Constructive Feedback, Thinking Process and Cooperation: Assessing the Quality of Classroom Interaction

Tahir Sousa[1], *Lucie Flekova*[2], *Margot Mieskes*[3,4], *Iryna Gurevych*[2,4]

[1]University of Minnesota, USA
[2]Technical University of Darmstadt, Germany
[3]University of Applied Sciences Darmstadt, Germany
[4]German Institute for Educational Research (DIPF), Germany

sousa009@umn.edu, {flekova, gurevych}@ukp.informatik.tu-darmstadt.de, margot.mieskes@h-da.de

Abstract

Analyzing and assessing the quality of classroom lessons on a range of quality dimensions is a number one educational research topic, as this allows developing teacher trainings and interventions to improve lesson quality. We model this assessment as a text classification task, exploiting linguistic features to predict the scores in several lesson quality dimensions relevant for educational researchers. Our work relies on a variety of phenomena, amongst them paralinguistic features, such as laughter, from real classroom interactions. We used these features to train machine learning models to assess various quality dimensions of school lessons. Our results show, that especially features focusing on the discourse and semantics are beneficial for this classification task.

Index Terms: discourse relation, classroom interaction, interaction quality

1. Introduction

Educational researchers extensively analyze the interaction between teachers and students in all age groups in order to find components of school effectiveness. One of the employed methods is based on videography, in which acoustical and visual elements of the lessons are recorded. Researchers analyze these recordings and assess the quality based on the interaction among all participants in order to design teacher trainings or educational interventions. This assessment is a complex process, as multiple experts have to evaluate the quality independently (see e.g. [1]).

In close cooperation with educational researchers, we used the transcripts of a publicly available[1], multimodal corpus of mathematics lessons as a first step to create a machine learning prototype, that models the task of annotating lessons on a range of quality dimensions as a text classification task. To this end, we use features related to e.g. discourse and sentiment analysis. Our key contribution lies in using this type of data for the first time and analyzing the verbal behaviour as expressed in the transcripts, which affect the perception of lesson quality aspects. Our initial results show that features related to e.g. discourse and sentiment analysis allow for good lesson classification within the quality dimensions studied here.

The paper is organized as follows: Section 2 presents related research and Section 3 provides a description of the data set we used. In Section 4, we describe our text classification approach. Section 5 presents the results along with the suggested interpretation of our findings and its discussion. Section 6 concludes our work and addresses future extensions.

This work was carried out while all authors were at DIPF.

[1]http://www.fachportal-paedagogik.de/

2. Related Work

In the following, we present work from the broad range of educational Natural Language Processing (NLP), which is most relevant to our work. The presented works used spoken, transcribed and written material.

Group Interaction An important phenomenon in learning is the interaction among the students. Work such as [2] automatically predict student activity levels in group meetings using only average student talk time and overlap. With these features, they reliably differentiate between the students taking lead in conversation and the ones back-channeling. Others such as [3] 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 [4]. 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 (based on annotated dialog acts (DA)). Machine-learning techniques were used to automatically classify elements of group interaction in written German conversations [5]. The results (using a range of linguistic features) indicate that phenomena of group interaction can be reliably detected based on textual information.

Tutor/Teacher-Student Interaction and Feedback Feedback has been studied extensively, as an important mechanism in teaching (see for example [6] and [7]). Studies on the impact of feedback on mistake vs. feedback when correct found that feedback types were not predictive of post-test results [8]. Other studies on the correlation between DA and learning gain in informatics [9] found "several DA sequences that significantly correlate with learning gain"[9, p.73], such as a prompt followed by an instruction and feedback. The effective DA sequences varied per topic studied. 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" [10, 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.

Student Emotion Detection Several research studies have been conducted on automatically detecting the emotion of students in various situations. Researchers studied the emotion of students in human-human tutoring dialogues as opposed to human-computer ones [11]. 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. Others demonstrated that student uncertainty negatively correlates with learning success [12]. In another work, the authors additionally found that student disengagement negatively affects learning success [13]. 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 and the answer is a strong predictor for student disengagement.

3. The Pythagoras Data set

The data used in this paper originates from a bi-national (Germany and Switzerland) study [14], in which 40 classes of 8^{th} (Germany) and 9^{th} (Switzerland) grade students were videotaped during 5 of their mathematics lessons. During 3 of the lessons each class was introduced to the Pythagorean Theorem (*Theory*) and during 2 lessons each class dealt with problem solving (*ProbSol*) in general. The whole study contains 200 videos, each lesson being 40-50 minutes long. Educational researchers in both Germany and Switzerland analyzed the videos over a several-year period for a range of research questions. During this analysis, 193 videos were manually transcribed. The transcripts include elements such as laughter, coughing and door slams. Pauses were not marked specifically, beyond using "..." for short pauses and splitting segments 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). Table 1 shows a snippet of the transcription. In order to rate the interaction between teachers and students and among students, an annotation scheme where each aspect of the interaction (dimension) is described as a basic idea and a list of indicators such as "Students do not mock each other" was developed [15]. For each of the 28 dimensions defined, each lesson was rated on a 4-point Likert scale by 2-3 expert annotators. The judgments were based on: frequency or duration of the specific behavior during a lesson, intensity of this behavior and distribution across students. These annotations are also called "high-inference ratings".

Time	Speaker	Dialog
00:12:41:01	S	Wenn es um den Pythagroas 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.

Only 187 transcripts (115 Theory and 72 ProbSol) could be used, as the data in the remaining were corrupted beyond repair. Our final data set thus contains 78,242 transcribed conversation segments from a total of 140 hours of recordings.

Dimensions Analyzed From the 28 available dimensions, we used the following three in this first study: **Objective and constructive Feedback** (FEED) rates the amount and quality of feedback, for example the teacher should be benevolent, provide guidance through the improvement path and show no sarcasm. **Exploration of thinking of students** (THINK) rates the aptitude of the teacher to request detailed explanations. The teacher shall actively encourage students to justify their answers. Additionally, **Cooperation** (COOP) relates to how well the students support each other during work in smaller groups. Teacher shall show appreciation for team work, students shall appear accustomed to work together. A more detailed description of these, the remaining dimensions and underlying motivation can be found in the original annotation guidelines [15].

4. Experimental Setup

For our experiments we use only freely available tools, such as the TreeTagger [16], and Support Vector Machines (SVM-SMO) included in Weka [17], in its default settings.

4.1. Classification Task

Given the relatively small data set, we have approached the problem as a binary classification task and have divided the transcripts 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 for each dimension. The classification model is built separately for each of the quality dimensions. Additionally, we perform cross-dimensional tests for each model. Due to the data set size, we use a Leave One Out Cross-Validation approach. In order to prevent learning teacher- or class-specific phenomena, we modify it by excluding lessons of the same teacher 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 2) with respect to the accuracy.

	Feedback	Cooperation	Thinking
high [3.0 - 4.0]	115 (76%)	72 (44%)	38 (25%)
low [1.0 - 2.0]	36 (24%)	92 (56%)	113 (75%)

Table 2: Class distribution of lessons in the 3 analyzed dimensions.

Table 2 shows the dimensions this work focuses on, along with the number of high and low rated lessons that they each contain and that are further used as our experimental data. 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).

	Feed	Coop	Think	Feed	Coop	Think
Feedback	1	.34	.42	100%	46%	40%
Cooperation	.34	1	.29	43%	100%	54%
Thinking	.42	.29	1	40%	58%	100%

Table 3: First (left) part of the table represents the Pearson's pair correlation coefficient for the dimension ratings. Second (right) part 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.

Relations between the dimensions are displayed in Table 3. The left half of the table shows the Pearson's correlation for the ratings. The right half shows the percentage overlap of training, resp. test lessons between these dimensions. We also analyzed the performance of the classifier when trained on one dimension and tested on another. Results for dimensions with a higher rate of overlap are better than for dimensions with lower rate of overlap.

4.2. Features Used

Our features are assorted into six groups, described briefly below, while details, including references and examples are provided in Table 4. We empirically selected the strategy of normalizing count-based features per dialog 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 them.

Ngram (Ng) features consist of the 500 most frequent word uni-, bi- and trigrams from each folds training set after stopword cleaning[2]. **Surface (Su)** features measure common text length ratios, temporal proportions and informativeness.

Stylistic (Sty) features capture aspects such as the level of formality in the wording of an utterance, which we expect to influence student engagement, and the usage of modals and conditionals, which we assume to indicate student uncertainty [12].

Feature group	**Details**
N-grams	**(Ng)**
top 500 n-grams	of the most freq. 1,2,3-grams
Surface	**(Su)**
average length of ...	token, sentence and utterance
transcript metadata	avg utterance *t*, avg speaker *t*
$tf * idf$	summed over tokens per lesson
type-token ratio	on lemmas
Syntax & Style	**(Sty)**
each POS ratio	e.g. ratio of pronouns to tokens
formality score	[18]
ratio of modals	modal verb list[2]
grammatical mood	e.g. imperative, conditional
Semantic	**(Se)**
German LIWC lists	[19][3], e.g. Anger, Communication
15 polarity changers	[20][1,2], *no, barely*
politeness words	freq. of *thank* and *please* words
Discourse	**(Di)**
PDTB lexicon	[21][1], e.g. *but, although*
DiMLex lexicon	[22], e.g. *therefore, however*
non-verbal expressions	sighs, laughter and noise
attentive back-channels	nod, *mhm*, *ach so*, etc.
noun repetition	overlap by 2 consec. speakers
speaker changes	count&ratio of S-T,S-S,T-S, S-SN
Phonetic	**(Ph)**
frequency of ...	plosives, fricatives, vowels...[4]
avg. no of syllables	

Table 4: Features used. (1)Translated from the English original into German by a German native speaker. (2)List (German) available on our website. (3)See also `www.liwc.net`. (4)See also `mary.dfki.de`

Semantic (Se) features are mainly based on the German version of the Linguistic Inquiry and Word Count utility (LIWC). The 88 word lists in LIWC contain valuable semantic 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.

Discourse features (Di) tend to capture the intentions of speakers and their interaction. First, we model the Boolean presence and normalized count of individual discourse markers (DM) in the utterances, using German discourse marker lexicons such as DiMLex [22]. DMs are lexical items, annotated in their lexicon with a particular discourse relation they tentatively express, such as `Cause`, `Reason` or `Opposition`. Each of these discourse relations is one word-count-based feature in our model. We also count the occurrence of pairs of two consecutive discourse relations appearing in the same utterance. Additionally, we capture the repetition of nouns between consecutive speakers, assuming that higher overlap demonstrates better understanding [10]. We also measure the frequency and type of speaker changes as an indicator of student initiative turns.We further measure individual transcribed non-verbal expressions and attentive back-channeling.

Phonetic (Ph) features have been used in text processing before (e.g. machine translation [23] or normalization [24]), but are unexplored for more abstract tasks such as the prediction of lesson quality. We phonetize the transcripts using a standard text-to-speech tool [25] and analyze the frequency of each type of phonemes (e.g. plosives, fricatives, glottal stops, etc.). 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 [26].

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

5. Results

This section presents our findings for each dimension as well as the analysis of their relations.

5.1. System performance comparable to human assessment

	Theory			ProbSol		
Dim	IAA	SysAA	Acc	IAA	SysAA	Acc
Think	.66	.84 (.80-.87)	.92	.95	.73 (.72-.75)	.78
Feed	.69	.79 (.63-.89)	.88	.83	.90 (.90-.90)	.90
Coop	.77	.82 (.79-.84)	.87	.99	.83 (.83-.83)	.86

Table 5: Results comparing our system to human performance using percentage agreement within the annotators (IAA), considering the system as another annotator (SysAA) and comparing the systems results to the Gold standard (Acc).

Table 5 shows the comparison of the outcome of the system to human annotators using percentage agreement (SysAA). Our system performs comparably to a human annotator on every dimension, suggesting, that these highly abstract tasks may include computationally measurable clues. Our best results differ from the baseline significantly ($p < 0.05$) in all dimensions. Statistical significance of differences was computed using an approximate randomization approach; a non-parametric test suitable for F-scores [27]. For human annotators (IAA), the ProbSol lessons were not as challenging to rate as the Theory lessons, possibly due to a more straightforward student-teacher interaction. For the system, this issue does not arise (Acc).

5.2. Ablation tests are not enough

A detailed examination of ablation tests for each dimension revealed that features from different groups are in many cases mutually substitutive, indirectly representing the same phenomenon. 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, emotions are partially captured through syntactic cues such as interjections and adverbs, back-channels are reflected in n-grams and word length etc. Other features turned out to be not predictive at all, such as the paralinguistic information on laughter. Therefore, we examine the ranking of individual features based on information gain, correlation to rating, and classifier weights 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.

Exploration of thinking of students (THINK) benefits from all feature groups. High rated lessons are characterized

by the following features (examples for each and their translations are given in brackets):

- frequent DM of type Reason and Cause on student side (*weil*[because], *so dass*[so that])
- frequent DM of type Comparison and Elaboration on teacher side (*insbesondere*[especially], *das heisst*[that means])
- words of category Communication on teacher side (*sagen*[say], *fragen*[ask], *meinen*[mean], *beschreiben*[describe])
- frequent question words from students (*wie*[how], *wo*[where], *woher*[where from], *warum*[why])
- long student utterances (over 25 words)

We conclude that these features approximate behaviour observable by educational researchers: In highly rated lectures, 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.

Objective and constructive feedback (FEED) is best predicted by the following features:

- frequent affirmative back-channels of both teachers and students (reflected also through n-grams, interjection frequency and phonetic features)
- frequent question words, negations and words from the Discrepancy group (LIWC) on student side (e.g. *Das wollen wir aber nicht, oder?*[We don't want this, do we?]
- DM pair Comparison and Specification on student side (e.g. *oder*[or] ... *beispielsweise*[for example])
- frequent positive words of teacher (e.g. *Ja, gut!*[Yes, good]!)
- longer sentences and larger part-of-speech variety on student side

We hypothesize that these features are indicative of behaviour observed by educational researchers, such as: 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.

Cooperation (COOP) is best predicted through the following features:

- frequent speaker pattern S-SN (student - another student)
- frequent use of *we* rather than *I* on student side and frequent use of *You* on teacher side
- DM pairs 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])
- long student utterances (over 30 words), more frequent verbs and pronouns
- frequent cognition words on student side (e.g. *erkennen*[recognize], *konstruieren*[construct], *wissen*[know])
- frequent communication words on teacher side
- frequent back-channels

These features capture aspects usable by 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 explanations.

5.3. Result summary and discussion

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 markers and semantic word categories show high information gain for predicting such environments, which is a promising path e.g. for a qualitative evaluation of tutoring systems. While we acknowledge that the discourse markers are known to be highly ambiguous [22], e.g. the word *while* can represent a contrast as well as temporal co-occurrence, we believe that our findings open a new route to deeper semantic analysis of discourse structures as predictors of lesson quality.

6. Conclusion and Future Work

Predicting the quality of classroom lessons and analyzing the interaction between teachers and students and among students is an ongoing educational research topic. In this paper, we present initial experiments on the task of assessing three 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, previously unused 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 in 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 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 [28]), which require signal analysis both on the visual and the acoustical part of the data. Our next steps include extending the presented procedure to other dimensions and using automatic speech recognition methods in order to see how stable these methods are in light of noisy, ASR-output.

7. Acknowledgements

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1. The authors thank Prof. Klieme, Dr. Katrin Rakoczy and Petra Pinger for their support with the data and in questions of educational research.

8. References

[1] K. Rakoczy, "Motivationsunterstützung im Mathematikunterricht – Unterricht aus der Perspektive von Lernenden und Beobachtern," Ph.D. dissertation, Johann Wolfgang Goethe-Universität, Frankfurt, Germany, 2006.

[2] G. Gweon, R. Kumar, and C. P. Rosé, "Grasp: The group learning assessment platform," in *Proceedings of the 9th International Conference on Computer Supported Collaborative Learning – Volume 2*, ser. CSCL'09. International Society of the Learning Sciences, 2009, pp. 186–188. [Online]. Available: http://dl.acm.org/citation.cfm?id=1599503.1599566

[3] C. Kersey, B. Di Eugenio, P. Jordan, and S. Katz, "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*. Amsterdam, The Netherlands: IOS Press, 2009, pp. 325–332. [Online]. Available: http://dl.acm.org/citation.cfm?id=1659450.1659501

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

[5] C. P. Rosé, Y.-C. Wang, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer, "Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning," *Computer Supported Collaborative Learning*, vol. 3, no. 3, pp. 237–271, 2008.

[6] J. Hattie and H. Timperley, "The Power of Feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, March 2007.

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

[8] B. Di Eugenio, X. Lu, T. C. Kershaw, A. Corrigan-Halpern, and S. Ohlsson, "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*. Amsterdam, The Netherlands: IOS Press, 2005, pp. 798–800. [Online]. Available: http://dl.acm.org/citation.cfm?id=1562524.1562645

[9] L. Chen, B. Di Eugenio, D. Fossati, S. Ohlsson, and D. Cosejo, "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*, Portland, Oregon, June 2011, pp. 65–75. [Online]. Available: http://www.aclweb.org/anthology/W11-1408

[10] A. Ward and D. Litman, "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, 2007.

[11] D. J. Litman and K. Forbes-Riley, "Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors," *Speech Communication*, vol. 48, no. 5, pp. 559–590, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639305002189

[12] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Commununications*, vol. 53, no. 9-10, pp. 1115–1136, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2011.02.006

[13] K. Forbes-Riley, D. Litman, H. Friedberg, and J. Drummond, "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*, Montréal, Canada, 2012, pp. 91–102. [Online]. Available: http://www.aclweb.org/anthology/N12-1010

[14] F. Lipowsky, K. Rakozcy, C. Pauli, B. Drollinger-Vetter, E. Klieme, and K. Reusser, "Quality of Geometry Instruction and its Short-Term Impact on Students' Understanding of the Pythagorean Theorem," *Learning and Instruction*, vol. 19, no. 6, pp. 527–537, 2009.

[15] K. Rakoczy and C. Pauli, *Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse*, ser. Materialien zur Bildungsforschung. Gesellschaft zur Förderung Pädagogischer Forschung (GFPF)/Deutsches Institut für Internationale Pädagogische Forschung (DIPF), 2006, no. 15, ch. 13, pp. 206–233.

[16] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees," in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[17] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, 2005.

[18] F. Heylighen and J.-M. Dewaele, "Variation in the Contextuality of Language: An Empirical Measure," *Foundations of Science*, vol. 7, no. 3, pp. 293–340, 2002.

[19] M. Wolf, A. B. Horn, M. R. Mehl, S. Haug, J. W. Pennebaker, and H. Kordy, "Computergestützte quantitative Textanalyse," *Diagnostica*, vol. 54, no. 2, pp. 85–98, 2008.

[20] J. Steinberger, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vázquez, and V. Zavarella, "Creating Sentiment Dictionaries via Triangulation," *Decision Support Systems*, vol. 53, no. 4, pp. 689–694, 2012.

[21] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber, "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, 2008.

[22] M. Stede and C. Umbach, "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*, Stroudsburg, PA, USA, 1998, pp. 1238–1242. [Online]. Available: http://dx.doi.org/10.3115/980691.980771

[23] S. Vogel, Y. Zhang, F. Huang, A. Tribble, A. Venugopal, B. Zhao, and A. Waibel, "The CMU Statistical Machine Translation System," in *Proceedings of MT Summit IX,* New Orleans, USA, vol. 9, 2003, pp. 54–63.

[24] B. Han, P. Cook, and T. Baldwin, "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*, 2012, pp. 421–432.

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

[26] V. Nastase, M. Sokolova, and J. S. Shirabad, "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, 2007, pp. 406–410.

[27] E. W. Noreen, *Computer intensive methods for hypothesis testing: An introduction*. Hoboken, NJ: Wiley-Interscience, 1989.

[28] R. Rakov and A. Rosenberg, ""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, 2013.