Preliminary Experiments on Crowdsourced Evaluation of Feedback Granularity

Nitin Madnani[1], Martin Chodorow[2], Aoife Cahill[1], Melissa Lopez[1], Yoko Futagi[1] and Yigal Attali[1]

[1]*Educational Testing Service, Princeton, NJ*
[2]*Hunter College and the Graduate Center, City University of New York, NY*

Abstract

Providing writing feedback to English language learners (ELLs) helps them learn to write better, but it is not clear what type or how much information should be provided. There have been few experiments directly comparing the effects of different types of automatically generated feedback on ELL writing. Such studies are difficult to conduct because they require participation and commitment from actual students and their teachers, over extended periods of time, and in real classroom settings. In order to avoid such difficulties, we instead conduct a crowdsourced study on Amazon Mechanical Turk to answer questions concerning the effects of type and amount of writing feedback. We find that our experiment has several serious limitations but still yields some interesting results.

1 Introduction

A core feature of learning to write is receiving feedback and making revisions based on the information provided (Li and Hegelheimer, 2013; Biber et al., 2011; Lipnevich and Smith, 2008; Truscott, 2007; Rock, 2007). However, an important question to answer before building automated feedback systems is what type of feedback (and degree of interactivity) can best support learning and retention. Is it better to restrict the system to providing feedback which indicates that an error has been made but does not suggest a possible correction? Or is it better for the learner to receive feedback, which provides a clear indication of the error location as well as the correction itself, or even an explanation of the underlying grammatical rule? In this study, we refer to the first type of feedback as *implicit* feedback and to the second type as *explicit* feedback.

To the best of our knowledge, there is no empirical study that directly compares several different amounts of detail (granularities) in automatically generated feedback in terms of their impact on learning outcomes for language learners. This is not surprising since the ideal study would involve conducting controlled experiments in a classroom setting, requiring participation from actual language learners and teachers.

In this paper, we examine whether a large-scale crowdsourcing study conducted on Amazon Mechanical Turk, instead of in classrooms, can provide any answers about the effect of feedback granularity on learning. Our experiments are preliminary in nature but nevertheless yield results that — despite not being directly applicable to ELLs — are interesting. We also report on lessons we have learned about the deficiencies in our study and suggest possible ways to overcome them in future work.

For the purpose of this study, we define an "improvement in learning outcome" as an improvement in the performance of the Turkers on a specific task: detecting and correcting preposition selection errors in written text. Obviously, learning to use the correct preposition in a given context is only one, albeit an important, aspect of better writing. However, we concentrate on this single error type since: (a) doing so will allow us to remove any unintended effects of interactions among multiple errors, ensuring that the feedback message is the only variable in our experiment, and (b) automated systems for correcting

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 162–171,
Denver, Colorado, June 4, 2015.

preposition selection errors have been studied and developed for many years. Reason (b) is important since we can use the output of these automated systems as part of the feedback.

Briefly, a high-level description of the study is as follows:

1. Over multiple sessions, Turkers detect and correct preposition errors in sentences.
2. We provide sub-groups of Turkers with different types of feedback as they proceed through the sessions.
3. We measure the differences in Turker performance and see if the differences vary across feedback types.

Section 2 describes related work. Section 3 describes the experimental design of our study in more detail. Section 4 presents our analysis of the results from the study and, finally, Section 5 concludes with a summary of the study along with the lessons we learned from conducting it.

2 Related Work

One automated writing evaluation tool that helps students plan, write and revise their essays guided by instant diagnostic feedback and a score is Criterion. Attali (2004) and Shermis et al. (2004) examine the effect of feedback in general in the Criterion system and find that students presented with feedback are able to improve the overall quality of their writing. Those studies do not investigate different feedback types; they look at the issue of whether feedback in general is a useful tool. We propose to look at varying levels of detail in feedback messages to see what effect this has on student learning.

We have found no large-scale empirical studies comparing the types of feedback on grammatical errors in the field of second language acquisition, and no work at all on using computer-generated corrections. In the field of second language acquisition, the main focus has been on explicit vs. implicit feedback in a general sense.

The major focus of studies on Corrective Feedback, or "CF", for grammatical errors has been on whether CF is effective or not following the controversial claim by Truscott (1996) that it may actually be harmful to a learner's writing ability.

Russell and Spada (2006) used 56 studies in their meta-analysis of CF research, and of those, 22 focused on written errors and one looked at both oral and written errors. Meihami and Meihami (2013) list a few more studies, almost all of which are from 2006 or later. Some of the studies were conducted in classroom settings, while others were in "laboratory" settings. In all of the studies, corrective feedback was given by humans (teachers, researchers, peers, other native speakers), so the sample sizes are most likely limited (unfortunately, that information is missing from the Russell and Spada meta-analysis).

Doughty and Williams (1998) summarize the findings of the Lyster and Ranta (1997) classroom study of the effectiveness of various feedback techniques. Lyster and Ranta (1997) found that one of the effective types of feedback for stimulating learner-generated repairs was a repaired response from the teacher. There were also several other feedback types that were found to be effective including meta-linguistic cues, clarification requests and repetition of the learner error. Carroll and Swain (1993) found that in general some kind of feedback is better than no feedback.

There are very few studies that have compared the effectiveness of different types of written corrective feedback. Bitchener et al. (2005) and Bitchener (2008) seem to show that direct feedback (oral or written) is more effective than indirect, while in (Bitchener and Knoch, 2008; Bitchener and Knoch, 2009), which have larger sample sizes, the difference disappeared. Bitchener and Knoch (2010) investigated different types of corrective feedback over a 10-month period and also show that there are no differences among different types of feedback. However, Sheen (2007) found that the group receiving meta-linguistic explanations performed better than the one who received direct error corrections in the delayed post-test 2 months later. All of these studies focused only on English articles.

Biber et al. (2011) present a synthesis of existing work on the influences of feedback for writing development. One point from this report that is very relevant to our current work is that "Truscott (2007) focuses on the quite restricted question of the extent to which error correction influences writing accuracy for L2-English students. This study concluded

that overt error correction actually has a small negative influence on learners' abilities to write accurately. However, the meta-analysis was based on only six research studies, making it somewhat difficult to be confident about the generalizability of the findings." Biber et al. (2011) also mention that "In actual practice, direct feedback is rarely used as a treatment in empirical research."

The work most directly relevant to our study is that of Nagata and Nakatani (2010), who attempt to measure actual impact of feedback on learning outcomes for English language learners whose native language is Japanese. At the beginning of the study, students wrote English essays on 10 different topics. Errors involving articles and noun number were then flagged either by a human or by two different automatic error detection systems: one with high precision and another with high recall. A control group received no error feedback. Learning was measured in terms of reduction of error rate for the noun phrases in the students' essays. Results showed that learning was quite similar for the human-supplied feedback and the high-precision automated feedback conditions, and that both were better than the no-feedback condition. In contrast, the high-recall automated feedback condition actually yielded results worse than the no-feedback condition. This latter finding supports the commonly held assumption that it is better to provide less feedback than to provide incorrect feedback. Note, however, that their study only compares providing implicit feedback to providing no feedback.

3 Experimental Setup

We designed a crowdsourcing experiment to examine the differences in learning effects resulting from different types of feedback. The overall design of the experiment consists of three phases:

1. **Phase 1**. Recruit Turkers and measure their pre-intervention preposition error detection and correction skills. All Turkers are provided with the same minimal feedback during the pre-intervention session, i.e., they are on equal footing when it comes to writing feedback.
2. **Phase 2**. Divide the recruited Turkers into different, mutually exclusive groups. Each group participates in a series of intervention sessions where the Turkers in that group receive one specific type of feedback.
3. **Phase 3**. Measure the post-intervention performance for all Turkers. Similar to the pre-intervention session, the same minimal feedback is provided during the post-intervention session.

We chose to use five different feedback granularities in our study, which are outlined below. The first one represents implicit feedback and the last four represent explicit feedback.

1. **Minimal Feedback**. Messages are of the form: *There may be an error in this sentence*.
2. **Moderate Feedback**. The incorrect preposition is highlighted and the feedback message is of the form: *The highlighted preposition P_1 may be incorrect*.
3. **Detailed Feedback 1**. The incorrect preposition is highlighted and the feedback message is of the form: *The highlighted preposition P_1 may be incorrect; the preposition P_2 may be more appropriate*, where P_2 is a human expert's suggested correction for the error.
4. **Detailed Feedback 2**. The incorrect preposition is highlighted and the feedback message is of the form: *The highlighted preposition P_1 may be incorrect; the preposition P_2 may be more appropriate*, where P_2 is the correction assigned the highest probability by an automated preposition error correction system (Cahill et al., 2013).
5. **Detailed Feedback 3**. The incorrect preposition is highlighted and the feedback message is of the form: *The highlighted preposition P_1 may be incorrect; the following is a list of prepositions that may be more appropriate*, where the list contains the top 5 suggested corrections from the automated error correction system.

For all three detailed feedback types, Turkers were told that the corrections were generated by an automated system. Table 1 shows the design of our experimental study wherein all recruited Turkers were divided into five mutually exclusive groups, each corresponding to one of the feedback types described above.

For our pre-intervention/recruitment session (Session 1), we collected judgments from 450 Turkers

	Session 1	Session 2	Session 3	Session 4	Session 5
Group 1	Minimal		Minimal		Minimal
Group 2	Minimal		Moderate		Minimal
Group 3	Minimal		Detailed 1		Minimal
Group 4	Minimal		Detailed 2		Minimal
Group 5	Minimal		Detailed 3		Minimal

Table 1: The experimental design of the study. Turkers were divided into five mutually exclusive groups and always shown the same type of feedback during the intervention (sessions 2–4). All Turkers were shown the same minimal feedback during the pre- and the post-intervention (sessions 1 and 5, respectively).

	Session 1	Session 2	Session 3	Session 4	Session 5
Group 1	82	78	76	74	72
Group 2	82	72	70	68	66
Group 3	82	72	70	70	65
Group 4	83	74	72	70	70
Group 5	83	75	74	73	72
Total	412	371	362	355	345

Table 2: The number of Turkers that participated in each group for each session.

without regard for qualification requirements. One Turker's work was rejected for carelessness, and the remaining 449 received approved payments of $1. After scoring the responses, removing questionable work, and reviewing the distribution of scores, we reduced this number to 412 (approximately 82 Turkers per group). We then randomly assigned Turkers to one of the five feedback groups.[1] We administered Session 2 approximately two weeks after Session 1. We created a unique task for each feedback group, and Turkers were only permitted to access the task for their assigned group. Upon review, their work was approved for payment, and a new qualification score was assigned for entrance into the next session. The remaining sessions were posted every other day up to Session 5, and each task remained available for two weeks after posting. The payment amount increased by 50 cents for each new session, adding up to a total of $10 per Turker if they completed all five sessions. Table 2 shows the number of Turkers assigned to each group who participated in each of the five sessions.

We used the CLC-FCE corpus (Yannakoudakis et al., 2011), which has been manually annotated for preposition errors by professional English language instructors. We randomly selected 90 sentences with preposition errors and 45 sentences without errors and manually reviewed them to ensure their suitability. Unsuitable sentences were replaced from the pool of automatically extracted sentences until we had a total of 135 suitable sentences. We annotated each sentence containing an error with a correct preposition. The 135 sentences were then randomly divided into 5 HITs (Human Intelligence Tasks, the basic unit of work on MTurk), one for each of the five sessions. Each HIT was generated automatically, with manual human review. Given a sentence containing an error and a correction, we automatically extracted the following additional data:

- A version of the sentence where the only error is the preposition error (specifically errors where an incorrect preposition is used).

[1] An MTurk feature that was essential to this study was the ability to designate "qualifications" to recruit and target specific Turkers. MTurk requesters can use these qualifications to assign Turkers to conditions and keep a record of their status. After Turkers completed Session 1, we were able to use our own qualifications and a range of qualification scores to assign Turkers to groups and control the order in which they completed the sessions. Although the Turkers were assigned randomly to groups, we manually ensured that the distributions of Session 1 scores were similar across groups.

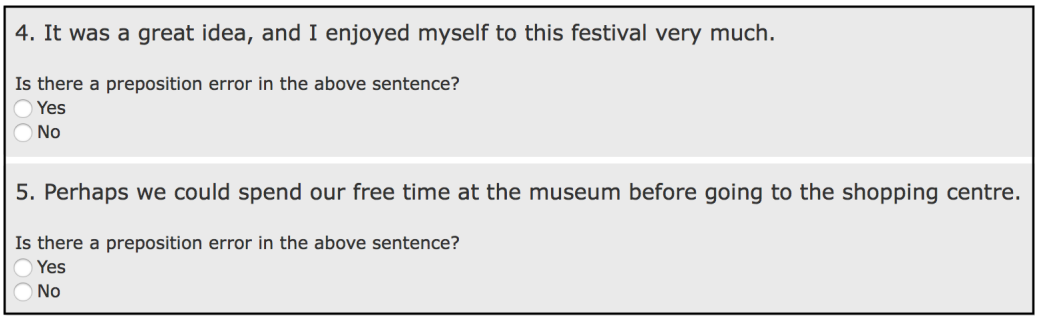

Figure 1: A partial screenshot of the HIT shown to the Turkers. The first sentence contains a preposition error and the second does not.

- The incorrect preposition, its position in the sentence, and the human correction.
- A version of the sentence that has the preposition error corrected.

The pre- and post-intervention HITs consisted of 30 sentences and the intervention sessions consisted of 25 sentences each. About a third of the sentences in each HIT contained no errors (to measure detection ability) and the remaining contained a single preposition error (to measure correction ability). Turkers were first asked to indicate whether or not there was a preposition error in each sentence, in order to test their error detection skills. Once the Turker answered, they received a feedback message of the appropriate granularity directing them to correct the error in the sentence, if there was one. If there were no errors annotated in the sentence, Turkers received a message saying that the sentence contained no errors. Figure 1 shows a partial screenshot of a HIT.

In order to understand more about our participants, we geo-located Turkers using their IP addresses. A significant majority of the Turkers — 319 out of the 345 who participated in all five sessions — were from the United States with the remaining located in India (21), Mexico (3), Ireland (1), and Sweden (1).

4 Analysis

To prepare data for analysis, we automatically scored the Turker responses and manually adjusted these scores to account for sentences where more than one correction was appropriate. Scoring for each sentence depended on the presence of an error. For sentences with errors, Turkers could score a maximum of two independent points: 1 point for detection and 1 point for correction. Because Turkers were not asked to correct sentences without errors, these were only worth 1 point for detection.

4.1 Prepositions Used

Before examining the Turker responses, we analyzed the actual prepositions that were involved in each erroneous sentence in each session. Figure 2 shows this distribution. We observe that not all prepositions are represented across all sessions and that the distributions of prepositions are quite different. In fact, only three prepositions errors (“of”, “in” and “to”) appear in all five sessions.

4.2 Turker Motivation

One of the most common problems with using crowdsourcing solutions like MTurk is that of quality control. In our study, we excluded 37 Turkers at the pre-intervention stage for quality control. However, after that session, no Turkers were excluded since we wanted all recruited Turkers to finish all five sessions. Therefore, it is important to examine the recruited Turkers’ responses provided for all three intervention sessions for any strange patterns indicating that a Turker was trying to game the task by not providing good-faith answers. For example, a Turker who was only motivated to earn the HIT payment and not to make a useful contribution to the task could:

- answer ‘yes’ or ‘no’ to *all* error detection questions or at random
- *always* accept the suggested preposition
- *always* use a random preposition as their answer

- *always* pick the first preposition from a given list of prepositions

We analyze the Turkers' error detection responses all together and their error correction responses by feedback type.

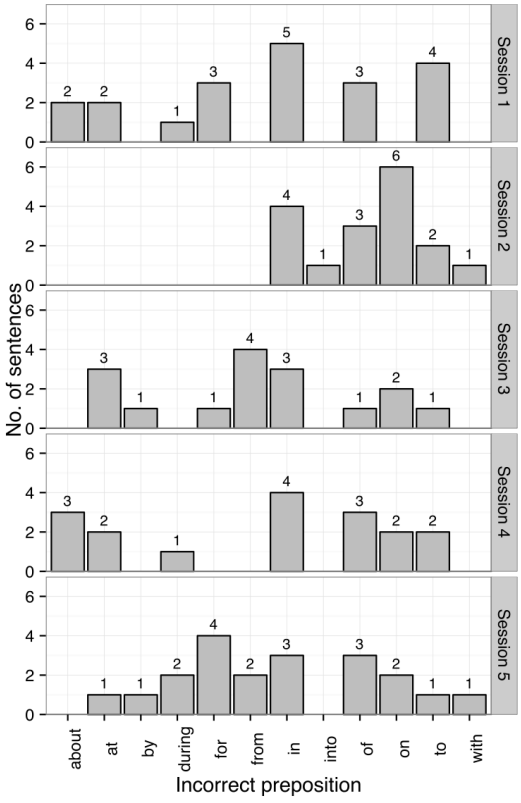

Figure 2: The distribution of prepositions involved in the erroneous sentences for each session.

4.2.1 Analyzing Detection Responses

First, we examine the possibility that Turkers may have answered 'yes' or 'no' at the error detection stage for all questions or may have selected one of those answers at random for each question. To do this, we simply compute the proportion of sentences for which each Turker accurately detected the error, if one was present. The faceted plot in Figure 3 shows that almost all of the Turkers seem to have answered the error detection questions accurately, and without trying to game the system. Each facet shows a histogram of the average accuracy (across all sentences) of the Turkers from one of the five feedback groups and for each of the five sessions. The dotted line in each plot indicates the accuracy that would

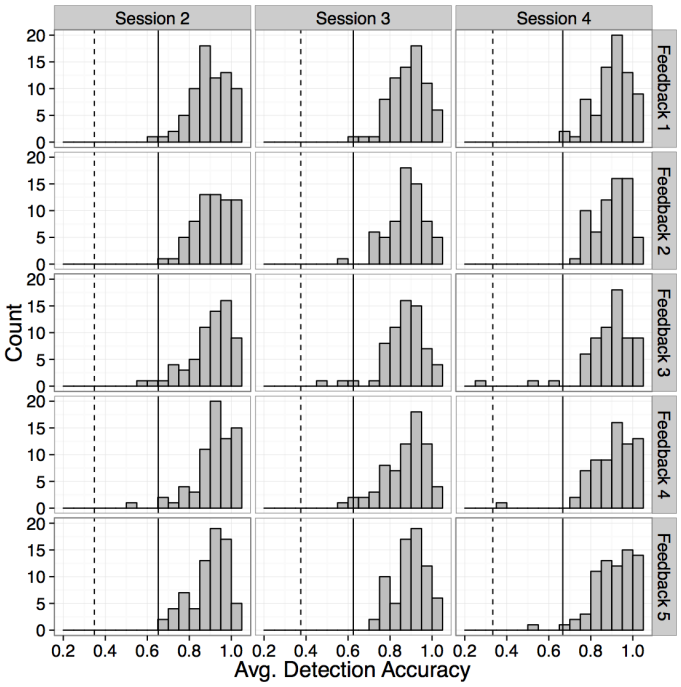

Figure 3: A histogram of the Turkers' average error detection accuracy for the three intervention sessions. The dotted and solid lines indicate accuracies that a Turker would have obtained had they answered every question in a session with 'No' or 'Yes', respectively.

have been obtained by a Turker had they simply said 'no' to all the error detection questions and the solid line indicates the accuracy that would have been obtained by answering 'yes' to all of them. Note that these lines are the same across feedback groups because the sentences are the same for a session, irrespective of the feedback group.

4.2.2 Analyzing Correction Responses

In this section, we analyze the Turker error correction responses by feedback type. First, we examine the responses from the Turkers in Group 3, i.e., those who received messages of the **Detailed Feedback 1** type. Figure 4 shows that most of the Turkers accepted the suggested correction. Note that since Turkers were not informed that the suggestion came from an expert, this is still an indicator of good Turker performance. Furthermore, the figure shows that even a majority of the Turkers who decided not to accept the suggestion actually answered with an alternative correct preposition of their own. The "*Not Accepted - Incorrect (Other)*" category in the

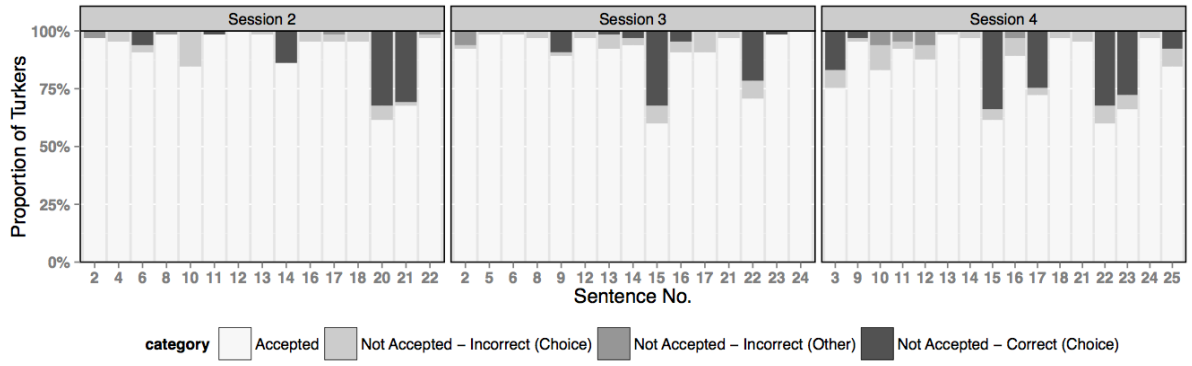

Figure 4: By session and sentence, the proportion of Turkers in Group 3 accepting the (always correct) suggested preposition, and, if not accepting it, the correctness of their repairs.

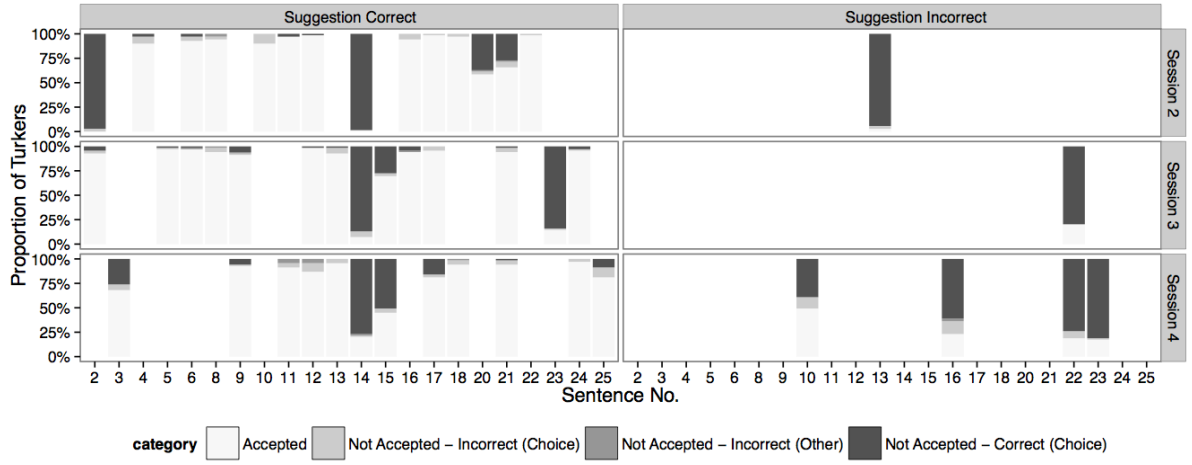

Figure 5: By session and sentence, the proportion of Turkers in Group 4 accepting the (possibly incorrect) suggested preposition, and, if not accepting it, the correctness of their repairs.

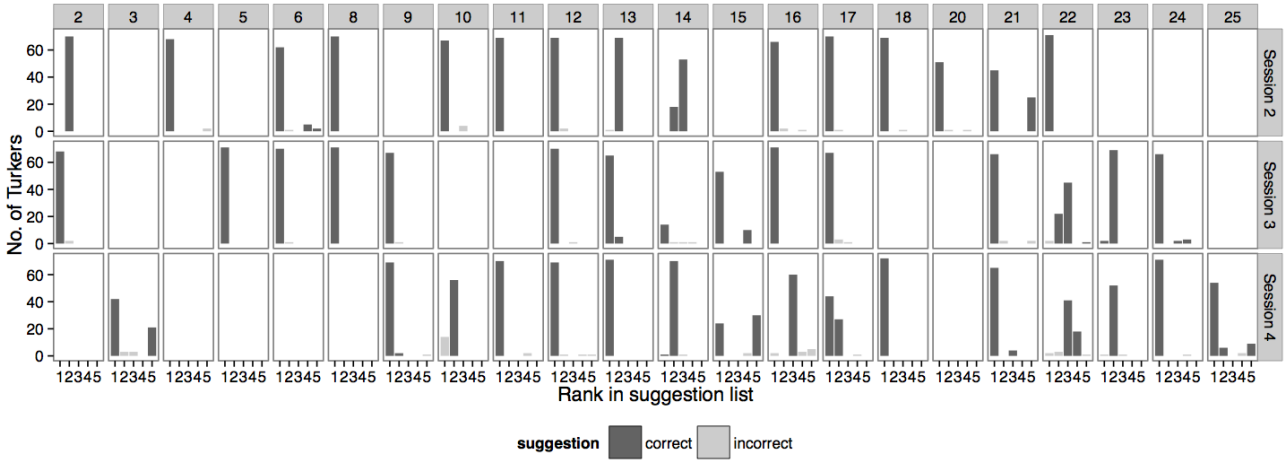

Figure 6: By session and sentence, the number of Turkers in Group 5 selecting a preposition at each rank position in the suggestion list and the correctness of their selection.

figure refers to rare cases where the Turkers deleted the erroneous preposition or made other changes in the sentence instead of fixing the preposition error.

Next, we examine responses from Turkers in Group 4. Figure 5 shows – similar to Figure 4 – the proportion of Turkers that simply accepted the suggestion provided as compared to those who did not. However, in this case, we have the additional possibility of the suggested preposition being incorrect, since it is generated by an automated system. Again, we see that most Turkers accept the suggested preposition when it's correct, but when it's incorrect, they answer with a different correct preposition of their own.

Our analysis for Group 5 shows similar trends, i.e., most Turkers take the time to find a contextually accurate answer even if it's not on the list of suggested prepositions. Therefore, we do not include a corresponding plot for Group 5 in the paper.

Instead, we thought it would be interesting to examine the Turkers' responses from another angle. Since a correct suggestion may not always be the top-ranked preposition in the suggestion list, it would be interesting to include suggestion ranks into the analysis. Figure 6 shows, for each sentence in each session, the number of Turkers that accepted each ranked suggestion. The color of the bar indicates whether that particular suggestion was correct or incorrect. Note that there may be multiple correct suggestions in a list. Again, we observe that, although there are some Turkers who accepted the top ranked answer even if it was incorrect, the great majority took the time to select a correct preposition no matter what its rank was. Note that the blank facets in the figure represent sentences for a session that did not contain any errors.

4.3 Learning Effects

In this section, we attempt to answer the primary question for the study, i.e., assuming that sessions 2-4 constitute the intervention, is there a significant difference in the pre-intervention and post-intervention Turker performance across the various feedback conditions?

To answer this question, we first compute the log-odds of Turkers accurately detecting (and correcting) errors for the pre-intervention and post-intervention sessions — sessions 1 and 5 respectively — and plot them in Figure 7. We observe that for detection, the changes in performance between pre- and post-intervention are similar across feedback groups and no group seems to have performed better than Group 1, post-intervention. As far as correction is concerned, there is improvement across all feedback conditions, but the change in Group 3's performance seems much more dramatic than that for the other groups.

However, we need to determine whether these improvements are statistically significant or instead can simply be explained away by sampling error due to random variation among the Turkers or among the sentences. To do so, we use a linear mixed-effects model.[2] The advantages of using such a model are that, in addition to modeling the fixed effects of the intervention and the type of feedback, it can also model the random effects represented by the Turker ID (Turkers have different English proficiencies) and the sentence (a sentence may be easier or more difficult than another). In addition, it can also help us account for further random effects, e.g., the effect of Turkers in different groups learning at different rates and the sentences being affected differently by the different feedback conditions. Specifically, we fit the following mixed-effects logistic regression model using the `lme4` package in R:

$$\begin{aligned}\texttt{accurate} \sim\ & \texttt{group} * \texttt{session} \\ & + (1 + \texttt{session} \,|\, \texttt{mturkid}) \\ & + (1 + \texttt{group} \,|\, \texttt{sentnum})\end{aligned}$$

where `accurate` (0 or 1) represents whether a Turker accurately detected or corrected the error in the sentence, `group` represents the feedback type, and `session` is either the pre- or the post-intervention session (1 or 5). The $*$ in the model indicates the inclusion of the interaction between `group` and `session`, which is necessary since our model is focused on a second order measure (the differences between changes in performance). We fit two models of this form, one for detection and one for correction. Examination of the results indicates:

1. In the detection model, there was a significant effect of `session` ($p < 0.05$). However, neither the effect of `group` nor any of the interactions of

[2] cf. Chapter 7, Baayen (2008).

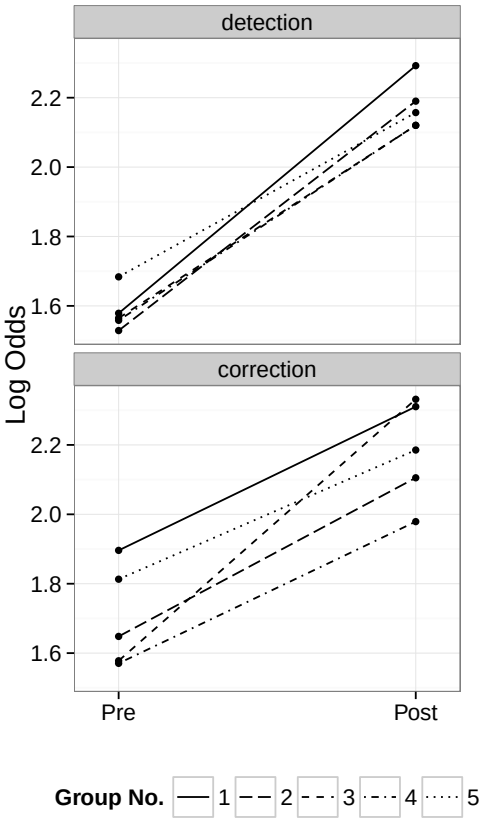

Figure 7: Log-odds of Turkers accurately detecting or correcting preposition errors, pre- and post-intervention.

`group` by `session` were significant.

2. In the correction model, the effect of `group` was significant only for Group 3. In addition, the interaction of `group` by `session` was also significant only for Group 3.

From the above results, we can conclude that:

- Irrespective of the feedback type they were shown, Turkers exhibited significant improvements in their detection performance between the pre- and post-intervention sessions, probably due to practice. This was *not* the case for correction.
- Only Turkers from Group 3 (i.e., those shown expert suggestions as feedback - **Detailed Feedback 1**) exhibited a significantly larger improvement in correction performance due to the intervention, as compared to the Turkers that were shown minimal feedback (no explicit feedback).

5 Summary

In this paper, we presented a study that uses crowdsourcing to evaluate whether the granularity of writing feedback can have a measurable impact on learning outcomes. The study yields some interesting results. In particular, it provides some evidence to support the finding from Nagata and Nakatani (2010) that only high precision feedback can help learners improve their writing. However, the study is quite preliminary in nature and focuses on the outcomes for a *single* writing skill. In addition, there were several other deficiencies:

- The distributions of preposition errors across sessions varied considerably which might have made it harder for Turkers to generalize what they learned from one session to another. Another possible confounding factor may have been the fact that the Turker population we recruited was largely located in the U.S. whereas the sentences were chosen from a corpus of British English.
- It is clear from the high levels of pre-intervention error detection and correction performance that the recruited Turkers are not English language learners. We had hoped to recruit Turkers with varied English proficiencies by not restricting participation to any specific countries. However, a more explicit strategy is likely necessary.
- Even though we were fortunate that the Turkers were well-motivated throughout our task, enforcing quality control in a study of this type is challenging.
- Note that in our experimental set up, Turkers receive, as part of the feedback message, an explicit indication of whether or not their detection answers were correct, but no such indication is provided for their correction answers. This could be why `session` had a significant effect for detection but not for correction.

We believe that our study, along with all its deficiencies, represents a useful contribution to the field of assessing the impact of writing feedback, and that it can help the community design better studies in the future, whether they be conducted using crowdsourcing or with actual students in a classroom.

Acknowledgments

We would like to thank the three anonymous reviewers. We would also like to thank Keelan Evanini, Beata Beigman Klebanov and Lin Gu for their comments.

References

Yigal Attali. 2004. Exploring the Feedback and Revision Features of *Criterion*. Paper presented at the National Council on Measurement in Education (NCME), Educational Testing Service, Princeton, NJ.

R. H. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Douglas Biber, Tatiana Nekrasova, and Brad Horn. 2011. The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. Research Report RR-11-05, Educational Testing Service, Princeton, NJ.

J. Bitchener and U. Knoch. 2008. The Value of Written Corrective Feedback for Migrant and International Students. *Language Teaching Research Journal*, 12(3):409–431.

J. Bitchener and U. Knoch. 2009. The Relative Effectiveness of Different Types of Direct Written Corrective Feedback. *System*, 37(2):322–329.

J. Bitchener and U. Knoch. 2010. *The Contribution of Written Corrective Feedback to Language Development: A Ten Month Investigation*.

J. Bitchener, S. Young, and D. Cameron. 2005. The Effect of Different Types of Corrective Feedback on ESL Student Writing. *Journal of Second language Writing*.

J. Bitchener. 2008. Evidence in Support of Written Corrective Feedback. *Journal of Second Language Writing*, 17:69–124.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust Systems for Preposition Error Correction Using Wikipedia Revisions. In *Proceedings of NAACL*, pages 507–517, Atlanta, GA, USA.

S. Carroll and M. Swain. 1993. Explicit and Implicit Negative Feedback. *Studies in Second Language Acquisition*, 15:357–386.

C. Doughty and J Williams. 1998. *Pedagogical Choices in Focus on Form*.

Z. Li and V. Hegelheimer. 2013. Mobile-assisted Grammar Exercises: Effects on Self-editing in L2 Writing. *Language Learning & Technology*, 17(3):135–156.

Anastasiya A. Lipnevich and Jeffrey K. Smith. 2008. Response to Assessment Feedback: The Effects of Grades, Praise, and Source of Information. Research Report RR-08-30, Educational Testing Service, Princeton, NJ.

R. Lyster and L. Ranta. 1997. Corrective Feedback and Learner Uptake. *Studies in Second Language Acquisition*, 19:37–66.

B. Meihami and H. Meihami. 2013. Correct I or I Dont Correct Myself: Corrective Feedback on EFL Students Writing. In *International Letters of Social and Humanistic Sciences*, page 8695.

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating Performance of Grammatical Error Detection to Maximize Learning Effect. In *Proceedings of COLING (Posters)*, pages 894–900, Beijing, China.

JoAnn Leah Rock. 2007. The Impact of Short-Term Use of Criterion on Writing Skills in Ninth Grade. Research Report RR-07-07, Educational Testing Service, Princeton, NJ.

J. Russell and N. Spada. 2006. The Effectiveness of Corrective Feedback for the Acquisition of L2 Grammar: A Meta-analysis of the Research. In J. D. Norris and L. Ortega, editors, *Synthesizing Research on Language Learning and Teaching*, pages 133–164. John Benjamins, Philadelphia.

Y. Sheen. 2007. The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners Acquisition of Articles. *TESOL Quarterly*, 41:255–283.

Mark D. Shermis, Jill C. Burstein, and Leonard Bliss. 2004. The Impact of Automated Essay Scoring on High Stakes Writing Assessments. In *Annual Meeting of the National Council on Measurement in Education*.

J. Truscott. 1996. The Case against Grammar Correction in L2 Writing Classes. *Language Learning*, 46:327–369.

John Truscott. 2007. The Effect of Error Correction on Learners' Ability to Write Accurately. *Journal of Second Language Writing*, 16(4):255–272.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the ACL: HLT*, pages 180–189, Portland, OR, USA.