lesson quality.

All 28 lesson quality dimensions were rated on a four-point Likert scale (see the annotation guidelines of Rakoczy and Pauli (2006) for details). Educational researchers have averaged the annotation values from raters per lesson. These averages serve as the gold-standard values for our evaluation. For the purpose of achieving reliability in this study, we selected only those dimensions out of all available ones, that had the Generalization Coefficient value (for details see Section 3.2 above) of inter-rater

agreement > 65%, as calculated by Rakoczy and Pauli (2006) (see Table 2). As the two different lesson types (ProbSol and Theory) had two different sets and numbers of annotators, they are listed separately here.

Given the relatively small data set, we have approached the problem as a binary classification task and have divided the lessons into high- and low-rated lessons. A score of [1.0-2.0] indicates a low rating, while a score of [3.0-4.0] indicates a high rating. Lessons with an average score between 2.0 and 3.0 were ignored for a given quality dimension.

Based on the quality criteria stated above and after determining highly correlated dimensions, we selected six dimensions with sufficiently wide rating distribution (i.e. more than 35 lessons in each of the rating intervals 1-2 and 3-4) to conduct our machine learning experiments. Most of these dimensions relate to how the teacher treats the students, as explained below in this section.

Dimension	ProbSol	Theory
Relevance of lesson content (RELEV)	.83	.89
Recognition of the student (RECOG)	.85	.74
Objective and constructive feedback (FEED)	.89	.70
Learning community (LEARN)	.84	.70
Cooperation (COOP)	.99	.80
Exploration ofhinking processes (THINK)	.95	.65

Table 2: Values for Inter-Annotator Agreement using the Generalization Coefficient as calculated by Rakoczy and Pauli (2006) for the dimensions we analyzed.

Figure 1 shows the dimensions this work focuses on, along with the number of high and low rated lessons they each contain. Note that the majority class is different for individual dimensions (high for FEED and low for THINK) and some dimensions show stronger imbalance than others (THINK vs. COOP). Table 3 displays basic length-based statistics of the data in the dimensions used.

No. of	$\sum$(T)	$\sum$(S)	$\sum$(all)	avg(T)	avg(S)	avg(all)
Utterances	45 546	32 696	78 242	243.56	174.84	209.20
Sentences	73 509	27 380	100 889	393.10	146.42	269.76
Tokens	705 467	253 213	958 680	3772.55	1354.08	2563.22

Table 3: Dataset statistics in terms of total and average numbers of teacher(T), student(S) and combined(all) utterances, sentences and tokens in the data.

Below we give a brief description of each dimension used in a study. A more detailed description of these, the remaining dimensions and underlying scientific motivation can be found in the original annotation guidelines by Rakoczy and Pauli (2006).

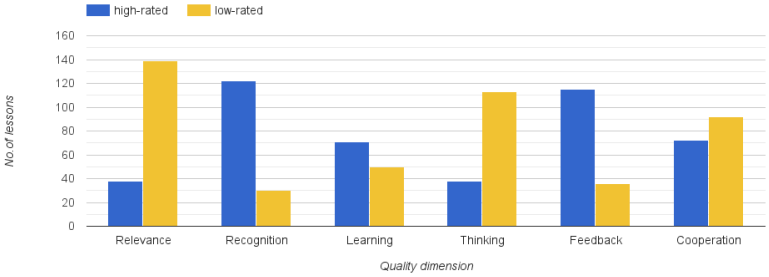

Figure 1: Distribution of lessons in the dimensions under analysis.

4.1 Objective and constructive Feedback (Feed)

This dimension rates the amount and quality of feedback, for example the teacher should be benevolent, provide guidance through the improvement path and show no sarcasm. Constructive feedback shall motivate students and improve their methodology by correcting the student but not discouraging him to try again. Objective means that the feedback only focuses on the topic and not on the person.

4.2 Exploration of thinking processes of students (Think)

This dimension rates the aptitude of the teacher to request detailed explanations. The teacher shall actively encourage students to justify their answers. This allows for an easier understanding of the material by the students. Additionally, this dimension rates whether the teacher attempts to learn the background of student's answers. Suggested indicators for this dimension are *why-* and *how-* questions. The teacher encourages the students to explain phenomena in his/her own words rather than repeating what the teacher has said.

4.3 Cooperation (Coop)

This dimension relates to how well the students support each other during work in smaller groups. The teacher shall show appreciation for team work, students shall appear accustomed to work together.

4.4 Relevance of lesson content (Relev)

This dimension rates how much the teacher tries to present the students the relevance of what he/she is teaching. Suggested indicators for this dimension are examples taken directly or indirectly from every day life of the students, they are allowed to work with every day items or the phenomenon relates to something the students know from their

every day life. But the historical context of what is being presented should not be forgotten. The quality of the examples chosen by the teacher for this purpose is more important than the quantity.

4.5 Recognition of the student (Recog)

Ratings in this dimension show how respectfully a teacher treats the students, whether he/she takes them seriously or mocks them using cynicism and sarcasm. This is especially obvious when students give wrong answers. Suggested indicators for this dimension are the confidence and the interest which the teacher shows towards the students, when they describe their perspectives and opinions. Presence of stinging jokes results in a lower score.

4.6 Learning Community (Learn)

Here, the working atmosphere in the class is rated, which is supposed to be supportive. The teacher is not supposed to take the role of a superior, but is part of the group, which tries to achieve a learning goal. Suggested indicators for this include the teacher using the first person plural ("wir" in German), teacher and students listening to each other and interacting without mocking each other.

5 Experimental Setup

We first build one classification model for each of the quality dimensions separately, i.e. for feedback quality, cooperation support etc., and later combine these into a global model. Since our goal is to identify features particularly characteristic for each dimension, we assume that a model which successfully learns useful information from the data shall perform better on predicting ratings for the dimension in question than on predicting the other ratings. To verify this assumption, we perform cross-dimensional tests for each model, i.e. examining how an all-features classification model trained on one dimension performs on another dimension.

Due to the small data set size, we use a Leave One Out Cross-Validation approach (LOOCV). In order to prevent learning phenomena specific to a certain teacher-class-combination, we modify the LOOCV-approach by excluding lessons of the same teacher-class-combination from the training set. We hereafter call this approach Leave One *Classroom* Out Cross-Validation (LOCOCV). We compare our results to the majority class baseline (Table 1) with respect to the accuracy.

Relations between the dimensions are displayed in Table 4. The first value of each cell shows the Pearson's correlation for the ratings. There is a strong correlation between the dimensions Feedback quality and Learning environment, Thinking processes and Receptive understanding, and Feedback quality and Receptive understanding.

The second value in each cell shows the percentage overlap of lessons used for each dimension, based on the average rating score in the dimension. It shows that for example

about 46% of the data from the Feedback dimension is shared with the dimension Cooperation.

	Feed	Coop	Think	Relev	Learn	RecUnd
Feed	1 / 100%	.34 / 46%	.42 / 40%	.22 / 41%	.66 / 54%	.61 / 43%
Coop	.34 / 43%	1 / 100%	.29 / 54%	.09 / 55%	.39 / 42%	.20 / 35%
Think	.42 / 40%	.29 / 58%	1 / 100%	.08 / 62%	.48 / 50%	.77 28%
Relev	.22 / 35%	.09 / 51%	.08 / 53%	1 / 100%	.35 / 38%	.21 / 35%
Learn	.66 / 67%	.39 / 57%	.48 / 62%	.35 / 56%	1 / 100%	.36 / 37%
RecUnd	.61 / 40%	.20 / 35%	.77 / 26%	.21 / 38%	.36 / 28%	1 / 100%

Table 4: First value in the table represents the Pearson's pair correlation coefficient for the dimension ratings. Second value shows the percentage share of lessons from one dimensions of experimental data (row) in the experimental data of another dimension (column), i.e. overlap of identical lessons with rating 1.0-2.0, resp. 3.0-4.0 between dimensions.

5.1 System Architecture

Our experimental setup is based on the Darmstadt Knowledge Processing (DKPro) (Eckart de Castilho and Gurevych, 2014) software repository, an open-source Natural Language Processing (NLP) toolkit building upon the Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004). We publish the Java code used in this experiment as open-source on our website (see Section 1).

We preprocess the data using the TreeTagger (Schmid, 1994) components for German lemmatization, Part of Speech (POS) and chunk tagging. Our machine learning configuration is based on the Waikato Environment for Knowledge Analysis (WEKA) toolkit (Witten and Frank, 2005) and consists of a Support Vector Machine (SVM-SMO) classifier with polynomial (quadratic) kernel and default parameters, wrapped in a cost sensitive meta classifier with error costs empirically set to account for the class imbalances.

We examine the contribution of individual features, described in Section 5.2, using the ablative analysis, the Information Gain scores and the WEKA visual analytics capabilities, and summarize our findings in Section 6.

5.2 Features Used

Our features are clustered into seven groups, described below. Details, including references and examples, are provided in the individual sections.

We empirically selected the strategy of normalizing count-based features per utterance (as opposed to normalization per time or sentence), averaged per lesson.

For every feature, we additionally measure values for teacher and student utterances in the lessons individually, as well as the ratios between a feature value for the teacher and the student.

5.2.1 Ngram features (Ng)

This group of features consists of the 500 most frequent word uni-, bi- and trigrams from each fold of the training set after stopword cleaning[4]. We also derived phrase triplets based on constituency parsing, i.e., the 500 most frequent word trigrams extracted from the NP-VP-NP triples in an utterance (Levy et al., 2013), as these appeared to represent the content of short questions and answers well in automated question-answering tasks.

5.2.2 Surface features (Su)

These features are commonly used in text classification tasks. We measure text length ratios, expressed in terms of length of a word, sentence or utterance, information-based values such as the average $tf * idf$ score per utterance, and the type-token ratio. Additionally, this group includes surface features based on meta data taken from the transcripts, such as the time taken per utterance and per speaker, and the number of speaker changes. We hypothesize that longer monologues of the teacher might lead to student disengagement (Forbes-Riley et al., 2012), thus lower rating on LEARN. Unfortunately, the transcriptions did not allow us to capture pauses between speakers, which were helpful for Forbes-Riley et al. (2012).

5.2.3 Stylistic features (Sty)

These features capture aspects such as the level of contextuality (CF) in the wording of an utterance, measured based on POS tags used (Heylighen and Dewaele, 2002) :

CF = (noun frequency + adjective freq. + preposition freq. + article freq. - pronoun freq. - verb freq. - adverb freq. - interjection freq. + 100) * 0.5

We expect it to influence student engagement, i.e., our hypothesis is that more casually speaking teachers could be more engaging, and that an increased usage of pronouns, interjections etc. at the expense of nouns can serve as a proxy to capture this casual way of speaking (Heylighen and Dewaele, 2002). We measure also the usage of modals and conditionals, which we assume to indicate student uncertainty (Forbes-Riley and Litman, 2011), and the proportion of each Part-of-Speech tag to other ones in teacher's and students' utterances. We excluded syntactic features based on dependency parsing and Part-of-Speech bi- and trigrams, as they negatively impacted our classification results on the development data. We hypothesize that this is due to spoken language transcription, which has specific properties leading to low parsing accuracy (see for example work by Bechet et al. (2014) for experiments on parsing spontaneous speech).

5.2.4 Sentiment and other lexicon-based features (Se)

These features are mainly based on the Linguistic Inquiry and Word Count (LIWC) utility (Pennebaker et al., 2001) in its German adaptation (Wolf et al., 2008). The

[4] `http://snowball.tartarus.org/algorithms/german/stop.txt`

88 word lists in LIWC contain valuable information not only on emotion (e.g. words expressing anger, sadness or fear), but also social processes (e.g. friends, family, communication) and cognitive processes (e.g. certainty, insight or discrepancy), validated by expert judges. LIWC additionally counts several syntactic aspects, e.g. pronoun type or verb tense.

We hypothesize that teachers using harsh words score consistently lower in the quality dimensions, in comparison to teachers using a lot of encouraging words. Therefore, we employ also the SentimentWortschatz lexicon for an additional, more fine-grained sentiment polarity measure of the lessons, on top of LIWC. SentimentWortschatz (Remus et al., 2010) lists polarity bearing German words weighted within an interval of [-1; 1] plus their Part-of-Speech tag, and if applicable, their inflections. In addition, we also use amplifying and down-toning intensifiers (such as *very*) translated by a German native speaker from their English form (Taboada et al., 2011). We also add polarity changers, such as *not* (Steinberger et al., 2012), and compute alternative sentiment polarity measures with negations and intensifiers included, in the same way as the authors of these intensifiers do (Steinberger et al., 2012; Taboada et al., 2011). However, we do not observe any difference in performance compared to using the plain lexicons without the polarity changers and intensifiers.

Next, we add three features based on grammatical mood (Eisenberg et al., 1998). A grammatical mood of verbs (subjunctive, imperative, indicative) allows speakers to express their attitude to the statement (for example desire, doubt or command). We calculate the proportions of these utterances in the lessons based on syntactic annotations of verbs, combined with the sentiment polarity of the utterance, i.e., counting for example negative and positive imperative as two separate features.

Finally, we monitor the occurrence of several custom-defined words indicating politeness, such as *Danke* (Thank you) and *Bitte* (Please).

A brief overview of lexicons used in this section is provided in Table‘5

Resource	Example
German LIWC	*Anger: hate, kill, annoyed, Insight: think, know, consider*
SentimentWortschatz	*harmonic: +0.5243 ADJ, crisis: 0.3631 NN*
168 intensifiers	*very, slightly, somewhat, extraordinarily*
15 polarity changers	*not, none, any*
grammatical mood	doubt vs. command
politeness words	*thank you, please*

Table 5: Lexicon-based features coming from various German and English resources, of which the latter have been translated for our purpose.

5.2.5 Features based on discourse indicators (Di)

These features capture the discourse-relevant information. We model the Boolean presence and the normalized count of individual discourse markers in the utterances as features. To detect these, we utilize two lexicons of discourse markers - the German

DimLEx (Stede and Umbach, 1998) and the 15 most frequent discourse markers of the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) translated by a German native speaker. Discourse markers are lexical items, annotated in their lexicon with a particular discourse relation they tentatively express, such as `Cause`, `Reason` or `Opposition`. Each of the 173 Discourse Markers are grouped according to their discourse relations and are one word-count-based features in our model. We also count discourse marker bigrams, i.e., the occurrence of pairs of two consecutive discourse relations appearing in the same utterance.

Additionally, we calculate the ratio of nouns repeated by one speaker from the previous speaker in proportion to the total number of nouns used by the speaker in that utterance, assuming that higher overlap demonstrates better understanding, based on the work by Ward and Litman (2007) on lexical convergence. This number is then averaged per lesson, i.e. how many nouns on average do the students repeat after previous speaker and how many nouns on average does the teacher repeat after the previous speaker.

We also measure the frequency and the pattern type of speaker changes as an indicator of student initiative turns, demonstrated as predictors of the knowledge co-construction (Kersey et al., 2009). The pattern is defined as a bigram of speaker annotations (column Speaker in Table 1), for example S-T, T-S or S-SN (student followed by new student). We further monitor individual transcribed non-verbal expressions and attentive back-channeling. These include sighs, laughter, but also so-called social noise, such as *mhm*. Back-channeling is a way of showing a speaker that you follow and understand their contribution, often through interjections, such as "I see" or "ok" (Gweon et al., 2009).

5.2.6 Topic Features

We hypothesize that due to certain expressions (e.g. expert mathematical terms) students may perceive the content as harder, causing disengagement (Forbes-Riley et al., 2012) and impacting the rating of the lesson. Hence, we crafted custom word lists to measure the frequency of terms related to topics discussed in the lessons. Among others, these were on *mathematics*, *school* and the *Pythagorean theorem*. A separate word list was used to capture expressions unrelated to the lesson content and possibly indicating creativity of the teacher – this contained words such as *princess, castle, river* and so on. Additionally, we built topic models on the teacher's and students' speech in high and low rated lessons using Mallet topic modelling tools (McCallum, 2002) empirically set to 50 topics. See Table 6 for an example of topics used in this feature group.

5.2.7 Phonetic (Ph)

Phonetic features have been used in text processing areas such as machine translation (Vogel et al., 2003) or normalization (Han et al., 2012), however they remain unexplored for more abstract tasks, such as the prediction of lesson quality. Our intuition behind this group of features is that certain phoneme combinations may be difficult to understand or certain phoneme occurrences may point to the sentiment of a speaker (Nastase et al.,

Topic	Example
mathematics	*comma, squared, surface*
school	*homework, pupil, lesson*
pythagorean theorem	*triangle, right angle, theorem*
storytelling	*farmer, river, house, tree*
Mallet topic 1:	*five, two, fourteen, seven*
Mallet topic 2:	*mhm, ah, aha, hm, okay, oh*

Table 6: Topical features based on the transcripts and the discussed topics.

2007). We phonetize the transcripts using a German text-to-speech tool (Schröder and Trouvain, 2003) and analyze the frequency of each type of phonemes (e.g. plosives, fricatives, glottal stops, etc. - see Table 7 for examples).

Phonetic	Example (in SAMPA notation)
frequency of plosives	p, t, k, ...
frequency of fricatives	f, S, Z, x, ...
frequency of vowels	I, E, a, Y, ...
...	

Table 7: Phonetic Features in SAMPA notation, taken from the automatically phonetized transcripts.

6 Results

In the following, we present our findings for the six dimensions examined (Objective and constructive feedback, Exploration of thinking processes, Cooperation, Relevance, Recognition of a student, Learning community) as well as the analysis of their relations both on the gold-standard and classification-model levels. In order to illustrate the usefulness of the presented methods in supporting the analysis of classroom interaction through educational researchers, we discuss in detail the results on the three dimensions with the highly informative features beyond lesson ngrams. These three dimensions are: Exploration of thinking processes, Objective and constructive feedback and Cooperation. Other dimensions were omitted since their generalizable features are related to the ones in selection (e.g. Recognition of a student shows similarities to constructive feedback) and their unique features are of less interest (e.g. the Relevance dimension is specific to the lesson content).

6.1 Summary of classification results

Table 8 shows the comparison of the outcome of the classification system (Support Vector Machines, detailed in 4.1) to human annotators using percentage agreement (SysAA). Our system performs comparably to a human annotator on every dimension, suggesting, that these highly abstract tasks include computationally measurable clues.

	Theory			Problem Solving			Combined	
Dim	IAA	SysAA	Acc	IAA	SysAA	Acc	BaseF_1	BestF_1
Think	.66	.84 (.80-.87)	.92	.95	.73 (.72-.75)	.78	.647	.855
Relev	.87	.86 (.85-.88)	.87	.99	.75 (.74-.75)	.89	.691	.873
Recog	.70	.71 (.64-.77)	.72	.95	.84 (.67-.87)	.73	.505	.716
Feed	.69	.79 (.63-.89)	.88	.83	.90 (.90-.90)	.90	.609	.883
Learn	.65	.82 (.79-.86)	.83	.78	.88 (.88-.88)	.88	.434	.851
Coop	.77	.82 (.79-.84)	.87	.99	.83 (.83-.83)	.86	.403	.866

Table 8: Results comparing our system to human performance.
IAA: percentage agreement on Theory and Problem Solving within the three and two human annotators respectively.
SysAA: percentage agreement when considering the system as an additional annotator (in brackets is the minimal and maximal pair agreement with individual annotators).
Acc: percentage agreement between the system results and the average human annotator rating (taken as our Gold standard).
F_1: Results for the majority baseline and the best F_1 score achieved through the machine learning setup for each studied dimension.

Our best results for binary classification of high- and low-rated lessons (see F_1 in column Combined in Table 8) differ from the weighted majority baseline (BaseF_1) significantly ($p < 0.05$) in all dimensions. Statistical significance of differences was computed using an approximate randomization approach (Noreen, 1989). For human annotators (IAA), the Problem Solving lessons were not as challenging to rate as the Theory lessons, possibly due to a more straightforward student-teacher interaction, more explicit feedback etc. For the system (Acc), this issue does not arise – performance on both lesson content types is comparable.

6.2 Global Model and comparison between dimensions

Using the ratings in the six quality dimensions above, we also trained a global model. We selected only lessons rated high [3-4] in at least 3 dimensions and low in at most two (70 lessons), resp. rated low [1-2] in at least 3 dimensions and high in at most two (61 lessons). We then perform binary classification to discriminate such overall high-rated from the overall low-rated lessons and examine feature rankings.

Distribution of individual dimensions in the global model is very similar to the one described in Figure 1. The most frequently appearing high ratings in the overall high-rated lessons are in the dimensions Feed and Learn, while the low-rated lessons are most frequently rated low for the dimensions Think and Relev.

We achieve the best global result ($F_1 = 0.885$) using the combination of discourse, surface and phonetic features and topics. Discourse features are the most predictive ones based on an ablation test. In the detailed analysis, we found that in the global model high rated lessons were characterized by:

- increased reasoning and elaboration discourse markers from the student side (as in LEARN, COOP and THINK)

Feature		Dimension					
Ngrams, topics (Ng/To)	Relev	Recog	Feed	Learn	Coop	Think	Global
wie, warum				•			
besprechen, resultat						•	
mhm, ja, hm, genau		•	•				
Storytelling: felder, wasser...	•						
Number topic: komma, null...	•						•
Surface (Su)	Relev	Recog	Feed	Learn	Coop	Think	Global
(S)No.of dialog turns				•			•
Talk ratio T/S				•			•
(S)Answers <5 words		•		•	•	•	•
(S)Answers >25 words		•		•	•		•
Syntax (Sy)	Relev	Recog	Feed	Learn	Coop	Think	Global
(T)Interjection ratio		•	•				
(S)Verb ratio		•	•		•		
(S)Adverb ratio		•	•				
(S)Pronoun ratio		•	•		•		
(S)Conjunction ratio			•				
LIWC & Sentiment (Se)	Relev	Recog	Feed	Learn	Coop	Think	Global
(T,S)Positive		•	•				
(T)Communication					•	•	
(T,S)Cognitive				•	•		
(T,S)Cause				•		•	
(S)Question words		•	•			•	
(S)We, Self					•		
(T,S)Assent					•		
(T)You					•		
(S)Discrepancy		•					
(S)Negate		•					
Discourse (Di)	Relev	Recog	Feed	Learn	Coop	Think	Global
(T)Attentive signs (*ja*,nods)		•	•	•			•
(S)Attentive signs			•				•
(S)Cause				•	•	•	•
(S)Reason				•		•	
(T)Reason	•	•				•	
(S)Compare	•	•		•		•	•
(T,S)Elaboration	•			•			•
(T)Circumstance	•						
Dialog sequence S-S		•			•		
Phonetic (Ph)	Relev	Recog	Feed	Learn	Coop	Think	Global
(T)Syllables per word		•	•		•		

Table 9: Detailed presentation of features with the highest information gain (top 20) for the best classification model in each dimension. Teacher (T) or Student (S) indicates features measured for the respective utterances only.

- more back-channels from both sides (similarly to FEED and LEARN)
- longer student utterances (similarly to COOP and THINK)
- relatively more student utterances in comparison to the teacher (similarly to LEARN)

To validate that our models learn aspects relevant to the dimension in question, we measured the performance of classifiers trained on one dimension (row in Table 10)

and tested on another (column in Table 10). Expectably, results for dimensions with a higher rate of correlation and data overlap (see Table 4) are better than for dimensions with lower rate of correlation and data overlap, however, all cross-dimensional train-test results (see Table 10) were still significantly different (worse) than a cross-validation on each single dimension itself (see Table 8). This suggests that each of our three models learned features predominantly relevant for its dimension of interest.

	Cross Relev	Recog	Feed	Learn	Coop	Think
Relevance	1	.36	.38	**.62**	**.61**	**.67**
Recognition	.40	1	**.92**	**.69**	.46	.37
Feedback	.44	**.87**	1	**.68**	.49	.42
Learning	.59	**.69**	**.74**	1	**.62**	**.64**
Cooperation	.57	.47	.51	.60	1	**.62**
Thinking	**.65**	.37	.43	**.69**	**.65**	1

Table 10: Percentage accuracy for the best classification model from each dimension (row) tested on another dimension (column). Accuracy is chosen here over F-score for better comparison with the correlation and data overlap values presented in Table 4. Accuracy scores $\geq$0.6 are highlighted.

Across all dimensions, students in high rated lessons are given a chance to express themselves in more elaborated and argumentative manner, while teachers extensively demonstrate their attention and stimulate the communication. Already the simple discourse features and word categories related to sentiment and cognitive processes show high predictive power in our task, which is a promising path for an automated assistance in qualitative evaluation of teaching.

6.3 Detailed feature analysis on selected dimensions

In previous work, Rosé et al. (2008) (see Section 2) analyzed the possibility to employ NLP features for the task of classification of cooperation in learning environments. The authors note that in the future it would be beneficial to evaluate "which features provide the greatest predictive power for the classification" (Rosé et al., 2008, p.266). To explore this question, we discuss below in detail the results on the three dimensions with the highest potential of being generalized to other studies. These are: the Exploration of thinking processes, Objective and constructive feedback and Cooperation.

A detailed examination of ablative tests with different feature groups for each dimension revealed that features from different groups are in many cases mutually substitutive, indirectly representing the same phenomenon (see Table 9). For example, the length of sentences, captured in surface features is also apparent through a larger variety of POS tags and discourse markers present in the utterance. Similarly, back-channels are reflected in ngrams and word length etc. Therefore, we further examine the ranking of individual features based on information gain in order to understand the underlying phenomena. For each dimension, the features consistently scoring high across classification folds are listed in the following, together with our suggested interpretation.

6.3.1 Exploration of thinking processes of students (THINK)

Features	Ngram	Surface	Syntax	Sentiment	Discourse	Phonetic	All	Base
F-score	.809	.767	.678	.640	.855	.647	.779	.647

Table 11: Performance (F-score) of individual feature groups for the dimension *Exploration of thinking processes*. Baseline is a majority baseline weighed accordingly to the lesson proportion in the high and low class.

Table 11 reveals that especially the discourse features, ngrams and surface features were important for this dimension. Table 12 sheds additional light on this classification task by highlighting individual features with the highest information gain for this dimension. We found that high rated lessons are characterized through student features such as Reason and Cause (*weil (because), so dass (so that)*), question words (*wie (how), wo (where), woher (where from), warum (why)*) and long utterance (more than 25 words). On the teacher side these are characterized through features such as Comparison and Elaboration (*insbesondere (especially), das heisst (that means)*) and Communication (*sagen (say), fragen (ask), meinen (mean), beschreiben (describe)*).

Feature	Examples
For speaker: Teacher	
$\text{Communication}_{LIWC}$	*sagen (say), fragen (ask), meinen (mean), beschreiben (describe)*
Compare_{PDTB}	*insbesondere (especially)*
$\text{Elaboration}_{DiMLex}$	*das heisst (that means)*
For speaker: Student	
Cause_{PDTB}	*weil (because)*
Reason_{DiMLex}	*so dass (so that)*
Utterances >25 words	
Questions_{LIWC}	*wie (how), wo (where), woher (where from), warum (why)*

Table 12: Features predictive for high rating in THINK.

We conclude that these features approximate the behaviour observed by educational researchers: In highly rated lessons the teacher encourages the students to communicate and students ask more questions. Both students and teachers use more reasoning, especially students have longer utterances and compare and explain concepts.

6.3.2 Objective and constructive feedback (FEED)

In the ablative tests (Table 13), this dimension is best predicted by discourse-based, sentiment- and cognition-oriented lexicon-based features, and ngrams. The most informative individual features are listed in Table 14. Teachers use positive words and both teachers and students give back-channels, which is reflected also through ngrams, interjection frequency and phonetic features. Students use question words, but

Features	Ngram	Surface	Syntax	Sentiment	Discourse	Phonetic	All	Base
F-score	.872	.739	.735	.757	.831	.739	.883	.659

Table 13: Performance (F-score) of individual feature groups for the dimension *Objective and constructive feedback*. Baseline is a majority baseline weighed accordingly to the lesson proportion in the high and low class.

also words from the group Discrepancy and negation (e.g. *Das wollen wir aber nicht, oder? (We don't want this, do we?)*), Comparison and Specification (e.g. *oder (or) ... beispielsweise (for example)*). Similar to THINK, long sentences and a high variation in POS are very predictive for this dimension.

Feature	Examples
For speaker: Teacher	
Positive$_{SentiWS,LIWC}$	*gut (good), perfekt (perfect), wunderbar (wonderful)*
Interjection rate (high)	
For speaker: Student	
Questions$_{LIWC}$	*wie (how), wo (where), woher (where from), warum (why)*
Discrepancy$_{LIWC}$	*aber (but), hoffe (hope), sollte (should)*
Negation$_{LIWC}$	*nicht (not), keine (none)*
Comparison+ Specification$_{DiMLex}$	*oder (or)...beispielsweise (for example)*
Pronoun rate (high)	
Verb rate (high)	
For all speakers	
Back-channels	*hm, ja (yes), mhm, genau (exactly)*
One-syllable words	

Table 14: Features predictive for high rating in FEED.

Lessons rated **high** on FEED are also characterized by higher frequency of indicative verbs, demonstrative pronouns and subordinate conjunctions, hinting towards an increased number of complex factual sentences. Together with the higher question word ratio observed in students' turns, this suggests, that an environment with constructive feedback may encourage students to pursue the problems further and discuss them with the teacher. Lessons scored **low** in this dimension show on average lower frequencies of positive sentiment words and in several cases even some negative tone, e.g. *Das ist falsch* (That is wrong).

We hypothesize that these features are indicative of the behaviour observed by educational researchers: Both students and teachers actively listen to each other. The teacher encourages the students to proceed and the students express opinions and voice questions. Additionally, they do not hesitate to ask even when they are unsure.

Tentatively, an environment with constructive feedback appears to support students to pursue the problems with more confidence and discuss them with the teacher.

6.3.3 Cooperation (COOP)

Features	Ngram	Surface	Syntax	Sentiment	Discourse	Phonetic	All	Base
F-score	.662	.739	.735	.754	.712	.564	.725	.403

Table 15: Performance (F-score) of individual feature groups for the dimension *Cooperation*. Baseline is a majority baseline weighed accordingly to the lesson proportion in the high and low class.

This dimension benefits from the entire range of features. Among the best performing useful feature groups are the syntactic, surface, and sentiment and other lexicon-based features. Table 16 lists individual features with the highest information gain for this dimension.

The use of back-channels on the teacher side was also useful in FEED, whereas the use of communication words on the teacher side was useful in THINK. The other features that best predicted this dimension are: The specific speaker pattern of student speaker followed by a new student speaker (S-SN), indicating that students discuss things amongst themselves. This is supported by the use of *we* rather than *I* on the student side. The teacher uses *you* (German second person plural, referring to the group). Also, the students use words from Alternative and Comparison (*oder (or) ... obwohl (although)*, Alternative and Elaboration (*oder (or) ... beispielsweise (for example)*), Contrast and Elaboration (e.g. *anderseits (on the other hand) ... und (and)*) in their utterances. Also, the students use a range of cognition words, such as *erkennen (recognize), konstruieren (construct), wissen (know)*. Also, the students use long utterances.

In the lessons rated **low** for COOPERATION, students often use very short sentences, as in most of the previously discussed dimensions.

Feature	Examples
For speaker: Teacher	
Communication$_{LIWC}$	*sagen (say), fragen (ask), meinen (mean), beschreiben (describe)*
Back-channels	*hm, ja (yes), mhm, genau (exactly)*
Pronoun You$_{LIWC}$	
For speaker: Student	
Cognition$_{LIWC}$	*erkennen (recognize), konstruieren (construct), wissen (know)*
Alternative+Comparison$_{DiMLex}$	*oder (or) ... obwohl (although)*
Alternative+Elaboration$_{DiMLex}$	*oder (or) ... beispielsweise (for example)*
Contrast+Elaboration$_{DiMLex}$	*anderseits (on the other hand) ... und (and)*
Speaker pattern S-SN	student - new student
Pronoun We$_{LIWC}$	
Utterances >25 words	

Table 16: Features predictive for high rating in COOP.

These results capture the following aspects relevant for educational researchers: Students communicate among each other and perceive themselves as a team, using *we* rather than *I*. They speak more and make their own suggestions. The teacher encourages this behavior by showing attention, while letting the students provide the explanations.

6.4 Discussion

In general, it can be concluded that a **high**-rated lesson in our data set is characterized by the following: Firstly, increased **reasoning activity** of the students, visible through the discourse markers. Secondly, increased **engagement of the students**, apparent through length-based features. And thirdly, increased **encouragement and attention** from both sides, detected through specific interjections, adjectives and adverbs and through non-verbal signs such as nodding. Particularly, the reasoning activity analysis could benefit from an extended, focused NLP research on argumentation, which is beyond the scope of this paper.

In accordance with Kersey et al. (2009), we find speaker changes to be a good predictor of cooperation. Also, back-channeling appears important for indicating collaborative thinking in our data set, in line with Gweon et al. (2009) inter alia. Usage of plurals (*we* vs. *I*) and words indicating cognition and communication were also prevalent in high-rated lessons along multiple dimensions. Short student utterances in general predict a lesson of low quality, which may indicate disengagement as described by Forbes-Riley and Litman (2012).

Previous research in educational NLP (see Sections 2 and 3) indicated that feedback does not influence the performance in post tests. Despite this, our results suggest that it may influence the way students express themselves – lessons, where students obtain more and better feedback, correlate with lessons where students speak more often. However, the direction of the causation needs to be further studied.

Our study is of relatively small scale, partly due to the necessity of manual transcription of video texts. Future research could overcome these limitation by using automated speech recognition tools or focusing on multimodal features. Additionally, an extension to teaching domains other than mathematics would be welcome to verify the generalizability of our findings.

While the identification of specific misclassification sources on the document level is challenging in our data, due to the high level of abstraction in ratings, we identified two areas of improvement.

Lexical ambiguity Lexical resources such as LIWC are lacking additional annotations to support sense disambiguation. Hence, particular words were occasionally misrepresented, as they were often used in a sense irrelevant to the category. For example the word *(S|s)chau* ([the] show|looking at) was grouped in the LIWC category `TV` even though it never occurred in the lessons in relation to television. Also, the discourse markers are known to be highly ambiguous (Stede and Umbach, 1998), e.g. the word *while* can represent a contrast as well as temporal co-occurrence.

Loss of audio/video information Restricting ourselves to the analysis of text deprives our experiments of information from the acoustic and visual parts of the data, which was available to the expert raters. We hypothesize that the maximum attainable performance is lower than for a multimodal system. For example, sarcasm in speech, which was often present in our data, is more easily discernible through prosodic and facial gesture features (see for example Rakov and Rosenberg (2013)), and so is user disengagement (Forbes-Riley and Litman, 2012) or convergence between conversants (Ward and Litman, 2008). Furthermore, we found that several misclassified documents contained markers of dialog instances inaudible to the transcriptionists.

7 Conclusion and Future Work

Predicting the quality of classroom lessons and analyzing the interaction between teachers and students and among students is an educational research topic of great importance. In this paper, we present initial experiments on the task of assessing several quality dimensions of classroom interaction, employing an existing data set of this kind for the first time. We model this as a text classification task, demonstrating the high potential of automated quality prediction systems to assist educational researchers. We present a freely available data set of German classroom transcripts and expert ratings on quality dimensions such as constructive feedback, thinking process and cooperation.

We defined a broad range of features from diverse NLP areas, reflecting the analysis of the verbal behaviour of the teachers and students, such as discourse analysis, phonetics and emotion detection. We applied machine learning techniques to classify lessons according to the dimensions highly relevant for educational researchers.

We carefully examined the relation between each of the measured phenomena and the quality dimensions, and suggested an interpretation of the most remarkable findings. We successfully built classifiers comparable to human annotators on this data set.

Our findings on the relevance of various feature groups offer room for extension both on the NLP and the educational researchers' side. On the latter, it would be worthwhile to analyze the correlation between the students' performance and the features which possibly influence the quality of a lesson, e.g. the back-channeling of a teacher. In continuation of our collaboration, it would be interesting to examine the benefit for the educational researchers of using a semi-automatic approach based on this work in the annotation of future data sets.

We hypothesize that the maximum attainable performance of our approach is lower than for a multimodal system. For example, sarcasm in speech, which was often present in our data, is more easily discernible through prosodic and facial gesture features (see for example Rakov and Rosenberg (2013)), which require signal analysis both on the visual and the acoustical part of the data. Our approach can be applied to further data sets of similar kind [5]. Automatic speech recognition (ASR) could be used in order to see how stable our results are in light of noisy, ASR-output.

[5] such as `http://www.timssvideo.com/timss-video-study`

Acknowledgement

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant No. I/82806 and by the German Research Foundation under grant No. GU 798/14-1. The authors thank Prof. Klieme, Dr. Katrin Rakoczy and Petra Pinger from the German Institute for Educational Research for their support with the data and educational research questions.

References

Ahrenberg, L. and Tarvi, L. (2014). Translation Class Instruction as Collaboration in the Act of Translation. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42, Baltimore, Maryland.

Bechet, F., Nasr, A., and Favre, B. (2014). Adapting dependency parsing to spontaneous speech for open domain spoken language understanding. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech), 14.–18. September 2014, Singapore*.

Chen, L., Di Eugenio, B., Fossati, D., Ohlsson, S., and Cosejo, D. (2011). Exploring Effective Dialogue Act Sequences in One-on-one Computer Science Tutoring Dialogues. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–75, Portland, Oregon.

Cheng, J., Zhao D'Antilio, Y., Chen, X., and Bernstein, J. (2014). Automatic Assessment of the Speech of Young English Learners. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–21, Baltimore, Maryland.

Clausen, M., Reusser, K., and Klieme, E. (2003). Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen. Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. *Unterrichtswissenschaft*, 31(2):122–141.

Di Eugenio, B., Lu, X., Kershaw, T. C., Corrigan-Halpern, A., and Ohlsson, S. (2005). Positive and Negative Verbal Feedback for Intelligent Tutoring Systems. In *Proceedings of the 2005 Conference on Artificial Intelligence in Education: Supporting Learning Through Intelligent and Socially Informed Technology*, pages 798–800, Amsterdam, The Netherlands. IOS Press.

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In Ide, N. and Grivolla, J., editors, *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Eisenberg, P., Gelhaus, H., Wellmann, H., Henne, H., and Sitta, H. (1998). *Duden, Grammatik der deutschen Gegenwartssprache*, volume 4. Bibliographisches Institut & F. A. Brockhaus AG, Mannheim, 6 edition.

Farra, N., Somasundaran, S., and Burstein, J. (2015). Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74, Denver, Colorado.

Ferrucci, D. and Lally, A. (2004). UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348.

Forbes-Riley, K. and Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Commununications*, 53(9-10):1115–1136.

Forbes-Riley, K. and Litman, D. (2012). Adapting to Multiple Affective States in Spoken Dialogue. In *Proceedings of the SIGdial 2012 Conference: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226, Seoul, Korea.

Forbes-Riley, K., Litman, D., Friedberg, H., and Drummond, J. (2012). Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 91–102, Montréal, Canada.

Gweon, G., Kumar, R., and Rosé, C. P. (2009). Grasp: The group learning assessment platform. In *Proceedings of the 9th International Conference on Computer Supported Collaborative Learning – Volume 2*, CSCL'09, pages 186–188. International Society of the Learning Sciences.

Han, B., Cook, P., and Baldwin, T. (2012). Automatically Constructing a Normalisation Dictionary for Microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432.

Hattie, J. and Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1):81–112.

Hattie, J. A. C. (2009). *Visible Learning: A synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge, New York, USA.

Hausmann, R. G. M., Chi, M. T. H., and Roy, M. (2004). Learning from collaborative problem solving: An analysis of three hypothesized mechanisms. *26th Annual Conference of the Cognitive Science society*, pages 547–552.

Heylighen, F. and Dewaele, J.-M. (2002). Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science*, 7(3):293–340.

Kersey, C., Di Eugenio, B., Jordan, P., and Katz, S. (2009). Knowledge co-construction and initiative in peer learning interactions. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, pages 325–332, Amsterdam, The Netherlands. IOS Press.

Kharkwal, G. and Muresan, S. (2014). Surprisal as a Predictor of Essay Quality. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–60, Baltimore, Maryland.

Levy, O., Zesch, T., Dagan, I., and Gurevych, I. (2013). UKP-BIU: Similarity and Entailment Metrics for Student Response Analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 285–289, Atlanta, Georgia, USA.

Lipowsky, F., Rakozcy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., and Reusser, K. (2009). Quality of Geometry Instruction and its Short-Term Impact on Students' Understanding of the Pythagorean Theorem. *Learning and Instruction*, 19(6):527–537.

Litman, D. J. and Forbes-Riley, K. (2006). Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors. *Speech Communication*, 48(5):559–590.

Loukina, A., Zechner, K., and Chen, L. (2014). Automatic Evaluation of Spoken Summaries: The Case of Language Assessment. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 68–78, Baltimore, Maryland.

Loukina, A., Zechner, K., Chen, L., and Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–19, Denver, Colorado.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit.

Napoles, C. and Callison-Burch, C. (2015). Automatically scoring freshman writing: A preliminary investigation. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 254–263, Denver, Colorado.

Nastase, V., Sokolova, M., and Shirabad, J. S. (2007). Do happy words sound happy? A Study of the Relation between Form and Meaning for English Words Expressing Emotions. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007),* Borovets, Bulgaria, pages 406–410.

Noreen, E. W. (1989). Computer intensive methods for hypothesis testing: An introduction.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count.*

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC),* Marrakech, Morocco, 26 May – 1 June 2008.

Rakoczy, K. (2006). *Motivationsunterstützung im Mathematikunterricht – Unterricht aus der Perspektive von Lernenden und Beobachtern.* PhD thesis, Johann Wolfgang Goethe-Universität, Frankfurt, Germany.

Rakoczy, K. and Pauli, C. (2006). *Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse*, chapter 13, pages 206–233. Number 15 in Materialien zur Bildungsforschung. Gesellschaft zur Förderung Pädagogischer Forschung (GFPF)/Deutsches Institut für Internationale Pädagogische Forschung (DIPF).

Rakov, R. and Rosenberg, A. (2013). "Sure, I Did The Right Thing": A System for Sarcasm Detection in Speech. In *Proceedings of the 14th Annual Conference of the International Speech Communcation Association,* Lyon, France, 25–29 August 2013.

Remus, R., Quasthoff, U., and Heyer, G. (2010). SentiWS – a Publicly Available German-language Resource for Sentiment Analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, pages 1168–1171.

Rosé, C. P., Wang, Y.-C., Arguello, J., Stegmann, K., Weinberger, A., and Fischer, F. (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning. *Computer Supported Collaborative Learning*, 3(3):237–271.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Schröder, M. and Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *Journal of Speech Technology*, 6:365–377.

Stede, M. and Umbach, C. (1998). DiMLex: A Lexicon of Discourse Markers for Text Generation and Understanding. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – Volume 2*, pages 1238–1242, Stroudsburg, PA, USA.

Steinberger, J., Ebrahim, M., Ehrmann, M., Hurriyetoglu, A., Kabadjov, M., Lenkova, P., Steinberger, R., Tanev, H., Vázquez, S., and Zavarella, V. (2012). Creating Sentiment Dictionaries via Triangulation. *Decision Support Systems*, 53(4):689–694.

Swanson, B., Yamangil, E., and Charniak, E. (2014). Natural Language Generation with Vocabulary Constraints. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Baltimore, Maryland.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., and Waibel, A. (2003). The CMU Statistical Machine Translation System. In *Proceedings of MT Summit IX,* New Orleans, USA, volume 9, pages 54–63.

Ward, A. and Litman, D. (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education,* Farmington, PA, USA, October 1-3, 2007.

Ward, A. and Litman, D. (2008). Semantic Cohesion and Learning. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems, ITS* Montreal, Canada, June 23-27, pages 459–469.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., and Kordy, H. (2008). Computergestützte quantitative Textanalyse. *Diagnostica*, 54(2):85–98.