How Should Online Teachers of English as a Foreign Language (EFL) Write Feedback to Students?

Cecilia Aguerrebere
Fundación Ceibal, Uruguay
caguerrebere@ceibal.edu.uy

Monica Bulger
Future of Privacy Forum
mbulger@fpf.org

Cristóbal Cobo
Fundación Ceibal, Uruguay
ccobo@ceibal.edu.uy

Sofía García
Fundación Ceibal, Uruguay
sgarcia@ceibal.edu.uy

Gabriela Kaplan
Plan Ceibal, Uruguay
gkaplan@ceibal.edu.uy

Jacob Whitehill
Worcester Polytech. Inst., USA
jrwhitehill@wpi.edu

ABSTRACT

We analyze teachers' written feedback to students in an online learning environment, specifically a setting in which high school students in Uruguay are learning English as a foreign language. How complex should teachers' feedback be? Should it be adapted to each student's English proficiency level? How does teacher feedback affect the probability of engaging the student in a conversation? To explore these questions, we conducted both parametric (multilevel modeling) and non-parametric (bootstrapping) analyses of 27,627 messages exchanged between 35 teachers and 1074 students in 2017 and 2018. Our results suggest: (1) Teachers should adapt their feedback complexity to their students' English proficiency level. Students who receive feedback that is too complex or too basic for their level post 13-15% fewer comments than those who receive adapted feedback. (2) Feedback that includes a question is associated with higher odds-ratio (17.5-19) of engaging the student in conversation. (3) For students with low English proficiency, slow turnaround (feedback after 1 week) reduces this odds ratio by 0.7. These results have potential implications for online platforms offering foreign language learning services, in which it is crucial to give the best possible learning experience while judiciously allocating teachers' time.

1. INTRODUCTION

For decades, teacher feedback has been shown to be one of the greatest drivers of student learning [9]. The research focus has shifted from assessing whether feedback is effective to identifying the most powerful strategies [20]. Because of the complex nature of the feedback process, the answer to this question remains deeply tied to the particular context in which it happens. One particular learning domain that is growing fast in terms of number of learners and learning platforms (e.g., Duolingo, Babbel, Learning English at Coursera) is online learning of English as a foreign language (EFL). Despite its increasing prominence, teacher feedback in the online EFL context has received limited research attention [20, 5].

In this paper we seek to contribute to the understanding of how teacher feedback influences students' behavior in the online EFL context. In particular, we focus on an EFL program in which students learn English with the help of a remote teacher (RT), who is a native English speaker, with whom students communicate online using discussion forums. Within this context, we seek to build an understanding of how the feedback the RTs give to their students affects their posting behavior: (1) How complex should the RT feedback be? (2) Should it be somehow adapted to their student's English proficiency level? (3) How does RT feedback affect the probability of engaging the student in a conversation? This research has potential implications for the countless online platforms offering foreign language learning services, aiming to enhance students' learning experience.

Learning environment: This study is conducted in the context of a program for EFL learning created for secondary school students who attend the public school system in Uruguay. Uruguayan secondary school students (native Spanish speakers) often struggle with English, having very disparate proficiency levels when they enter high school. This program, known as Tutorials for Differentiated Learning (TDL), was conceived to help tackle this problem by providing the students with the option to learn and practice English at their own pace. For this purpose, a set of resources and exercises for EFL learning are made available online through an LMS system.The students are encouraged by their classroom English teachers (CT) to explore the material and complete the exercises, but participation in the program is not mandatory. Completing an exercise consists of reading the material and posting a comment in English in a discussion forum. Exercises are organized in topics (e.g., music, sports, fashion, national parks, travel, etc) and there is one discussion forum per exercise. A RT, assigned to each classroom, reviews the students' posts and gives them individualized feedback.

RT-student interactions: The student always starts the thread by posting a comment about a given topic in the LMS discussion forum. The RT replies giving the students personalized feedback on what they wrote. Then, the conversation may or may not continue depending on whether the student posts a new comment on the given thread. If the

Cecilia Aguerrebere, Monica Bulger, Cristóbal Cobo, Sofía García, Gabriela Kaplan and Jacob Whitehill "How Should Online Teachers of English as a Foreign Language (EFL) Write Feedback to Students?" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019),* Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 234 - 239

student doesn't, that conversation ends there, and the student may start new threads when doing new exercises. Here is an example of an interaction where the student engaged in the conversation:

Student: *I do not have favorite music I like to listen to everything a little.*
RT: *That's great Alicia. What's your favourite song right now?*
Student: *At this moment I've heard a song from Michael Jackson that I loved its name is Thriller.*
RT: *Ok Alicia, thank you for sharing that :)*

and another example where she didn't:

Student: *I see six oceans: Atlantic, Indian, Pacific, Atlantic, Arctic and Southern ocean.*
RT: *Very well Andrea.*

Learning English: In TDL, the RTs' feedback is intended less as a way of correcting students' mistakes and more as a way to encourage students to participate. Participation in the discussion forums is expected to be conducive to better learning since doing the exercises requires *reading* the material in English as well as *writing* the response in English. Therefore, two measures of interest are: the total comments the student posts and whether the student engages in a given conversation with the RT.

2. PREVIOUS WORK

Even though there is no unified definition of feedback, the seminal work by Hattie and Timperley [9] conceptualizes feedback as *information provided by an agent regarding aspects of a student's performance or understanding.* It can be provided effectively, but it is dependent on several factors such as the task, the learning context and the learners [9, 20]. It may improve learning outcomes when it has a direct use (e.g. correct the task), or it may increase motivation when only expressing praise for the student [20].

In the online language learning context, feedback has been reported as a fundamental aspect in skills development [11]. Teacher feedback in online language learning environments can also inform development of data-driven personalized feedback. Emerging data-driven learning systems adapt feedback to individual student needs, and have been shown to improve learning outcomes [17]. Furthermore, data mining has been used to understand the effects that polarity (positive vs. negative comments) and timing can have in different student's learning aspects [12, 16].

Research on feedback for EFL learning in computer-mediated (CM) environments has widely focused on *peer* feedback, often on EFL writing [18]. Jiang and Ribeiro [10] present a systematic literature review on the effect of CM peer written feedback on adult EFL writing. They confirmed the findings from previous research acknowledging the positive impact of CM peer feedback in this context.

As in many other subjects in educational data mining, most research on feedback has focused on higher education settings [20], leaving the fundamental need to build understanding of the primary and secondary education contexts unattended. We find previous work on secondary and primary education contexts on particular topics such as teachers' feedback strategies [3], student-generated feedback [8] among other more general examples [15].

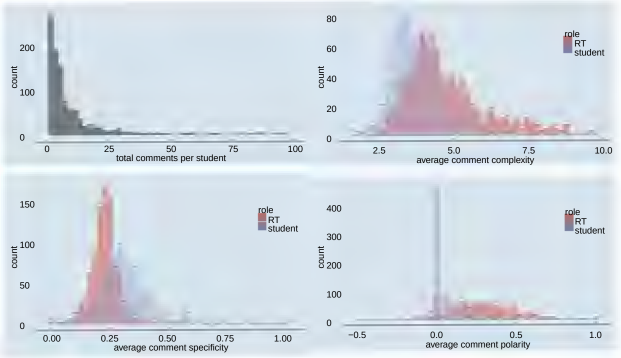

Figure 1: Histogram of the total comments posted by each student (top left). Histograms of the average complexity (top right), specificity (bottom left) and polarity (bottom right) of the RTs' feedback (red) and of the students' posts (violet).

Our work complements and enriches the previous work in several aspects: (1) it studies asynchronous teacher feedback in an online EFL environment, which has been seldom studied [19], (2) it considers a secondary education setting, also fundamental and rarely analyzed [20], (3) it follows a quantitative analysis exploiting a large scale dataset (in terms of number of students, teachers, classrooms, and school diversity) as opposed to most case studies which often include relatively few students or classrooms [19].

3. DATASET DESCRIPTION

The dataset under consideration was originally collected by Aguerrebere et al. [2] and includes all the comments (i.e., content, posting date, user id) as well as administrative information (for each student, who are her CT and RT) for the 1st secondary school classrooms (12-year-olds) that participated in the TDL program during school year 2017. In this work the dataset is extended to also include school year 2018. This includes a total of 27,627 comments exchanged between 1074 students, organized in 83 classrooms (in 49 public high schools located in 18 different states in the country), and 35 RTs. The dataset has a nested structure: students are organized into classrooms. Each classroom has a dedicated classroom teacher; in contrast, each remote teacher may serve multiple classrooms. Figure 1 shows the histogram of the total comments posted by each student during their corresponding school year. The dataset has been de-identified to preserve each participant's privacy and handled according to Uruguayan privacy protection legislation. After talking with the TDL stakeholders and the program leaders, a set of features characterizing each comment was defined: *complexity*, *specificity*, *polarity* and *response delay*. Each feature represents a different aspect of how elaborate a comment is (*complexity*, *specificity*), its tone (*polarity*) and how long the student had to wait to receive feedback (*response delay*).

Complexity (c) measures how elaborated a comment is, by adding its characters per word, words per sentence and total sentences: $c = \frac{1}{4}\frac{\#char}{\#words} + \frac{1}{5}\frac{\#words}{\#sent} + \#sent$ (weights are included to give similar relevance to all terms, 4 and 5 are the median characters per word and words per sentence respectively). Examples of low, medium and high complexity comments are: ($c = 2.4$) "*Well done!*", ($c = 3.7$) "*My favourite national park is Yellowstone.*", ($c = 8.9$) "*Hi Alberto! This is an accurate description of the different continents, but can you try again? The activity is asking about different volcanic landforms! Can you please look at the encyclopedia and read the part about volcanic landforms to find the names of the three types of volcanic landforms? Here's the link: [link]*".

Specificity (s) measures how specific, on average, the words are in the comment. It combines how deep each word w_i appears in the WordNet [13] structure and how frequent the word is in the dataset: $s = \frac{1}{W}\sum_{i=1}^{W} \frac{depth(w_i)}{Z} + \frac{1}{freq(w_i)}$, where W is the total words in the comment and Z a normalizing factor equal to the maximum average comment complexity [6]. Examples of comments with low and high specificity: ($s = 0.1$) "*Very good! Do you have any cats?*" and ($s = 1.2$) "*The skeleton of brontosaurus.*" .

Polarity (p) measures the tone of the comment (positive, negative) as the average of an index (-1 (negative) to +1 (positive))[1] assigned to each sentence based on the adjectives it contains (e.g., great, nice, awful). Examples of positive and negative comments: ($p = 1.0$) "*Great Carla! Awesome spelling!!*" and ($p = -0.6$) "*I would not like because they are dangerous.*".

Response delay (τ) is the timelapse between the student's post and the RT's response in days. Median τ is 3 days.

Figure 1 shows the histograms of the the average *complexity*, *specificity* and *polarity* of the RTs' feedback and the students' posts. The average specificity tends to be larger for students than for RTs; this is likely because students' comments are responses to exercises that elicit very specific words such as "*skeleton of brontosaurus.*", whereas the RTs' comments often contain basic words, e.g., "*Great work!*".

4. DATA ANALYSIS

We examine the effects of various feedback characteristics of RTs' feedback on students' posting behavior. We use two complementary approaches [14]: (1) multilevel linear regression that models the nested nature of the data; and (2) non-parametric bootstrap analysis. The latter is more complicated (e.g., requires a bin width parameter) but can model non-linear relationships and makes fewer assumptions (e.g., normality of residuals) than many parametric models.

4.1 How complex should the RT feedback be?

In this section we are interested in the question: Does RT feedback complexity (low vs. high) affect the total number of comments posted by the student? Note that we must consider the potential confound of the student's English proficiency level, as the effect of RT feedback complexity may vary for more or less proficient students.

4.1.1 Non-parametric approach

To answer this question we first follow a non-parametric approach, using bootstrap to test the null hypothesis:

$$H_0 : \quad E[T|S_c] = E[T|S_c, R_c], \tag{1}$$

versus $E[T|S_c] \neq E[T|S_c, R_c]$, where T is the total comments posted by the student, S_c is the student's English proficiency level (low/high) and R_c is the complexity level of the feedback the student received from his RT (low/high). Hypothesis (1) tests whether the total comments posted by the student depend on the complexity level of the feedback she received from her RT, after conditioning on her English proficiency level. If the null hypothesis is rejected, then the complexity level of the feedback that the students receive affects their average engagement with the program, measured by the total comments they post and conditioned on their English proficiency level. It is important to note that we must reject the null hypothesis if $E[T|S_c]$ differs statistically significantly from $E[T|S_c, R_c]$ for *any* values of S_c and R_c. For this reason, in this section we examine the potential impact of R_c on T for each possible combination of (S_c, R_c).

To estimate the student's English proficiency level we use the average complexity of the comments posted by the student. Students with average complexity below (above) the median are classified as having low (high) proficiency respectively. Similarly, the complexity level of the RT's feedback is low (high) when it is below (above) the median.

Bootstrapping for equality of means: How can we test whether $P(T|S_c, R_c = low)$ and $P(T|S_c, R_c = high)$ have the same mean? If the distribution $P(T|S_c, R_c)$ were Gaussian for all values of S_c and R_c, then we could just use a t-test to compute the p-value (or a nested ANOVA to take into account the nested structure of the data). However, in our case the data are not Gaussian (see [1]). Fortunately, the bootstrap procedure proposed by Efron & Tibshirani [7] provides a rigorous methodology. By sampling *with replacement* from our original dataset, we can *simulate* multiple data samples. We subselect the data for which $R_c = low$ and the data for which $R_c = high$ and then resample each of them to generate multiple bootstrap samples. To enforce the null hypothesis (i.e., equal means), we *set* the means of the two samples to be equal to the mean of the combined sample. We then compute the normalized difference in means between the two subsets in the bootstrapped sample. Over all B bootstrap iterations ($B = 10000$), we finally compute the fraction in which the normalized difference in means is at least as large as the observed statistic. See [1] for a detailed description of the algorithms.

Results: For high level students, more *basic* feedback is associated with more posting. Students who received high level feedback posted on average 22% fewer comments than those who received low level feedback (9.3 versus 12 total comments, $p = 0.006$). Examples:

Student A: *My favorite food is hamburguer.*

RT (low level): *Nice Lucia! What do you eat in your hamburger?*

Student B: *They are in Spain in Barcelona.*

RT (high level): *Hello Pablo! Yes they are in Spain. Very good. Here is a link in case you would like to know a bit more about Spain and their culture. In Uruguay there are a lot of people who have Spanish origins. It is very evident in the food :) I have never been to Spain. Would you like to go to Spain?*

A possible explanation for this behavior is that even the *high-level* students have weak English proficiency and might feel overwhelmed by feedback that is too complex. No statis-

[1] The *sentiment* function of the PATTERN.EN Python module was used: `www.clips.uantwerpen.be/pages/pattern-en`.

tically significant differences are observed for low level students (8.5 versus 7.6 total comments, $p = 0.14$).

4.1.2 Controlling for the RT

Another possible confound is the effect of the RT, as more motivated RTs may give better feedback that leads to more posts by their students. To take this into account, null Hypothesis (1) is reformulated as:

$$H_0 : \quad E[T|S_c, RT] = E[T|S_c, R_c, RT], \tag{2}$$

versus $E[T|S_c, RT] \neq E[T|S_c, R_c, RT]$, where RT is the student's remote teacher. If Hypothesis (2) is rejected, the complexity level of the feedback *a given RT* gives to his students has an effect on the total comments posted by them. A variation of the bootstrapping algorithm is used to test Hypothesis (2) where, instead of defining the $[low, high]$ RT feedback complexity levels globally from all samples, independent thresholds are defined for each RT.

Results: The result previously obtained remains valid even when conditioning on the RT (i.e., resampling within each RT), meaning that different feedback levels given by the RT to his students are associated with different total posts (8.9 versus 12.5 total comments, $p = 0.01$). No statistically significant differences are observed for low level students.

Even if controlling for the RT makes the result more *solid* from a statistical point of view, interpreting the results becomes more difficult, as the meaning of *low* and *high* complexity feedback changes from RT to RT. Because a fundamental goal is to translate these results into useful information for the teachers, this approach has the disadvantage that an *absolute* complexity level reference cannot be given to them as reference of what *low* and *high* means, and how to position the feedback they give with respect to that.

4.1.3 Parametric Approach

A parametric approach is conducted to complement the results obtained by the non-parametric analysis, allowing for the inclusion of additional predictors (possible confounds) and avoiding binning (in RT complexity). Binning is kept to determine student level (low/high), as separate models are computed for each case. In order to take into account the nested structure of the data, a multilevel modeling approach is employed where the CT and RT effects on the student's activity are modeled as nested random effects. Therefore, we model student i's total posts as a negative binomial random variable, to account for the fact that it is count data with overdispersion, with expected value μ_i given by:

$$\log(\mu_i) = \beta + \gamma_0 c_i + \gamma_1 y_i + \gamma_2 p_i + \gamma_3 s_i + \gamma_4 \tau_i + C + R, \tag{3}$$

(capital letters denote random variables and lower-case denote fixed values). β is the baseline total comments. c_i, p_i, s_i and τ_i are the average complexity, polarity, word specificity and response delay of the feedback comments student i received from his RT, respectively. y_i is the school year. The fixed effects $\gamma_0, \ldots, \gamma_4$ represent the effects of the corresponding covariates on the total comments. The nested random effect CT-RT is represented by the random variables C and R, assumed to follow zero-mean Gaussian distributions with standard deviations σ_C and σ_R. All the parametric models were fit using the R *lme4* package [4].

	stud. level	β		$\boldsymbol{\hat{\gamma}_0}$		$\hat{\gamma}_1$		$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$
Model 3	**low**	1.12	**	0.03		0.30	.	-0.03	0.76	0.01
	high	2.50	***	**-0.11**	*	0.26		-0.02	-0.02	0.04
Model 4	**low**	1.63	***	**-0.14**	**	0.25		-0.32	0.67	-0.01
	high	2.24	***	**-0.16**	**	0.22		-0.06	0.12	-0.002

Table 1: Effects of RT feedback on students' total comments for Models 3 and 4, in log scale. Signif. codes: 0 (***) 0.001 (**) 0.01 (*) 0.05 (.) 0.1

Results: Table 1 (Model 3) shows the computed effects for all the covariates. The parametric analysis, which includes other possible confounds, confirms the same tendency observed with the non-parametric approach: a negative statistically significant effect of the RT feedback complexity level is observed for high level students ($\exp(-0.11) = 0.9$, i.e., 10% less total posts per unit increase in RT average complexity) and no effect is observed for low level students. To compare this result to the one obtained by the non-parametric approach, we compute the equivalent per unit decrease in the non-parametric case (computed as the total decrease divided by the difference between the average RT complexity in the two compared levels, 3.8 and 5.9) which equals 10%.

4.2 Should RTs feedback complexity be close to that of their students?

Rather than the *absolute* complexity, we can also consider the *relative* complexity of the RTs' feedback compared to the complexity of students' comments. Put another way: should the feedback complexity be somehow *adapted* to the student? To answer this question we propose to model the total comments posted by the student as a function of the *distance* between the average complexity of the student's comments and the average complexity of the feedback the student received from his RT.

4.2.1 Parametric approach

We model each student i's total posts as a negative binomial random variable with expected value μ_i given by:

$$\log(\mu_i) = \beta + \gamma_0 |c_i - c_{s_i} - \alpha| + \gamma_1 y_i + \gamma_2 p_i + \gamma_3 s_i + \gamma_4 \tau_i + C + R, \tag{4}$$

The fixed effect γ_0 represents the effect of the absolute value of the difference between the student's (c_{s_i}) and the RT's (c_i) average comments complexity, where α is introduced as an offset to account for the fact that the feedback may need to be close to that of the student but not necessarily equal. See Model 3 for the definition of the rest of the variables.

Setting α: Model 4 is fitted for different values of α and the one corresponding to the largest log-likelihood is selected. For low student levels the maximum log-likelihood is obtained at $\alpha = 0.25$, whereas for high student level it is at $\alpha = -0.79$. Hence, this analysis suggests that even if for both low and high level students feedback complexity should be close to the student level, low level students benefit from feedback slightly more complex than theirs whereas high level students benefit from feedback slightly below theirs. Recall that low level students post very basic comments, whereas those of high level students tend to be more elaborated but remain still simple.

Results: Table 1 (Model 4) shows the computed fixed effects for the different covariates. The distance between the average complexity of the student's comments and the average complexity of the feedback the student received from his RT has a negative stat. sig. effect on the total comments posted by the student. There is a 13% ($p = 0.008$) and 15% ($p = 0.003$) decrease in total comments with one unit distance increase, for the low and high level students respectively. We present in the following a series of examples in order to help the reader gain insight into what small and large student-RT complexity distance mean in practice.

Large distance:

Student A: *I would like to defile.*

RT: *Nice try Marcela, but I do not understand what you mean. There are two models in the above photo, can you tell me which model is from Brazil and which model is from the USA? or - can you tell me, who is your favourite model? My favourite model is Kate Moss.*

Small distance:

Student A: *My favorite sport is football.*

RT: *Very good! what is your favorite football team?*

In the latter example, the interactions seem closer to an online chat for students with basic English skills.

4.2.2 Non-parametric approach

A non-parametric approach is conducted to complement the results obtained by the parametric analysis. For this purpose, bootstrapping is used to test the null hypothesis:

$$H_0 : \quad E[T|S_c] = E[T|S_c, D], \tag{5}$$

versus $E[T|S_c] \neq E[T|S_c, D]$, where D is the distance between the student's and the RT's average comments complexity as defined in Model 4. The bootstrapping algorithm introduced in Section 4.1.1 is used, with a for loop on small and large D (below and above the median distance) instead of RT complexity, and α is set to the values obtained in Section 4.2.1 (see [1] for details).

Results: A negative statistically significant effect is observed for D on the total comments posted by the student, both for low and high level students. Students who received feedback less adapted to their level posted 15% and 34% less comments than those who received more adapted feedback, for low (6.6 versus 7.6 total comments, $p = 0.007$) and high (9.2 versus 12.3 total comments, $p < 0.001$) level students respectively. To compare these results to those obtained by the parametric approach we compute the equivalent per unit decrease (computed as the total decrease divided by the difference between the average D in the two compared levels) which equals 9% both for low and high student level.

4.3 Engaging students in conversation

The program aims at motivating the students to interact with others in English. Therefore, we are interested not only in their total posts but also in the probability of engaging them in a conversation. Following the same rationale as Section 4.1.3, we use a multilevel logistic regression model to explore this question. Let $(Y_j)_{j=1,\ldots,N}$ be Bernoulli variables with $P(Y_j = 1|\eta_j) = \exp(\eta_j)/(1 + \exp(\eta_j))$ with:

$$\begin{aligned}\eta_j =&\beta + \gamma_0 c_j + \gamma_1 \log(\tau_j) + \gamma_2 p_j + \gamma_3 s_j + \gamma_4 y_j + \\ &\gamma_5 q_j + \gamma_6 l_j + \gamma_7 e_j + S + C + R,\end{aligned} \tag{6}$$

$Y_j = 1$ if the student posts a second comment in conversation j and 0 otherwise. β is the baseline. c_j, p_j and s_j are the complexity, polarity and specificity of the RT's response to the first comment posted by the student who initiated conversation j. q_j, e_j and l_j are boolean variables taking value 1 if the RT's response asked the student a question, included an emoticon or shared a link respectively. τ_j is the timelapse between the moment the student started conversation j and the RT replied. y_j is the school year. $\gamma_0, \ldots, \gamma_7$ represent the fixed effects of the corresponding covariates. The nested random effect student-CT-RT is represented by the random variables S, C and R, assumed to follow zero-mean Gaussian distributions with standard deviations σ_S, σ_C and σ_R. N is the total number of conversation threads and there may be several threads per student.

Results: Table 2 shows the estimated covariate effects. By far, and maybe not surprisingly, the fact that the RT asks the student a question has the largest stat. sig. positive effect on the probability of getting the student to continue the conversation. When a question is asked, assuming the rest of the covariates remain fixed, the odds ratio for students of the same classroom is $\exp(2.94) = 19.0$ and $\exp(2.86) = 17.5$, for low and high level students respectively.

As more elaborated comments often include questions, the positive effect of complexity suggests that more elaborated comments increase the probability of engaging a student in a conversation. On the contrary, the negative stat. sig. effect of polarity is likely due to the fact that very positive comments such as *"Great work!"* tend to be quite basic in terms of complexity. For high level students a larger delay is associated with more responses, as after a one week delay the odds ratio is 1.2 (the rest of the covariates and random effects remaining constant). This is likely because for high level students RT response delay is positively correlated with complexity (Spearman 0.1) and negatively correlated with comments polarity (Spearman -0.12): writing more elaborated comments take longer. This is not observed for low level students (Spearman 0.02 for both complexity and polarity), for whom it seems important to reply as soon as possible, as after a one week delay the odds ratio is 0.7. Finally, for both high and low level students, results suggest that using less specific words is associated with higher probability of engagement in the conversation.

5. SUMMARY AND CONCLUSIONS

We conducted an observational analysis of 27,627 comments, exchanged between 1074 high school students and 35 RTs over 2 years, to study the effect that different RT's feedback characteristics have on the students' posting behavior in an online EFL learning environment. The research questions, as well as the features defined for the characterization of the comments, were discussed and validated with the stakeholders and the leaders of the program in order to take advantage of their wide experience on the topic. Both parametric (multilevel modeling) and non-parametric (bootstrapping) analyses were performed, controlling for the effect of possible confounds such as (1) the classroom teacher, (2) the remote teacher and (3) the students' English proficiency level. Our results suggest that:

(1) Teachers should observe the complexity of their students' comments and adapt the complexity of their feedback ac-

student level	β		$\hat{\gamma}_0$		$\hat{\gamma}_1$		$\hat{\gamma}_2$		$\hat{\gamma}_3$		$\hat{\gamma}_4$		$\hat{\gamma}_5$		$\hat{\gamma}_6$		$\hat{\gamma}_7$
low	-7.11	***	0.08		-0.15	***	0.36		-2.20	**	0.24		2.94	***	-0.53		0.11
high	-7.46	***	0.35	***	0.10	**	-0.47	*	-2.66	**	0.62		2.83	***	-0.77	.	-0.39

Table 2: Effects of RT feedback characteristics on the probability of engaging the student in conversation, in logarithmic scale. Significance codes: 0 (***) 0.001 (**) 0.01 (*) 0.05 (.) 0.1

cordingly. Students who receive feedback that is too complex or too basic for their level post 13% ($p = 0.008$) and 15% ($p = 0.003$) fewer comments than those who receive adapted feedback, for low and high level students respectively.
(2) According to some RTs who were consulted about the potential causes of the observed behavior, the students may be more motivated when the language of the RT is accessible to them because they understand it, they learn from it and are challenged by it, without this turning into frustration.
(3) The best way to engage the students in a conversation is to pose a question (this increases the odds by 19 and 17.5 for low and high level students respectively). The comments should be complex enough to include a question (i.e., *"Great work!"* won't be enough) yet remain simple in terms words specificity. Also, for low level students, it is important to respond as quickly as possible (after a one week delay the odds ratio is 0.7).

Even if no causal inferences can be made, this study generated enlightening insights which have potential implications for the countless online platforms offering foreign language learning services, in which it is crucial to give the best possible learning experience while judiciously allocating resources (e.g. teachers' time).

Acknowledgments: We thank the TDL team and RTs for their participation and helpful feedback. J. Whitehill was supported by the National Science Foundation under Grant No. #1822830.

6. REFERENCES

[1] Aguerrebere, C., Bulger, M., Cobo, C., García, S., Kaplan, G., and Whitehill, J. How should online teachers of english as a foreign language write feedback to students? Available at `https://users.wpi.edu/~jrwhitehill/AguerrebereEtAl_EDM2019_full.pdf` (2019).

[2] Aguerrebere, C., Cabeza, S. G., Kaplan, G., Marconi, C., Cobo, C., and Bulger, M. Exploring feedback interactions in online learning environments for secondary education. In *Proc. of Latin Amer. Workshop on Learn. Analytics* (2018), pp. 128–137.

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

[4] Bates, D., Mächler, M., Bolker, B., and Walker, S. Fitting linear mixed-effects models using lme4. *J Stat Softw 67*, 1 (2015), 1–48.

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

[6] Deshpande, S., Palshikar, G. K., and Athiappan, G. An unsupervised approach to sentence classification. In *COMAD* (2010), p. 88.

[7] Efron, B., and Tibshirani, R. J. *An introduction to the bootstrap.* CRC press, 1994.

[8] Harris, L. R., Brown, G. T., and Harnett, J. A. Analysis of New Zealand primary and secondary student peer-and self-assessment comments: Applying Hattie and Timperley's feedback model. *Assessment in Education: Principles, Policy & Practice* (2015).

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

[10] Jiang, W., and Ribeiro, A. Effect of computer-mediated peer written feedback on ESL/EFL writing: A systematic literature review. *Elec. Int. J of Educ., Arts, and Science 3*, 6 (2017).

[11] Kahraman, A., and Yalvac, F. EFL turkish university students' preferences about teacher feedback and its importance. *Procedia-Social and Behavioral Sciences 199* (2015), 73–80.

[12] Lang, C., Heffernan, N., Ostrow, K., and Wang, Y. The impact of incorporating student confidence items into an intelligent tutor: A randomized controlled trial. *Int. Educ. Data Mininig Soc.* (2015).

[13] Miller, G. A. Wordnet: a lexical database for english. *Comm. of the ACM 38*, 11 (1995), 39–41.

[14] Miyamoto, Y., Coleman, C., Williams, J., Whitehill, J., Nesterko, S., and Reich, J. Beyond time-on-task: The relationship between spaced study and certification in MOOCs. *Available at SSRN 2547799* (2015).

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

[16] Olsen, J. K., Aleven, V., and Rummel, N. Predicting student performance in a collaborative learning environment. *Int. Educ. Data Mining Soc.* (2015).

[17] Romero, C., and Ventura, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3*, 1 (2013), 12–27.

[18] Saeed, M. A., Ghazali, K., and Aljaberi, M. A. A review of previous studies on ESL/EFL learners' interactional feedback exchanges in face-to-face and computer-assisted peer review of writing. *Int. J. of Educ. Tech. in Higher Education 15*, 1 (2018), 6.

[19] Shang, H.-F. An exploration of asynchronous and synchronous feedback modes in EFL writing. *J. of Computing in Higher Education 29*, 3 (2017), 496–513.

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