Exploring Feedback Interactions in Online Learning Environments for Secondary Education

Cecilia Aguerrebere[1], Sofía García Cabeza, Gabriela Kaplan,
Cecilia Marconi, Cristóbal Cobo, and Monica Bulger[2]

[1] Plan Ceibal,
Avenida Italia 6201, 11500 Montevideo, Uruguay
{caguerrebere,sgarcia,gkaplan,cmarconi,ccobo}@ceibal.edu.uy
[2] Data & Society Research Institute
36 W 20th St, New York, NY 10011, United States

Abstract. For decades, teacher feedback has been found to significantly impact student learning. In recent years, research focus has shifted from trying to assess whether feedback is effective to determining whether and how it can be improved. Because of the complexity of the feedback process, the answer to this question is deeply dependent on several factors, such as the learning environment. The goal of this work is to study the feedback process in an online learning environment, in a secondary education setting of learning English as a second language. Using natural language processing techniques we propose to analyze the teacher-student interactions to identify the different types of feedback observed in this learning context, as well as to gain insights on the most effective strategies to improve the students' engagement with the learning process. This article presents the preliminary results of this on-going research.

1 Introduction

As the field of online learning matures, it is now possible to measure key aspects of successful learning and instruction, and to do so at a large scale. For decades, teacher feedback has been found to significantly impact student learning [Brophy and Good, 1970, Dweck, 2002, Hattie and Timperley, 2007]. In recent years, research focus has shifted from trying to assess whether feedback is effective to determining whether and how it can be improved [Van der Kleij et al., 2015]. Because of the complexity of the feedback process, the answer to this question is deeply dependent on factors such as context, learning environment, type and timing of feedback and the task being performed.

The goal of this work is to study the feedback process in an online learning environment, in a secondary education setting of learning English as a second language (ESL). Online learning adoption has witnessed a staggering increase in the last decades, and countless online platforms offer second language learning services. This has motivated several studies on feedback interactions under these settings, which helped gain insight into the feedback process but also raised new questions [Conrad and Dabbagh, 2015, Van der Kleij et al., 2015].

The educational setting to be studied is part of an educational program led by Plan Ceibal[3], an ambitious country-wide one-to-one laptop program deployed in Uruguay. Since 2007, Uruguay has provided Internet access and a laptop to every child and teacher in K-12 public education (about 85% of the student population in the country). Unlike any large-scale laptop program to date [Ames, 2016], it continues to grow and expand its scope. A key part of its success is that the program evolved from its initial goal of reducing the digital divide to providing a wide range of educational digital tools and developing educational programs.

The central research questions of this work are: What are the most relevant types of feedback observed in this online learning context? What feedback characteristics are associated with an increased engagement of the students with the learning activity? Moreover, we are interested in studying how artificial intelligence, more precisely natural language processing (NLP) techniques, can be used to learn the relevant types of feedback, considering aspects such as: the task being performed, the feedback comments content, the timing of feedback, the nature of communication flows and students' task-specific motivation. By answering these questions, we hope to contribute to the better understanding of the effectiveness of the feedback process in general, and to that of online learning in particular. Moreover, this research is particularly relevant because it concerns secondary education settings, which are seldom studied and present a profound need for further research [Van der Kleij et al., 2015]. Furthermore, this research is timely as more online learning environments that include variable quality of feedback are being developed and adopted in secondary educational settings.

This article presents the preliminary results of this on-going research. It is organized as follows. Section 2 summarizes the previous work. Section 3 describes the learning environment where the study is conducted. Sections 4 and 5 introduce the methodology and preliminary results respectively. Conclusions and future work are presented in Section 6.

2 Previous Work

Feedback in educational settings is a highly active research area. Various meta-analyses [Van der Kleij et al., 2015], including the seminal work by Hattie & Timperley [Hattie and Timperley, 2007], show that feedback can be provided effectively, but it is dependent on several factors. [Havnes et al., 2012] found that feedback is more effective if it has a direct use (e.g. correct the task). In a meta-analysis [Van der Kleij et al., 2015] on the effects of feedback in computer based learning environments, elaborated feedback significantly improved higher order learning outcomes. Different feedback classifications have been proposed in the literature. [Hattie and Timperley, 2007] classify feedback in terms of task completion, process, teacher regulation of task, and self. [Shute, 2008] focuses classification on whether the feedback serves to acknowledge the correctness of

[3] `https://www.ceibal.edu.uy/en/institucional`

an answer or provides more elaborated guidance. [Brown et al., 2012] classify feedback based on the teachers objectives: to improve learning, to report and compliance, or to motivate students. To [Van der Kleij et al., 2015], expressing praise for the student did not improve the learning outcomes, but improved the students' motivation and perseverance. [Van der Kleij et al., 2015] recommend that future research take into account the precise characteristics of feedback, the task, the learning context and the learners when assessing feedback effectiveness, particularly for secondary education settings.

To be effective, feedback needs to be processed mindfully, especially in online environments where students can easily ignore written feedback [Timmers and Veldkamp, 2011]. Regarding feedback timing, a meta-analysis by [Van der Kleij et al., 2015] found that students spent more time reading immediate feedback and that immediate feedback related to improved learning outcomes.

A majority of research on feedback has occurred in higher education settings, [Evans, 2013, Van der Kleij et al., 2015], leaving unanswered questions about its effectiveness in secondary educational settings, even as its practice expands. Prior research focuses on categorizing teachers' strategies for providing feedback in secondary schools [MacDonald, 2015], student-generated feedback in K-12 [Harris et al., 2015], the effects of expectancy-incongruent feedback on task performance [Baadte and Kurenbach, 2017], among other more general examples [Oinas et al., 2017].

3 Learning Environment

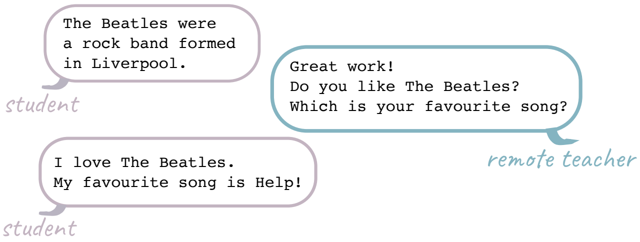
The Beatles were a rock band formed in Liverpool.
student
Great work! Do you like The Beatles? Which is your favourite song?
remote teacher
I love The Beatles. My favourite song is Help!
student

Fig. 1. Example of a student-teacher interaction in the TDL program.

This study will be conducted in the context of one of Plan Ceibal's educational programs for ESL teaching, the Tutorials for Differentiated Learning (TDL). The TDL are a series of resources and exercises for ESL learning, with varying complexity, created to support secondary school students by proposing activities tailored to their needs. The TDL are available online, through a Learning Management System (LMS), and students are encouraged to explore and complete exercises at their own pace. A remote teacher (RT) provides individualized feedback to students by posting comments on the LMS. The student-RT interactions under consideration follow the pattern: 1) the student posts a comment as a response to an exercise; 2) the RT posts a comment giving feedback to the student; 3) the conversation may or may not continue between the student and the RT. All conversations include at least two comments, the student's comment initiating the conversation and the teacher's reply, and in the best case scenario they include subsequent interactions. All interactions occur in discussion forums public to the class within the LMS. Figure 1 illustrates an example of these interactions.

4 Methodology

We propose to use NLP techniques to analyze the student-RT interactions, to identify the different types of feedback observed in this learning context, as well as to gain insights on the most effective strategies to improve the students' engagement with the learning process. For this purpose, each comment is mapped into a vector of features, composed of numerical and categorical variables, designed to represent relevant aspects of the feedback interaction. The proposed features are listed in Table 1. Two types of features can be differentiated: those depending on the teacher only (e.g., *timeToResp*, *complexity*, *asksQuestion*) and those dependent on other factors (e.g., *likes*, *origStResp*, *threadLength*). The former are of particular interest as they can be used to infer *recommendations* of teacher's actions associated with desired results.

The dataset under consideration includes 5073 comments posted by 20 teachers, from interactions with 520 students organized in 41 groups. Each comment includes: the comment content (text), the timestamp and the number of *likes* received, corresponding to the 2017 edition of the TDL program.

5 Preliminary Results

We present here a summary of the results obtained in a preliminary analysis of the dataset under consideration.

5.1 General dataset description

Figure 2 shows the average number of words per comment as a function of the total number comments for each user. Most students wrote under 50 comments in the year, except for some outliers who wrote above 150. Most RTs made under 500 comments, except for one teacher who made over 1600 comments. The vast majority of students made less than 50 comments (average 12.4) with less than 10 words per comment (average 8.3). There is a small group of students, more *active* than the rest, who wrote between 100 and 200 comments with 8 to 15 words, and another small group who wrote much longer comments (over 20 words per comment). Teachers' behavior is more diverse regarding comments length. There is a group of RTs with 5 to 15 words per comment on average, and another group with 20 to 30. The average words per comment of the teacher who made over 1600 comments is 5.

Table 1. Features defined to characterize each comment.

Numerical Features	
Variable	**Description**
`timeToResp`	Timelapse between the student's comment and the teacher's response (in days).
`sent_len`	Total number of sentences in the comment.
`likes`	Total number of *likes* assigned to the comment.
`actionability`	Comment's sentences are classified according to their grammatical mood into: indicative and non indicative. The `actionability` is the percentage of indicative sentences in the comment.
`sentiments`	Comment's sentences are assigned an index based on the adjectives they contain (e.g., good, bad, amazing, irritating), taking values -1 (negative) to +1 (positive). The `sentiments` feature is the average of this index among the comment's sentences.
`complexity`	Measures the complexity of the comment, operationalized as the automated readability index (combining number of sentences, words and characters).
`specificity`	How deep each word appears in the Wordnet structure.
`threadLength`	Length of the conversation (i.e., total number of posts in the conversation).
`totalComms`	Total number of comments posted by the user (used as a proxy of the user's activity intensity level)
Categorical Features	
Variable	**Description**
`hasLink`	Boolean variable taking value 2 if the comment includes a link (usually sharing information) and 1 otherwise.
`hasEmoticon`	Boolean variable taking value 2 if the comment includes an emoticon and 1 otherwise.
`onlyEmoticon`	Boolean variable taking value 2 if the comment **is only an emoticon** and 1 otherwise.
`asksQuestion`	Boolean variable taking value 2 if the comment includes a question to the student and 1 otherwise.
`origStResp`	Boolean variable taking value 2 if the comment originated a student's response and 1 otherwise.

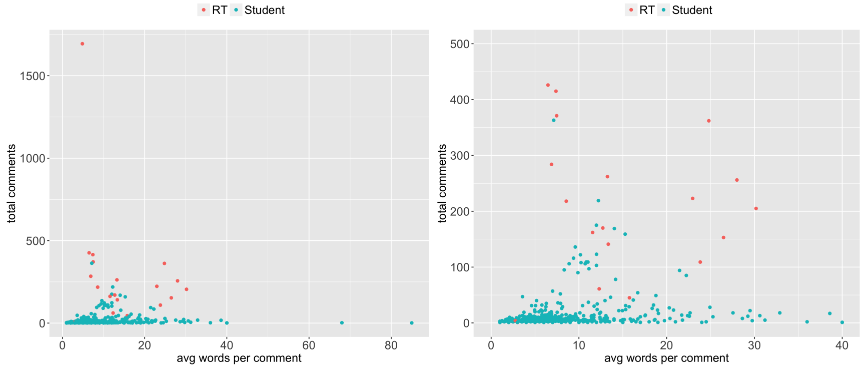

Fig. 2. Left: Average words per comment vs total comments for each user. Red points represent RTs and green points represent students. **Right:** Zoom of the image on the left.

5.2 Principal Component Analysis

Principal component analysis (PCA) for mixed type data [Chavent et al., 2017] is used to jointly study the numerical and categorical features. Numerical features are normalized to have zero mean and unitary standard deviation. Figure 3 shows the cumulative percentage of variance explained by the eigenvalues. Ten out of the original 13 components are needed to cover 90% of the samples variance, indicating that a large dimensionality reduction is not possible. Note that the feature `onlyEmoticon` is not considered as none of the comments in the dataset corresponded to emoticons only.

Figure 3 shows the correlation circle (correlation of each variable with the first two principal components) and the levels map (gives an idea of the pattern of proximities between the levels of different categorical variables) [Chavent et al., 2017]. The comment length is negatively correlated to the sentiment variable, as well as to the total number of comments done by the teacher, suggesting that teachers who post many comments tend to write shorter ones. These variables differentiate the comments in the first PCA component. The thread length is negatively correlated with the response time, suggesting that the chances of engaging the student in a conversation decreases as the teacher takes longer to respond. Longer conversation threads are associated with more *likes*, as well as with higher actionability and complexity although to a lesser degree. The levels map shows that comments without emoticon, that don't share a link and don't ask questions are more associated with no response from the student.

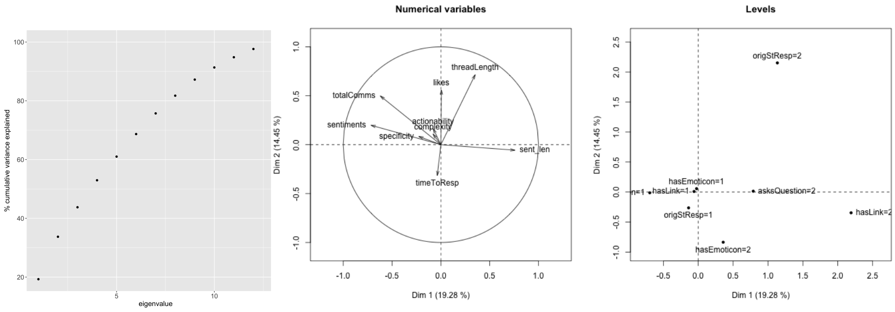

Fig. 3. Left: Cumulative proportion of variance explained by each eigenvalue. **Center:** Correlation circle. **Right:** Levels map for categorical variables.

A PCA analysis was conducted on teacher-dependent variables only (c.f. Section 4). Figure 4 shows the first two principal components of all comments, where the color in the top figure indicates whether the teacher's comment received a student's response (green) or not (red), whereas in the bottom figure it indicates the conversation length. The `origStResp` and `threadLength` variables were not used to build the PCA projection in this case. It can be observed that the first two PCA dimensions are not enough to predict the conversations length or whether the student continued the conversation or not.

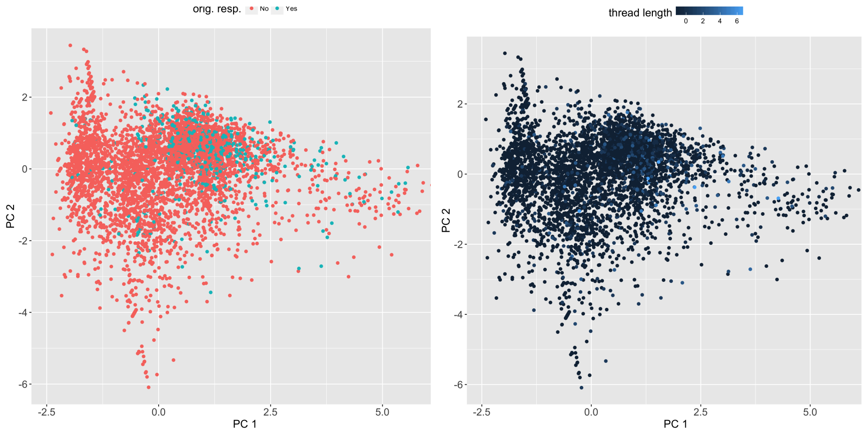

Fig. 4. PCA representation of all samples (conducted on teacher-dependent variables only). **Left**: The color indicates the number of sentences in the comment. **Right**: The color indicates the sentiments.

5.3 Features validation

The features defined in Section 4 will be used to characterize the feedback types observed in the TDL program. Because one of the goals of the program is to foster the student's integration in the English culture, an interaction student-RT is considered successful if, among other factors, it gets the student involved in a conversation. Hence, we would like the proposed features to be informative of whether a given feedback comment will originate a student response or not, and even further, to be predictive of the length of the conversation thread. In order to assess this, the defined features are used to train a classifier and predict whether the student will reply to the RT feedback comment or not. If the defined features are informative of the probability of a student following the conversation, this classifier will perform better than chance.

Experimental setup Three classification strategies are considered: decision trees, random forests and boosting trees. Decision trees are widely used because they are simple and easy to interpret. However, they are known to be outperformed by their counterparts which combine several trees: random forests and boosting trees. The latter increase performance at the cost of reduced interpretation. Nevertheless, they have variable importance definitions that shed light into the most relevant variables for classification, thus helping understand the latent phenomenon.

The three classifiers are trained to predict whether the student will continue the conversation or not (i.e., the feature `origStResp`) using the features that depend solely on the teacher (`timeToResp`, `sent_len`, `actionability`, `sentiments`, `complexity`, `specificity`, `hasLink`, `hasEmoticon` and `asksQuestion`).

The dataset is divided into a training and a testing subset, used to train and evaluate the algorithms respectively. The division is performed by randomly sampling the teachers, so as to ensure that the samples in the training and testing datasets are independent.

Two outliers are identified in the dataset: a student who posted 234 comments in a group with almost no activity (only two other students in his group posted 29 and 12 comments), and a group of 21 students who posted 1494 comments in total, with a median of 69 comments per student. Hence, we study classification performance using the complete dataset, the dataset removing the outliers, and also the behavior of the *active* group alone.

Results Table 2 summarizes the classification performance of the three evaluated approaches, measured by the area under the ROC curve (AUC), the true positive rate (TPR) and false positive rate (FPR), in each of the studied datasets. In all the tested configurations, the classifiers performance is above chance (TPR=0.5, FPR=0.5), showing the predictive power of the proposed features. The AUC results are not reported for the decision trees case because the implementation under consideration only reported the assigned labels (as opposed to the probability of belonging to each class required to build the ROC curve).

Table 2. True positive rate (TPR) and false positive rate (FPR) of the three evaluated classification approaches.

	decision trees		**random forest**			**boosting trees**		
dataset	TPR	FPR	TPR	FPR	AUC	TPR	FPR	AUC
complete	0.57	0.39	0.80	0.53	0.65	0.70	0.38	0.75
no outliers	0.56	0.45	0.81	0.54	0.67	0.64	0.52	0.71
active group	0.85	0.14	0.91	0.15	0.89	0.89	0.11	0.91

Decision trees: Despite performing better than chance, the variance of the decision trees obtained for different realizations of the training set is very large. In the case of the complete dataset the `asksQuestion` and `timeToResp` variables are the most important ones, whereas when removing outliers other variables also turn up as relevant (see Figure 5). Performance highly improves for the *active* group evaluation.

Random forests: Perform much better than decision trees. The `timeToResp` and `asksQuestion` features are always among the most relevant features.

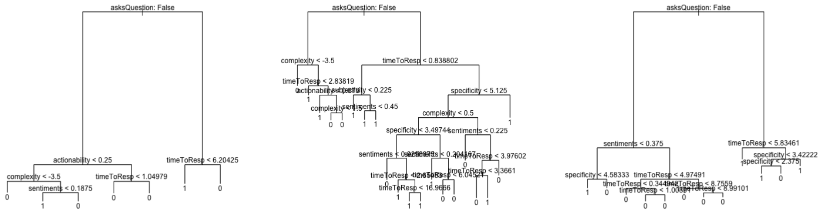

Fig. 5. Examples of the decision trees obtained for the different dataset samples. Whether the teacher asked a question and the response time are systematically among the most relevant variables for predicting the student response.

Boosting trees: Is the most stable of all tested alternatives, always with similar performance and most important variables, among which we find `timeToResp`, `specificity`, and `sentiments`.

6 Conclusions and Future Work

In this article, we presented the main ideas and preliminary results of an ongoing work to study feedback interactions in an online learning environment for secondary education. Despite exploratory, the first results already show the power of the proposed features to represent relevant aspects of the educational program, such as the probability of engaging the student in the conversation. This particular point is essential, as getting the student to continue the discussion with the RT encourages him to practise further. Classical machine learning classifiers show very good performance for student's response prediction with the defined features. The most relevant features vary with the approach, but the time the teacher takes to respond and whether he asks a question to the student are important variables for all methods. Longer response times are associated with no response from the student, thus ending the conversation with just one interaction. On the other hand, an important factor to increase the probability of engaging the student in the conversation is asking a question.

As future work, we will continue the exploration of the association of the defined features with students' behavioral aspects that are relevant for the TDL program. In addition to the students' response rate already analyzed, possible options could be the frequency of the students interaction with the TDL material (i.e., sporadic versus frequent) and the diversity of material consulted (i.e., how many different topics were consulted by the student). We hope this analysis will shed light into the given feedback process and guide the characterization of the different existing feedback interactions.

Bibliography

M. G. Ames. Learning consumption: Media, literacy, and the legacy of one laptop per child. *The Inform. Soc.*, 32(2):85–97, 2016.

C. Baadte and F. Kurenbach. The effects of expectancy-incongruent feedback and self-affirmation on task performance of secondary school students. *Eur. J. Psychol. Educ.*, 32(1):113–131, 2017.

J. E. Brophy and T. L. Good. Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *J. Educ. Psychol.*, 61(5):365, 1970.

G. T. Brown, L. R. Harris, and J. Harnett. Teacher beliefs about feedback within an assessment for learning environment: Endorsement of improved learning over student well-being. *Teach. and Teach. Educ.*, 28(7):968–978, 2012.

M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. Multivariate analysis of mixed data: The r package pcamixdata. 2017.

S. S. Conrad and N. Dabbagh. Examining the factors that influence how instructors provide feedback in online learning environments. *International Journal of Online Pedagogy and Course Design*, 5(4):47–66, 2015.

C. S. Dweck. Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). 2002.

C. Evans. Making sense of assessment feedback in higher education. *Rev. Educ. Res.*, 83(1):70–120, 2013.

L. R. Harris, G. T. Brown, and J. A. Harnett. Analysis of new zealand primary and secondary student peer-and self-assessment comments: Applying hattie and timperleyś feedback model. *Assessment in Education: Principles, Policy & Practice*, 22(2):265–281, 2015.

J. Hattie and H. Timperley. The power of feedback. *Rev. Educ. Res.*, 77(1): 81–112, 2007.

A. Havnes, K. Smith, O. Dysthe, and K. Ludvigsen. Formative assessment and feedback: Making learning visible. *Stud. Educ. Eval.*, 38(1):21–27, 2012.

V. A. MacDonald. *The application of feedback in secondary school classrooms: Teaching and learning in applied level mathematics.* PhD thesis, University of Toronto (Canada), 2015.

S. Oinas, M.-P. Vainikainen, and R. Hotulainen. Technology-enhanced feedback for pupils and parents in finnish basic education. *Computers & Education*, 108:59–70, 2017.

V. J. Shute. Focus on formative feedback. *Rev. Educ. Res.*, 78(1):153–189, 2008.

C. Timmers and B. Veldkamp. Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education*, 56(3):923–930, 2011.

F. M. Van der Kleij, R. C. Feskens, and T. J. Eggen. Effects of feedback in a computer-based learning environment on students learning outcomes: A meta-analysis. *Rev. Educ. Res.*, 85(4):475–511, 2015.