Linguistic and review features of peer feedback and their effect on the implementation of changes in academic writing: A corpus based investigation

Djuddah A.J. Leijen & Anna Leontjeva

University of Tartu | Estonia

Abstract: The inclusion of peer feedback activities into the academic writing process has become common practice in higher education. However, while research has shown that students perceive many features of peer feedback to be useful, the actual effectiveness of these features in terms of measurable learning outcomes remains unclear. The aim of this study was to investigate the linguistic and review features of peer feedback and how these might influence peers to accept or reject revision advice offered in the context of academic writing among L2 learners. A corpus-based machine learning approach was employed to test three different algorithms (logistic regression, decision tree, and random forests) on three feature models (linguistic, review, and all features) to determine which algorithm offered the best predictive results and to determine which feature model most accurately predicts implementation. The results indicated that random forests is the most effective way of modeling the different features. In addition, the feature model containing all features most accurately predicted implementation. The findings further suggest that directive comments and multiple peer comments on the same topic included in the feedback process seem to influence implementation.

Keywords: computer supported peer feedback, academic writing development, corpus analysis, machine learning, L2 learners

journal of WRITING RESEARCH

Leijen, D. A.J., & Leontjeva, A. (2012). Linguistic and review features of peer feedback and their effect on the implementation of changes in academic writing: A corpus based investigation. *Journal of Writing Research*, *4*(2), 177-202. http://dx.doi.org/10.17239/jowr-2012.04.02.4

Contact: Djuddah A.J. Leijen, Department of General Linguistics, University of Tartu, Jakobi 2, 50090 Tartu | Estonia – djuddah.leijen@ut.ee

1. Introduction

In academic writing pedagogy, the practice of utilizing peer feedback as a means to develop and support writing has gained enormous momentum in the past few years. It is generally agreed that feedback has to be of a specific nature (Flower, Hayes, Carey, Schriver, & Stratman, 1986) and include certain functions (Van den Berg, Admiraal, & Piloot, 2006; Van der Pol, Van den Berg, Admiraal, & Simons, 2008) to support the learning process of peer authors. In writing research, the characteristics of nature and functions of feedback have been organized into typologies (Cho, Schunn, & Charney, 2006; Chi, 1996) that can be used to measure both the usefulness and effectiveness of peer feedback on the writing process. For example, Cho et al. (2006) investigated the perceived usefulness of six different types of peer feedback: directive, non-directive, praise, criticism, summary, and off task. The results indicated that certain types of feedback have positive effects and others negative effects on perceived usefulness. It is reasonable to expect that if a type of feedback is considered useful, it should also have a positive effect; however, when comparing perceived usefulness and actual effectiveness of peer feedback, a discrepancy between the two is often found. For example, peers consider praise to be more useful in comparison to critical comments (Straub, 1997; Tseng, & Tsai, 2007; Hyland, 2000). However, in terms of effectiveness, the inclusion of praise is often found to have little effect on actual writing performance (Ferris, 1997; Cho & MacArthur, 2010; Cho, Chung, King, & Schunn, 2008). A similar discrepancy is also found for directive and non-directive comments (Clare, Valdes, & Patthey-Chavez, 2000; Beason, 1993). Although directive comments have been found to be useful, as they provide details of how to improve a given piece of writing (Knoblauch & Brannon, 1981; Moreno, 2004; Pridemore & Klein, 1991), non-directive comments are found to be more effective in improving writing performance (Cho & MacArthur, 2010; Strijbos, Narciss, & Dünnebier, 2010; Topping, 2010; Leki, 1990). Thus, although many features are found to have a positive effect on perceived usefulness, measuring the effectiveness of these features for implementation still remains ambiguous or produces comparatively conflicting results.

Recent research on peer feedback is making attempts to provide more insights into this question, and different features measuring effectiveness have already been offered. According to Nelson and Schunn (2009), for example, peer feedback is most effective when solutions are offered, the location of the problem is given, and a summary is included; however, peer feedback hinders implementation when the feedback includes an explanation of the problem. Their model was derived from an initial complex model, incorporating evidence from various studies investigating effective feedback, which included additional features such as praise and mitigation, a linguistic politeness strategy used to soften the effect of a direct criticism. These features were excluded from the revised model because no evidence was found that they affected implementation. Van der Pol et al. (2008) found that feedback instances that included

concrete revision advice led to the greatest amount of revision in a student's text. Although the use of praise received strong usefulness ratings by students, Van der Pol et al. showed that a considerable amount of these feedback instances did not include many revision suggestions and therefore their effect on implementation was minimal. Although students consider mitigation and the use of praise to be useful and include them in feedback instances, their effect on implementation is still unclear.

One of the main limitations of the studies described above is that they have not been able to determine how or to what extent more specific features of language, such as the formulation of praise, critical remarks, and the use of mitigation devices, influence the effectiveness of peer feedback. This is perhaps unsurprising, as most of this research has been carried out within an educational sciences rather than an applied linguistics research tradition. However, studies that have investigated more specific linguistic qualities represented in feedback of second language learners of English (L2) suggest that the way feedback is constructed does indeed have an impact on a student's reception of the feedback and understanding of the content of that feedback, and that this in turn determines implementation (Hyland & Hyland, 2001; Ferris, 1997; Nguyen, 2008). Hyland and Hyland investigated the use of mitigation as a device to soften criticism. According to Hyland and Hyland, the use of mitigation strategies often causes confusion amongst L2 students leading to critical feedback not being understood and therefore not used. This result is also supported by Nguyen's (2008) findings. In an intercultural pragmatics study comparing L2 and L1 students, Nguyen concludes that L2 students have the greatest difficulty in understanding mitigation as well as in being able to soften their own critical feedback using mitigation strategies. As a result, some critical feedback instances, which were not mitigated, were perceived by some students as rude. What this research shows is that the way one expresses oneself through language, whether it is by offering concrete advice, by using affective language, or by softening criticism, can have a strong impact on the reception and thus the implementation of the feedback. It is also clear from studies such as Nguyen's that L2 students may encounter difficulties in constructing the right tone when writing up their feedback, and in correctly interpreting the feedback offered to them.

Based on the above, it can be concluded that further research is needed to understand how linguistic features in peer feedback, specifically but not exclusively for L2 students, affect their writing performance. This paper aims to contribute to current research investigating the effectiveness of peer feedback by applying the methodology of corpus linguistics to identify specific linguistic features and more general review features of peer feedback, and to determine which of these features affect implementation.

Corpus linguistics is a rapidly growing field of research, and language learner corpora have proven to be an invaluable resource for the study of authentic, context-driven student language (Biber, Conrad, & Reppen, 2006; Gilquin, Granger, & Paquot, 2007). In addition, corpus-informed approaches are now becoming more commonplace in writing research (Schlitz, 2010; Hüttner, 2010). The advantage of

corpus-based approaches to the analysis of authentic learner data is that they allow the researcher to conduct studies on a much larger scale than hitherto, thereby yielding more reliable sets of quantitative results with a greater practical value (Pendar & Chapelle, 2008; Xiong, Litman, & Schunn, 2012). The research reported in this study combines corpus-based methods with machine learning techniques adopted from the neighbouring discipline of natural language processing in order to identify language specific patterns which generate the most effective results in peer feedback interactions.

2. Methods

2.1 Data

A corpus of peer feedback was compiled using data collected in an introduction to academic writing course using a web-based peer feedback tool SWoRD (Scaffolded Writing and Reviewing in the Discipline). The use of computer-supported peer-review systems, such as the one used for this study, is beneficial for the collection of corpus data (both large and small) suitable for more complex analysis of learner language. This is specifically beneficial for gaining a better understanding of learner corpora of non-native English language speakers. In addition, corpus-informed approaches can be expanded and further explored to study additional features included in the process of learning to write through L2 peer communication.

The academic writing course consisted of first year master students (N=13; male N=2; female N=11) from various disciplines, studying at the University of Tartu, Estonia. All students were English L2 learners and the entry requirement was set at a B2 level according to the Common European Framework of Reference. The official language of learning and instruction at the University of Tartu is Estonian. Students had not participated in any English academic writing courses or used peer feedback on writing as means to assist the writing process prior to this course. Students were informed and instructed about good peer feedback processes and practices at the beginning of the course using a web-based tutorial offered by the University of Pittsburgh's peerfeedback.net. The tutorial specifically targets students using SWoRD as a means to provide peer feedback on writing.

Students were asked to write an argumentative academic essay over a period of eight weeks. The writing process was broken down into three stages, and students were given a maximum of one week to complete and upload their text in SWoRD and an additional five days to provide feedback to three randomly selected peers. Once feedback was provided, students had another week to make revisions to their original uploaded text. Prior to the first stage of the writing process, students were asked to submit a structured outline of their topic, including a list of resources and references. To ensure a suitably narrow topic for writing, the instructor of the course assessed the outline before allowing the students to proceed any further. The first stage of the writing and feedback process was writing a draft introduction. Students were instructed to

submit their draft to the SWoRD system for peer feedback (average length of the text ranged between 115 and 310 words). Once the feedback process was completed, students were allocated time to make changes to their draft introduction before proceeding to the next stage of the assignment.

For the second stage of the writing process, students were asked to write a draft version of their body text. Students submitted their draft version of the body text in the online system (average length of the text ranged between 500 and 1100 words), including the revised version of the introduction text. Including the revised version of the introduction text meant that peers had a clear overview of the content of the text they were being asked to comment on. Students were prompted to comment only on the draft version of the body text. Once the peer feedback process on the draft was completed, students could make changes to the draft of the body text in order to proceed to the next stage.

The final stage of the writing process was writing a draft of the conclusion text. Again, students uploaded their draft to the online system (average length of the text ranged between 125 and 400 words), including the revised version of the body text and previously revised introduction text. Three peers were assigned for feedback and prompted to comment only on the draft conclusion. After the comments were received, students could make changes to the draft of the conclusion text. On completing these three stages, students were instructed to submit their completed text to the system (revised conclusion and the already revised introduction and body text). No additional peer feedback was given on the final submission. Proceeding through these stages ensured that for each stage of the writing process two versions of one text could be used for comparison: a draft version and a revised version after peer feedback.

Reviewing prompts

During each revision round, students were prompted to comment on the following aspects of the text under review: logic and support, and fundamental writing issues. Logic and support focused on the reasoning and support provided for the main argument. Fundamental writing issues focused on lower-level features such as spelling and grammar, word choice, and sentence structure. Overall the comment prompts served as a general guideline for students and to focus on general writing problems. In addition, students were asked to rate, on a seven-point scale (1=lowest; 7=highest), the text for audience awareness, sentence fluency, word choice, and writing conventions. Except for the introductory tutorial and discussion at the beginning of the course, no additional peer feedback guidance was offered during the course. In addition, the instructor of the course did not take part or intervene in the peer feedback process during the course.

SWoRD collected all the anonymous comments and ratings provided by students and compiled these in a downloadable database file. As this study focused on specific features of the reviews itself, feedback instances were segmented to include only a single reference to a particular suggested change to be made. A peer review could

make references to a number of changes, for example, pointing out concrete spelling mistakes and the need to include more concrete background information. In this case, the review was segmented into two feedback instances; one containing the part of the fragment related to the spelling mistakes and the other containing the fragment related to background information. After segmentation, the corpus contained 374 feedback instances.

Coding of data

The corpus was coded both manually and semi-automatically, using the 'Find' function in a text editor to locate the features that had previously been identified for analysis. These features for analysis were broken down into three different categories: linguistic features, review features, and task features (see Table 1). Manual coding was mainly carried out for the review features, and for the feature mitigation included in the linguistic features category. A semi-automatic coding procedure was performed on the linguistic features; specifically, the data were first located using the 'Find' function, and then checked manually to ensure the validity of the located feature. For example, specific nouns and verbs tagged could be part of a reference made to the text and therefore not included as a linguistic feature contained in the feedback.

Table 1. Features for analysis (including coding label)

Feature	
Linguistic features	Mitigation (SugType); Linguistic modality suggestion (modal) verbs (SUGmark); personal pronouns (PerPronoun); Location nouns and prepositions (LOCmark); Error nouns (ERRmark); Idea verbs (IDEmark); Negative words (NEGmark)
Review features	Feedback Type: directive, nondirective (ComType); appraisal, critical (ComAppraisal); Mentioned (Others); Solution offered (Solution).
Task features	Feedback Implemented (Implemented); Feedback length; Flesch reading ease (FleschRead).

2.2 Linguistic features

The selection of the linguistic features listed in Table 1 is based on previous research, which has either determined that these influence students' understanding of the reviews offered to them, or characterize a specific function of language.

Mitigation

Mitigation has been negatively associated with students' willingness or ability to implement changes offered in reviews (Hyland & Hyland, 1998; 2001). More specifically, the use of mitigation has been determined to be a source of confusion for

L2 learners who are not familiar with the constructs of mitigation, or would rather receive more concrete comments (Nguyen, 2008). As mitigation is primarily used to soften the strength of a given comment, the reception by the receiver may therefore also lessen the impact or necessity of the claim made. On the other hand, the use of mitigation has also been positively attributed to a student's agreement with the reviewer who, as a result of being more 'polite', is considered to have a higher personal integrity (Neuwirth, Chandhok, Charney, Wojahn, & Kim, 1994). In both cases, it can be assumed that the use, or non-use of mitigation may affect a student's willingness to make changes. Martínez-Flor's (2005) taxonomy of linguistic realization of strategies concerning the act of suggesting was used to code the feedback instances. Martínez-Flor identified three types of suggestions: direct, conventionalized, and indirect. Each type contains a number of strategies. For example the feedback instance: "There is a phrase many trends. It is quite hard to grasp the content. The author should at least bring some kind of examples of those trends." was coded as being a conventionalized type, using the should strategy. For the purpose of this study, only this type of mitigation was included in the analysis, as the coding of the strategies closely resembled the linguistic modality suggestion (modal) verbs feature described below. To test for inter-rater reliability, 50 randomly selected feedback instances were analyzed by two independent researchers for the type of suggestion resulting in an inter-rater reliability score of .90 (Cohen's Kappa).

Linguistic modality suggestion (modal) verbs

Linguistic modality is a form of mitigation; however, modality, in comparison to mitigation, can include strong forms of expressions (e.g. can and must). This feature only includes modal verbs ("may", "must", "would", "could", "should", "will", "can", "might", "need", "shall") and verbs which indicate suggestions (e.g. "suggest", "recommend", "advise", "encourage", "urge"). These modal and suggestive verbs express, in the context of the feedback, a subjective attitude and expression of certainty, probability, desirability, and obligation (Bybee, Perkins, & Pagliuca, 1994). Decoding these expressions may offer a challenge to the receiver and therefore influence implementation. For example the feedback instance: "Maybe it would have been better to mention different authors, whose work will be used to defend author's position." expresses an uncertain statement marked by both the use of maybe and would have been, which could either indicate that the reviewer is not sure if the comment is valid, or is merely being polite. As the comment can be interpreted differently, it is up to the receiver to decode the meaning. Given the background of the students and the use of English as a second language in the feedback instances, it can be assumed, based on L2 pragmatic studies, such as those conducted by Nguyen (2008), that the native language may influence the way students both use and interpret the L2. According to a politeness study conducted on Estonians, Estonian speakers, like Germans, use a higher level of directness in requests (in comparison to speakers of English), use fewer polite words, and communication tends to focus on content rather than building relationships

(Keevallik & Grzega, 2008; Keevallik, 2005). This being the case, it may be assumed that the reviewers in this study would therefore prefer a more direct approach to communicating their feedback, and as receivers of the feedback would be likely to value direct expression more than politeness. Feedback instances were semi-automatically coded for the inclusion or exclusion of this feature.

Personal pronouns

Personal pronouns have not been extensively studied in peer feedback, but are commonly analyzed in politeness studies as they measure a person's involvement in the act of communication (Brown & Levinson, 1987; Helmbrecht, 2001). In the case of peer feedback, the personal pronouns "I" and "me" are personal expressions relating to the comment made, whereas pronouns such as "the author", "you" or "none" may demonstrate a lack of personal involvement. A more personal involvement may lead to a much stronger review; however, there is a risk, if the review is too critical, that the receiver may reject the comment. On the other hand, less involvement is safer (or more polite) from a reviewer's perspective but might have the risk of weakening the advice. From a linguistic perspective, an important politeness strategy is to address people indirectly (Brown & Levinson, 1987). Based on the same politeness study referred to earlier (Keevallik & Grzega, 2008; Keevallik, 2005), Estonian speakers tend to use more frequently impersonal forms or avoid personal references altogether. As feedback instances could contain more than one personal pronoun, the part of the comment explicitly directing a need for improvement was used for coding. For example, the following feedback instance: "Introduction is overloaded. I can recommend to shorten this part and to delete some facts, because the author can paste them in future." was coded as first person pronoun. Instances were coded for second person if the review referred to the audience of the writer (marked by you/your) and third person if the review referred to anyone else (e.g. it, the author).

The following linguistic features are motived by and adapted from a study on understanding perceived peer-review helpfulness using natural language processing (Xiong & Litman, 2011).

Location Nouns and prepositions

Research suggests that an important characteristic of successful feedback (i.e. feedback that leads to implementation) is that it contains language features that explicitly locates the problem in question (Van der Pol, Admiraal, & Simons, 2006; 2010; Nelson & Schunn, 2009; Xiong & Litman, 2010; Xiong, Litman, & Schunn, 2012). These include prepositions and nouns such as "on", "in", "page", "paragraph", "sentence", "phrase", "before", "after", etc. Instances were coded for the inclusion or exclusion of this feature.

Error nouns

Error nouns may suggest that the review includes a problem. Problem nouns, such as "error", "mistake", "fault", "inaccuracy", "problem", "lack" etc., could have a negative or positive impact and may therefore influence implementation. The inclusion of these may emphasize the need for revision. Instances were coded for the inclusion or exclusion of this feature.

Idea verbs
The inclusion of idea verbs in feedback (e.g. "consider", "use", "look at", "note", etc.) offers the reader suggestions to a possible course of action to follow. Instances were coded for the inclusion or exclusion of this feature.

Negative words
Words, such as "bad", "wrong", "poor", "hard", "difficult" etc. included in the feedback instance could raise a greater awareness of a problem which needs changing. On the other hand, the use of negative words may also cause the writer to feel uncomfortable. It is unknown how this may affect implementation. Instances were coded for the inclusion or exclusion of this feature.

2.3 Review features

Review features refer to the type of feedback given and are partly motivated by previous research conducted by Cho et al. (2006). Review features include specific aspects related to the reviewing process and are not directly linked to any specific linguistic features.

Directive/Nondirective
As indicated, there seem to be some discrepancies between directive and nondirective comments in terms of their effectiveness on writing performance. This feature has been included in this study to serve as a cross-reference to previous studies. The coding scheme developed by Cho et al. (2006) was applied for the coding of this feature. For example, the following feedback instance: "The text doesn't really catch my attention. The author should try to make the research topic more interesting and give reasons why is this research new and special and what differentiates it from previous researches." was coded as directive as the reviewer suggests a specific change particular to the writer's paper. In contrast, the following example: "There was some spelling mistakes." was coded as nondirective as the reviewer suggests a nonspecific change that would apply to any paper and comments on a detail without suggesting a change.

Feedback instances were coded as either directive or nondirective. To test for inter-rater reliability, 50 randomly selected feedback instances were analyzed by two independent researchers resulting in an inter-rater reliability score of .91 (Cohen's Kappa).
Praise/Criticism

Analyses of praise and criticism, as indicated earlier, have also produced mixed results in previous research. The inclusion of this feature is, like the previous feature, included as a cross-reference to previous studies and uses a similar coding principle as referred to in the directive and nondirective comments. However, as feedback instances could include both praise and criticism, the coding was applied according to the pattern in which they were included. For example, feedback instances could first offer general praise, followed by criticism, or visa versa. As a result, feedback instances were either labeled as praise, criticism, praise and criticism, or criticism and praise. To test for inter-rater reliability, 50 randomly selected feedback instances were analyzed by two independent researchers resulting in an inter-rater reliability score of .91 (Cohen's Kappa).

Mentioned

As the peer feedback process included reviews from three peers, this feature indicates whether other reviewers have also made a comment to the same aspect that needs a writer's consideration. As has been pointed out in previous research (Cho and Schunn, 2007), receiving feedback from multiple peers can increase the persuasiveness of the feedback and therefore increase the likelihood of uptake. Accordingly, it can be assumed that this will have a positive impact on implementation. Feedback instances were coded as either being mentioned by others or not.

Solution offered

Independently from both the feature directive/nondirective and praise/criticism, which also indicate whether suggestions are offered or not, this feature strictly indicates whether the peer feedback instance includes a concrete, explicit solution which can be directly applied to the text by the writer. For example the feedback instance: "There is too little background information. The author should bring out more specific facts from previous researches." offers a suggestion, but does not concretely provide a solution such as highlighted in the following example: "In the 5th line the sentence ..."subsequently it is quite little stories related about man and women". Perhaps a better wording would be: subsequently there are quite a few stories related to love between a man and a woman.". To test for inter-rater reliability, 50 randomly selected feedback instances were analyzed by two independent researchers resulting in an inter-rater reliability score of .80 (Cohen's Kappa). The threshold for implementing a concrete change, as provided in the second example, may be much lower and may therefore have a positive impact on implementation.

2.4 Task features

Task features refer to the students' writing process and the length of the feedback instance. The changes students made to each single draft version of the three separate parts of the text were tracked using the 'Track Changes' function in Microsoft Word and compared to the feedback instance. The use of this feature in the essays provided by the

student meant that implementation as a task feature and dependent variable in this study could easily be identified and located (e.g. see appendix A). To ensure all revisions in the text were included in the analysis, texts were additionally compared using the 'Compare Document' feature in Word. A review was regarded as implemented when the changes in the text could be directly linked to the content of the feedback instance. In addition, a partial implementation was also regarded as implemented. For example, a feedback instance could comment on a general problem, such as "some grammar mistakes". In this case, if a student revised a grammar error, but missed a few other grammatical mistakes, it was still regarded as implemented, as the student had used the feedback instance to address the mistake to which their attention had been drawn. As the feedback instances were extracted to contain only a single reference to a single problem, all the changes made in the texts could be labeled as implemented or not implemented.

In addition to implementation, the length of the feedback was also included as a feature. Research has indicated that the length of sentences and the number of words included in sentences strongly affects the readability of the text (Flesch, 1948). Readability testing is commonly applied to gain a better understanding of the simplicity or complexity of the produced text and often compared to a specific grade level or reading ease. The readability test used for this study is the Flesch–Kincaid readability test (Flesch, 1948). This test uses a formula that includes the total number of words used, the total number of sentences, and the total number of syllables, and provides a score between zero and 100. A score between 60 and 70 is considered as plain English; a score above 70 is regarded, on a scale, to be easier to read; a score below 60 is regarded, on a scale, more difficult to read (Flesch, n.d.). The Flesch–Kincaid readability test provides some indication about the intelligibility of feedback, the assumption being that plain English could have a positive impact on implementation.

Both task features were included in the analysis of the linguistic features and review features; implementation being the dependent variable under investigation.

3. Statistical analysis

A corpus-based machine learning approach was employed to investigate which of the linguistic and review features influence a student's choice to implement changes offered by peer feedback. Machine learning is being increasingly used for the analysis of corpus data in general, and in the research area of Natural Language Processing in particular (Hu and Atwell, 2003; Xiong, Litman, & Schunn, 2012). In addition, machine learning has also proven to be a valuable tool for the investigation of learner corpora (Pendar and Chappell, 2008). Machine learning uses algorithms for the analysis of specific instances (data) to produce more generalizable models which can then in turn be reapplied to the knowledge discovery process. As a result, applying machine learning algorithms could assist in gaining a better understanding of online peer feedback processes (linguistic and review), as well as in predicting which features may

influence implementation. In order to make predictions, only the instances where students indicated a problem or suggested change were used in the analysis, which resulted in 253 reviews of the total 374. Of these 253 reviews, 89 resulted in implementation and 164 did not (approximately 35 and 65 percent of the instances). The data analysis involved three steps.

Firstly, as the study aimed to investigate the relationship between linguistic and review features (independent variables) and task feature implementation (the dependent variable), three statistical prediction algorithms were used to build three separate models to determine which of the algorithms offered the best predictive results. To increase the reliability of the analysis, the tests were carried out on two segments of the data: a training set, used to learn the provided model, and a testing set, used to validate the model. The training and testing sets were generated using a ten-fold cross-validation procedure. Ten-fold cross-validation increases the accuracy of the performance by partitioning the data into ten equal sized segments. During each round, 90 percent of the data is selected for training and the resulting model is tested on the 10 percent held out during the first iteration. This process is repeated ten times and a different fold of the ten percent of the data is held out for validation.

The statistical algorithms used to train and test the models were logistic regression, decision tree, and random forests. Logistic regression is a commonly used statistical method for the analysis of corpus linguistic data. Logistic regression determines how multiple independent variables (the different features included in this study) interact with a binary dependent variable (e.g. implementation). The outcome is expressed in a statistical model with predicted values. The predicted values express how well the model fits the actual observed data in the corpus. The second algorithm, decision tree, also referred to as classification and regression trees, is often used in machine learning, and like logistic regression is designed to mine a corpus to find all possible relationships between independent variables and a dependent variable. Starting at the root, decision tree divides the data into sets, called branches, by classifying the next best feature into different branches or leaves. For example, for this data, decision tree should calculate the input that best guesses whether feedback instances are implemented or not, and does so by finding the feature that best or next best predicts implementation until a tree has been built that demonstrates implementation. Decision tree, like logistic regression, is used to create a model predicting the outcome of implementation and is a promising method to use for the investigation of learner corpora data. Decision tree is specifically useful when a large number of features are used for exploration (Pendar and Chapelle, 2008). The third method, random forests (Breiman, 2001), is another frequently used algorithm in machine learning. Fundamentally, random forests is a development of decision tree modeling, in that it creates different sets of decision trees from the data and combines the predictions from all the trees. Random forests selects many samples from the data. A decision tree is then fitted to each of these samples, creating many decision trees. The accuracy of the analysis is calculated for each sampled observation and predictions are made for every

observation. These predictions are then averaged out for all the observations made. Random forests is a more robust classifier in comparison to decision tree, but the results are more difficult to interpret (Ho, 1998; Breiman, 2001).

Secondly, based on the outcome of the first step, the statistical algorithm was selected which best predicts the response for implementation. Once the method is selected, the results were compared for the feature models (linguistic, review, and all features) and the feature model was selected which performs the best: i.e., which predicts implementation most successfully. The selected model was then validated using the same procedure as described in step one; however, this time, the procedure of ten-fold cross-validation was repeated 10 times to eliminate any false positive errors. In addition, precision-recall and F-score were used to compare the quality of the classifiers produced by the cross-validation. Precision and recall are commonly used in information retrieval to compare the expected results (in this case predicted implementation) to the effective result (the actual analysis of implementation in the data) (Manning, Raghavan & Schütze, 2008).

Finally, in order to gain a better understanding which of the selected features influence implementation, a closer analysis was carried out on the best performing feature model (linguistic, review, and all features) using the best preforming algorithm. All statistical analyses were carried out in R (R Development Core Team, 2008).

4. Results

The first step of the analysis was to determine: 1) which statistical algorithm (logistic regression, decision tree, and random forests) would generate the best prediction of the three feature models, and 2) which feature model (linguistic, review, and all features) offers the best results in terms of predicting feedback implementation.

The results are graphically plotted as ROC (Receiver Operating Characteristics) curves, which are commonly used in machine learning. ROC curves display (on the x-axis) the rate of false positives (in the case of this data, instances where the analysis falsely predicts implementation), and (on the y-axis) the rate of true positives (instances correctly predicting implementation). In the case of classifying implementation correctly, the instances are sorted according to their rank, best ranking classification (instances correctly predicting implementation) appearing first. A good classification would therefore closely resemble a perfect corner (A perfect ROC curve rides the top left corner of the ROC plot). The area under the curve (AUC) measures the goodness of fit of the model, where an ideal model would have a value of 1.0. The generated ROC curve for this data assumes that if the included features in the model randomly predict implementation, the result would be an AUC of 0.500; however, if the features are ranked in such a way that they influence implementation positively, the results should be greater than 0.500: i.e., greater than chance. Table 2 shows the results of the three feature models and the three algorithms used to test the models (see appendix B for the ROC curves). Based on the results, we can deduce that of the three methods used,

random forests performs best on the model including only review features and the model including all features with an AUC of respectively 0.692 and 0.722. For the model including only the linguistic features, decision tree appears to score slightly better than random forests with an AUC and confidence interval of 0.590 (± 0.078) for random forests and 0.638 (± 0.077) for decision tree. However, as the difference in method accuracy between random forests and decision tree is not significant, random forests was used for further comparative analysis of the three feature models.

Overall, in comparison to the results obtained with random forests and decision tree, Table 2 shows that logistic regression performs poorly and appears to overfit the data. Logistic regression models with large numbers of features and limited amounts of training data are highly prone to overfitting. As a result, logistic regression is not further considered as a method for this dataset.

Table 2. Testing logistic regression, random forests, and decision tree on the three models.

Feature model	Logistic regression	Random forests	Decision tree
		AUC (± 95% of C.I)	
Linguistic features	.592 (± .074)	.590 (± .073)	.638 (± .077)
Review features	.521 (± .080)	.692 (± .066)	.650 (± .066)
All features	.501 (± .075)	.722 (± .062)	.644 (± .073)

As the study aimed to investigate which linguistic and review features influence the process of implementation, a comparative analysis was carried out on the three models: a model containing only linguistic features, a model containing only the review features, and a model containing both linguistic and review features. The results of the random forests on the three models, as presented in Table 2, indicate that the model including all features is the most accurate classifying correctly implementation (with an AUC of 0.722 ± .062). As might be expected, the model containing all features outperforms the other two models. It is worth mentioning that random forests tends to overfit less in comparison to decision tree (Breiman, 2001); therefore, adding more features does not influence the overall performance as might be in the case of decision tree or logistic regression.

In order to determine the reliability of the presented results, the results for the model containing all features was repeated on the ten-fold cross-validation using random forests 10 times. The main reason this step was carried out was to eliminate a false positive error that may occur in ten-fold cross-validation as the training sets overlap. It has been suggested (Bouckaert, 2003) that a repeated ten-fold cross-validation followed by a t-test with a degree of freedom equal to 10 is recommended for validation. Table 3 displays the results of the 10 times ten-fold cross-validation with corresponding t-value. The t-test determines whether there is a statistical difference between two folds. The results, however, did not show any statistically significant differences between each corresponding fold. The negative t-values indicate that the previous fold is more

accurate. Overall, the lack of any statistical difference between the folds suggests that the cross-validation of the data performs well.

Table 3. 10 times ten-fold cross-validation of the model containing all features using random forests and precision recall.

Fold	AUC (± 95% of C.I)	*t*-value	*p*-value	Precision	Recall	*f*-score
1	0.728 (± .069)			0.365	0.258	0.302
2	0.711 (± .063)	0.186	0.852	0.344	0.235	0.280
3	0.723 (± .057)	0.071	0.943	0.476	0.337	0.394
4	0.707 (± .066)	-0.032	0.973	0.315	0.202	0.246
5	0.716 (± .061)	-0.002	0.997	0.356	0.236	0.284
6	0.715 (± .073)	-0.063	0.949	0.373	0.247	0.297
7	0.718 (± .060)	-0.384	0.701	0.312	0.225	0.261
8	0.720 (± .067)	0.021	0.983	0.300	0.202	0.241
9	0.703 (± .060)	0.157	0.875	0.290	0.202	0.238
10	0.734 (± .055)	-0.124	0.901	0.482	0.303	0.372

The next step in the analysis process was to compare the quality of the obtained results using precision-recall and F-scores. ROC curves, as mentioned earlier, correspond to false positives (FP) and true positives (TP), and the AUC is interpreted as the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. As a result, it can be used to evaluate the quality of a classifier, but not the precision. Precision-recall and F-scores can be used to evaluate the result of a particular cut-off of 0.5. ROC and AUC show the results under all possible cut-offs.

Precision is a measure of the ability to retrieve the most precise results, and is calculated by dividing the number of correct results (TP) received by the number of correct results plus the number of unexpected results (FP). Higher precision means better relevance and more precise results, but may imply fewer results returned. Recall, on the other hand, measures the ability to retrieve as many instances as possible that match or are related to a query. Recall is calculated by dividing the correct results (TP) by the combination of correct results and missing results (False Negatives). Thus, recall measures the relevance. The F-score combines the score of precision and recall (Davis and Goadrich, 2006). As a measure to determine how good the classification is, precision is used as an indicator. The baseline measure for AUC is set at 0.530 when predicting the more common category. The results, as shown in Table 3, indicate that the performance of classification is very modest. In other words, although the model, according to the AUC score, performs relatively well, the precision by which we can predict the outcome is moderate.

The final step of the data analysis was to determining which features influenced implementation the most. Given that the model containing all features was determined to be the best model using random forests in previous steps, the analysis was first carried out on this model. A novel feature selection algorithm Boruta (Kursa & Rudnicki 2010) for finding relevant features in random forests was used on the full dataset to determine which features were considered to be the most relevant determining implementation. As can be seen in Figure 1, the results of this analysis identified five important features influencing implementation. Two with a relative high importance mean Z score: Feedback type (ComType), directive or nondirective (1.323) and mentioned by others (Others: 1.233). Three features with slightly lower importance mean Z score: location nouns and prepositions (LOCmark: 0.610), personal pronouns (PerPronoun: 0.491), and Solution offered (Solution: 0.370). The boxplots labeled randMin, randMean, and randMax are permutations of features that indicate random non-informal uncorrelated guesses with a response (Kursa & Rudnicki, 2010). Features which have a Z score higher than the Z score of the maximal shadow attribute (randMax) are claimed to be important. In addition to important feature selections,

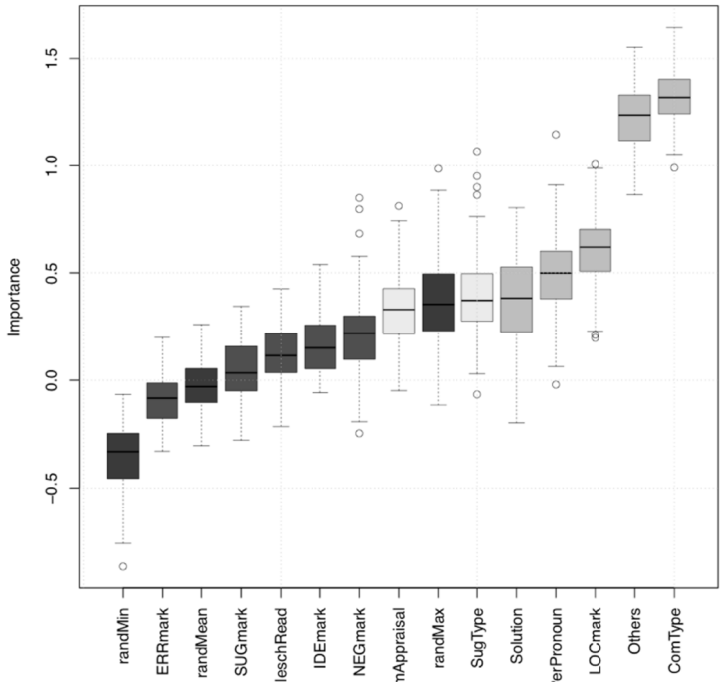


Figure 1: Boruta result plot for important feature selection for implementation for the model containing all features. The randMin, randMean, and randMax boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Boxplots plotted on the left and right of the randMax boxplot are respectively rejected and confirmed attributes.

Figure 1 also includes features rejected as important features for implementation. Rejected features include: Error nouns (ERRmark), Linguistic modality suggestion (modal) verbs (SUGmark), Flesch reading ease (FleschRead), Negative words (NEGmark), and Idea verbs (IDEmark). Those features are mainly represented in the model containing only linguistic features.

These results confirm that the model containing only linguistic features is a relatively weak model; however, two linguistic features, LOCmark and PerPronouns, were identified as being important for implementation, if only slightly so on the importance scale. As the model only containing review features performed fairly similar to all features, an additional analysis was carried out on the full dataset on only the model containing the review features to determine if the exclusion of linguistic features would improve or weaken the importance of the features. Figure 2 shows that the two features, ComType and Others, which received a relative high importance in the model containing all features, are also selected as important features in the model containing only review features. The Solution feature, which was selected as important in the model containing all features, has, however, been identified as not important in the model containing only review features. This may well be as Solution in the all-feature model is presented as a borderline case. If the feature were important, it should have also been selected as important in the feature containing only the review features.

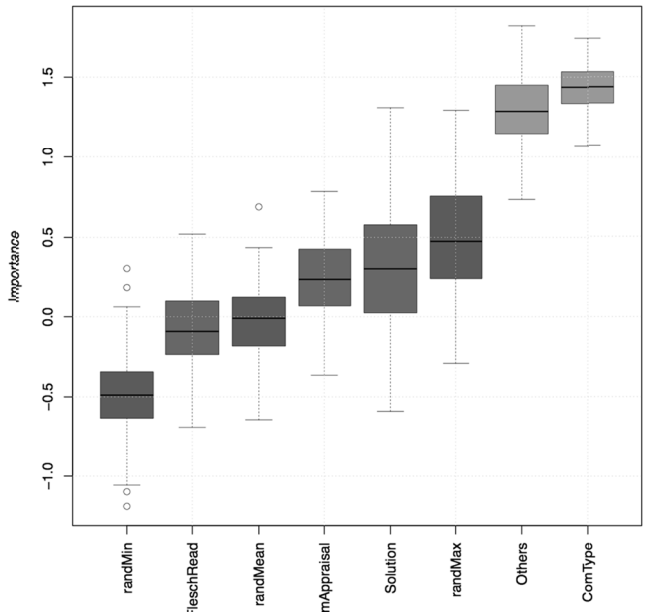


Figure 2: Boruta result plot for important feature selection for implementation for the model containing all features. The randMin, randMean, and randMax boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Boxplots plotted on the left and right of the randMax boxplot are respectively rejected and confirmed attributes.

For comparative purposes, a final model was tested on the full dataset and performed on only the features that were initially selected as important as shown in Figure 1. The results (see Figure 3) also confirm the findings of the analysis of only the review features: Solution was found to be not important for this dataset. One of the linguistic features included in the all feature model (PerPronoun) was also here selected as an important feature, whereas the other linguistic feature, LOCmark, was suggested as being tentative (represented as having a close Z score as the the randMax Z score).

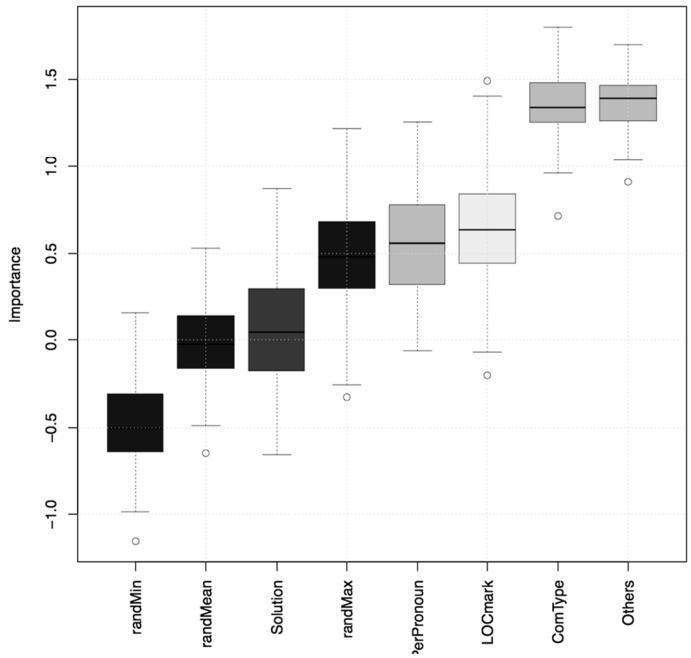


Figure 3: Boruta result plot for important feature selection for implementation for the model containing all features. The randMin, randMean, and randMax boxplots correspond to minimal, average and maximum Z score of a shadow attribute. Boxplots plotted on the left and right of the randMax boxplot are respectively rejected and confirmed attributes.

In brief, based on these three comparative feature selections, the feature ComType and Others seem to influence implementation the most for this dataset. As the results presented in the figures above do not indicate which type of comment (ComType) positively influences implementation: directive or nondirective; nor indicate whether multiple peers pointing out the same topic has a positive or negative influence on implementation, an additional analysis was carried out on the full dataset. Figure 4 shows how the random forests predictions for both ComType and Others are

distributed. The figure displays the density (probability for the variable to fall into a region) and the prediction score for all the trees by major voting. For example, a prediction score by random forests of .60 indicates that 60 percent of the trees voted for the feedback to be implemented. For the feature ComType, peers who were nondirective (light shaded area) are more frequently classified with smaller scores indicating that random forests tries to classify them as not implemented, and peers who were directive (dark shaded area) are more frequently classified with higher scores indicating that they are classified as having a positive influence on implementation. Similarly, for the feature Others, if no, (dark shaded area) it is more frequently classified with smaller scores indicating that random forests tries to classify them as not implemented, and, if yes, (light shaded area) it is more frequently classified with higher scores indicating that they are classified as having a positive influence on implementation. Based on these observations, we can conclude that both directive comments and mentioned by others have a positive influence on implementation.

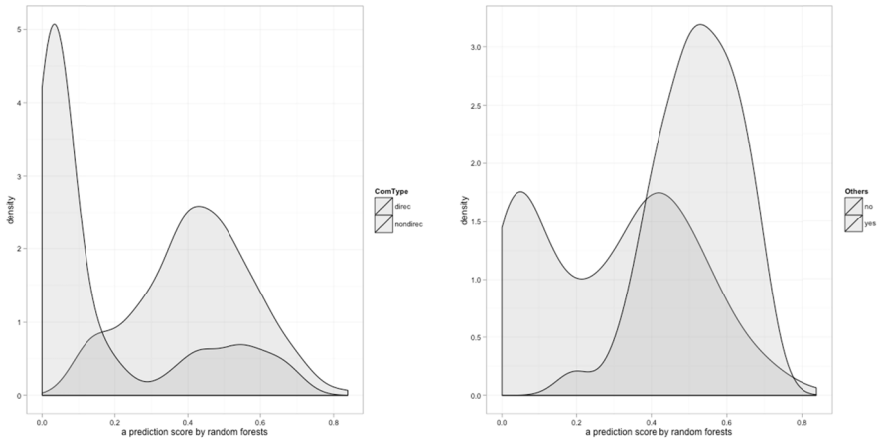


Figure 4: Distribution of predictions by random forests for the feature ComType and Others

5. Discussion

The inclusion of peer feedback on the academic writing process has become common practice in higher education. Given the complex nature of writing, the inclusion of peer feedback offers students multiple perspectives on their progress, besides the feedback given by the instructor. As previous research has shown, however, peer feedback is subject to a multitude of features which may support or hinder the writing process. The aim of this study was to investigate whether, how and to what extent a number of different linguistic and review features of feedback might influence peers to accept or reject revision advice offered in the context of academic writing among L2 learners.

In order to determine implementation, different algorithms were tested for their ability to build predictive feature models. Modeling the features proved to be useful for comparing the performance of the model containing only linguistic features, the model containing only review features, and the model containing all the features. The results indicated that the model containing only linguistic features performed poorly in comparison to the other models, indicating that the linguistic features only offered little evidence for implementation, except for one feature included in the model containing all the features as a predictor for implementation. Both the feature model containing only review features and the model containing all features seemed to perform equally well, which indicate that the review features would seem to have the greatest influence on implementation. In addition, comparing the three algorithms, decision tree and random forests offered a better insight into the performance of the features within the models, in comparison to logistic regression. Specifically for training and testing, logistic regression would need a much larger training set to reduce overfitting. Decision tree and random forests, due to their nature, would seem to be better suited to the analysis of smaller corpora such as that studied here.

Before the analysis, a number of assumptions were made about the linguistic and review features selected for this study and how these might influence implementation. Most of these assumptions were based on the evidence of previous studies, as explained in the data coding section. For example, for the linguistic features included in the analysis, the use of mitigation devices and linguistic modality suggestion and (modal) verbs were expected to influence implementation; however, neither the linguistic feature model, nor the model containing all features, included it as a predictor of implementation. Additionally, mitigation, due to its complex nature for L2 users was expected to negatively influence implementation. The results, however, provided no such evidence.

Although the linguistic model performed relatively poorly in comparison to the two other models, two linguistic features were included as influencing implementation, if only slightly: location nouns and personal pronouns. The identification of location nouns as a relevant feature seems to coincide with the results obtained by Nelson and Schunn (2009), who found that the inclusion of locating statements did influence implementation. The inclusion of personal pronouns in the all feature model as an influencing feature for implementation provided some additional insight, although it cannot be stated with certainty whether personal pronouns in general have a positive or negative influence on implementation, or whether certain individual pronouns (and not others) have a particularly strong effect. Given the linguistic background of the students, it may be the case that relatively impersonal feedback would be more positively valued. (As discussed earlier, politeness studies conducted on the Estonian language usage indicate a general cultural preference for a more impersonal style.) This assumption should, however, be approached with caution and would need to be further investigated on a larger corpus.

The analysis conducted on the models containing only review features and all features has perhaps provided the greatest insight into this small-scale corpus study. The features that were determined to influence implementation were: the type of comment used (directive/nondirective), the repetition of the comment by other peers, and, to a lesser degree, including a concrete explicit solution. As indicated, the type of comment, directive or nondirective, has received mixed results in terms of the effectiveness and usefulness. It can be claimed with some degree of certainty that the type of comment influences implementation; however, claiming that nondirective comments have a negative impact on implementation is not certain, although it does seem to be suggested in this study. This would indicate that, for this dataset, it supports the claim that directive comments are more effective, and refutes the claim made by previous studies that nondirective comments are more effective. Given the background of the students, this suggestion could have some credibility as Estonian speakers tend to be more direct and use less small talk; communication is content driven rather than for relationship building. (Keevallik & Grzega, 2008, Keevallik, 2005). Again, further investigation should provide more insight into this suggestion.

The other two features influencing implementation, as suggested by the two models (review and all features), seem to confirm the findings of Cho and Schunn (2007) who stated that multiple peers pointing to a similar problem benefit implementation. The inclusion of concrete solutions, as pointed out by Nelson and Schunn (2009), was suggested as important only by the all feature model, and rejected in both other models. While this does not rule out the possibility that the provision of concrete solutions may have a positive impact on implementation, it seems clear that reinforcement by other peers influences implementation more than the inclusion of concrete solutions.

Despite initial expectations that the inclusion, or exclusion of specific linguistic features and review features in peer feedback could generate specific predictions as to whether the receiver of the feedback would be expected to implement the changes or not, the analysis of the data has presented merely modest evidence. The models have provided some indication that review features predict better than linguistic features; however, certain linguistic features were nevertheless found to have some influence on implementation. The results have also revealed that, although the models perform relatively well, the precision of the classification is low. This has, however, set a baseline for comparative further research investigating similar features.

The algorithm chosen for this analysis (random forests) has been able to produce some positive results; however, there are some limitations to the exclusive use of random forests modelling. Random forests, in comparison to decision tree, or logistic regression is more difficult to interpret, although the Boruta package analysis used in this study was broadly successful in providing a more simplified visualization that could be used to measure feature importance. Further analysis using different methods, such as logistic regression or decision tree, may give additional interpretative support, thereby building a more detailed picture of how positively or negatively the features

influenced implementation. This would, however, require a much larger data set than was used in this study. As the analysis was conducted on a relatively small corpus (by machine learning standards), sparseness may have influenced the performance of the models; in fact, this was demonstrably the case with the logistic regression algorithm.

Regarding the limitations of the dependent variable, it should be noted that implementation should be carefully considered for further more large-scale corpus studies. Despite being able to link feedback instances to implementation relatively easily due to the extraction of feedback instances to contain only a single review feature, a small number of instances, specifically those referring to more general or higher order concerns in writing, were more difficult to locate or determine as implemented. This study addressed more general comments by making no distinction between partial implementation and full implementation. In addition, the use of the Track Changes function and an additional compare document function simplified labeling implementation; instances that were more difficult to determine may have benefited from an additional rater's perspective – although in mitigation it is worth pointing out that this would likely not have influenced the results as only a small number of these instances were encountered. Nevertheless, this remains an issue that would need to be addressed in future studies. In addition, implementation could be reconsidered so as to contain three levels (implemented, partially implemented, and not implemented) instead of implemented (which included partial implementation in this study) and not implemented.

Despite these limitations, both the method and feature analysis of peer feedback instances show promise. Further research on a much larger L2 learner corpus should offer more conclusive evidence supporting the influence linguistic features and review features have on the peer feedback process on academic writing. Specifically for linguistic features, further studies could expand the analysis of these features to contain additional principles of politeness, a deeper analysis of the various usages of mitigation and modality in peer feedback, as well as culture specific communication patterns influencing the feedback process.

References

Beason, L. (1993). Feedback and revision in writing across the curriculum classes. *Research in the Teaching of English, 27,* 395–421.

Biber, D., Conrad, S., & Reppen, R. (2006). *Corpus linguistics: Investigating language structure and use*. New York, NY, USA: Cambridge University Press.

Bouckaert R. R. (2003). Choosing between two learning algorithms based on calibrated tests. In *The Twentieth International Conference on Machine Learning* (ICML-2003). (pp. 51–58).

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. doi: 10.1023/A:1010933404324

Bybee, J. L. & Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect and modality in the languages of the world*. Chicago: The University of Chicago Press.

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10,* S33–S49. doi: 10.1002/(SICI)1099-0720(199611)10:7<33::AID-ACP436>3.0.CO;2-E

Cho, K., Chung, T. R., King, W. R., & Schunn, C. D. (2008). Peer-based computer-supported knowledge refinement: An empirical investigation. *Communications of the ACM, 51*(3), 83–88. doi: 10.1145/1325555.1325571

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20,* 328–338. doi: 10.1016/j.learninstruc.2009.08.006

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education, 48*(3), 409–426. doi: 10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260–294. doi: 10.1177/0741088306289261

Clare, L. C., Valdes, R., & Patthey-Chavez, G. G. (2000). Learning to write in urban elementary and middle schools: An investigation of teachers' written feedback on student compositions. In *Center for the Study of Evaluation Technical Report*, 526.

Davis, J. & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the International Conference on Machine Learning*. (pp. 233–240). New York: ACM.

Ferris, D. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly, 31,* 315–319. doi: 10.2307/3588049

Flesch, R. (n.d.). *How to write plain English*. Retrieved from http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221–233. doi: 10.1037/h0057532

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication, 37*(1), 16–55. doi: 10.2307/357381

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. In P. Thomson (Ed.). Corpus-based EAP Pedagogy. Special issue of *Journal of English for Academic Purposes, 6*(4), 319–335.

Helmbrecht, J. (2001). Politeness distinctions in pronouns. In M. Dryer, M. Haspelmath, D. Gil & B. Comrie (Eds.), *World atlas of language structures online*. Retrieved from http://wals.info/chapter/45.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 20*(8), 832–844. Retrieved from http://citeseer.ist.psu.edu/ho98random.html. doi: 10.1109/34.709601

Hu, R. & Atwell, E. (2003). A survey of machine learning approaches to analysis of large corpora. In *Proceedings of the Workshop on Shallow Processing of Large Corpora* (pp. 45–52). Lancaster University, UK.

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing, 7*(3), 255–286. doi: 10.1016/S1060-3743(98)90017-0

Hyland, F. (2000). ESL writers and Feedback: Giving more autonomy to students. *Journal of Language Teaching Research, 4*(1), 33–54.

Hyland, F. & Hyland, K. (2001). Sugaring the pill; Praise and criticism in written feedback. *Journal of Second Language Writing, 10*(3), 185–212. doi: 10.1016/S1060-3743(01)00038-8

Hüttner, J. (2010). Purpose-built corpora and student writing. *Journal of Writing Research, 2*(2), 197–218.

Keevallik, L. (2005), Politeness in Estonia: A matter of fact style. In L. Hickey & M. Stewart (Eds.), *Politeness in Europe* (pp. 203–217). Clevedon etc.: Multilingual Matters.

Keevallik, L. & Grzega, J. (2008). A few notes on conversational patterns in Estonian. *Journal for EuroLinguistiX, 5*, 80–87.

Knoblauch, C. H., & Brannon, L. (1981). Teacher commentary on student writing: The state of the art. *Freshman English News, 10*(2), 1–4.

Leki, I. (1990). Coaching from the margins: Issues in written response. In B. Kroll (Ed.), *Second Language Writing* (pp. 57-68). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781139524551.008

Manning, C. D., Raghavan, P., Schtze, H. (2008). *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press doi: 10.1017/CBO9780511809071

Martinez-Flor, A. (2005). A theoretical review of the speech act of suggesting: Towards taxonomy for its use in FLT. *Revista Alicantina de Estudios Ingleses, 18*, 167–187.

Kursa, M. B., & Rudnicki, W. R., (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36*(11), 1–13. Retrieved from http://www.jstatsoft.org/v36/i11/

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science, 32*(1), 99–113. doi: 10.1023/B:TRUC.0000021811.66966.1d

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 27*(4), 375–401. doi: 10.1007/s11251-008-9053-x

Neuwirth, C.M., Chandhok, R., Charney, D., Wojahn, P., & Kim, L. (1994). Distributed collaborative writing: A comparison of spoken and written modalities for reviewing and revising documents. In *Proceedings of the CHI'94 Conference on Computer-Human Interaction* (pp. 51- 57). Boston: ACM

Nguyen, T. (2008). Modifying L2 criticisms: How learners do it? *Journal of Pragmatics, 40*(4), 768–791. doi: 10.1016/j.pragma.2007.05.008

Pendar, N. & Chapelle, C. (2008). Investigating the promise of learner corpora: methodological issues. *CALICO Journal, 25*(2), 189–206.

Pridemore, D. R., & Klein, J. D. (1991). Control of feedback in computer-assisted instruction. *Educational Technology Research and Development, 39*(4), 27–32. doi: 10.1007/BF02296569

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, Retrieved from http://www.R-project.org.

Schlitz, S. A. (2010). Introduction to special issue: Exploring corpus-informed approaches to writing research. *Journal of Writing Research, 2*(2), 91–98.

Straub, R. (1997). Students' reactions to teacher comments: An exploratory study. *Research in the Teaching of English, 31,* 91–119.

Strijbos, J.W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency?. *Learning and Instruction, 20*(4), 291–303. doi: 10.1016/j.learninstruc.2009.08.008

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*(4), 339–343. doi: 10.1016/j.learninstruc.2009.08.003

Tseng, S., & Tsai, C. C. (2007). On-line peer assessment and the role of peer feedback: A study of high school computer course. *Computer & Education, 49*, 1161–1174. doi: 10.1016/j.compedu.2006.01.007

Van den Berg, I., Admiraal, W. F. & Pilot, A. (2006). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31*(1), 19–36. doi: 10.1080/02602930500262346

Van der Pol, J., Admiraal, W. F. & Simons, P. R. J. (2006). The affordance of anchored discussion for the collaborative processing of academic texts. *International Journal of Computer Supported Collaborative Learning, 1*(3), 339–357. doi: 10.1007/s11412-006-9657-6

Van der Pol, J., Admiraal, W. F. & Simons, P. R. J. (2010). Integrating an evaluation function in online anchored discussion to increase the local relevance of replies. *Journal for Computers & Human Behaviour, 26*(3), 288–295. doi: 10.1016/j.chb.2007.09.005

Van der Pol, J., Van den Berg, I., Admiraal, W. F. & Simons, P. R. J. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education 51,* 1804-1817. doi: 10.1016/j.compedu.2008.06.001

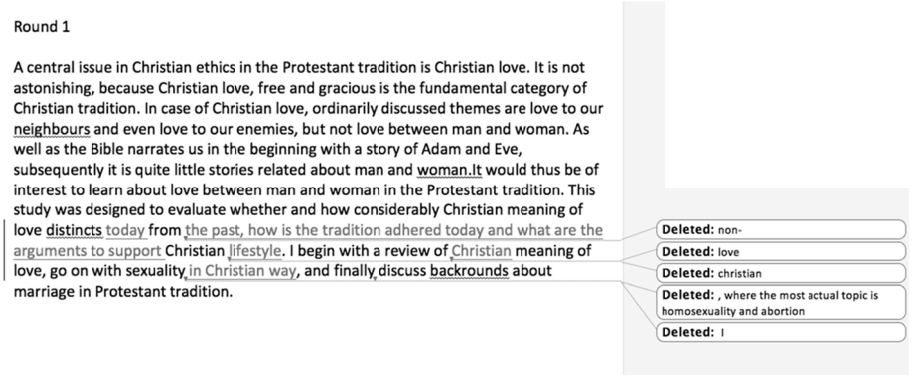
Xiong, W. & Litman, D. (2010) Identifying problem localization in peer-review feedback. In the Tenth International Conference on Intelligent Tutoring Systems. Pittsburg, PA.

Xiong, W. & Litman, D. (2011). Understanding differences in perceived peer-review helpfulness using natural language processing. In *Proceedings sixth workshop on innovative use of NLP for building educational applications* (pp. 1–10). Portland, OR.

Xiong, W. Litman, D. & Schunn, C.D, (2012). Improving research on and instructional quality of peer feedback through natural language processing. *Journal of Writing Research, 4*(2). 155-176.

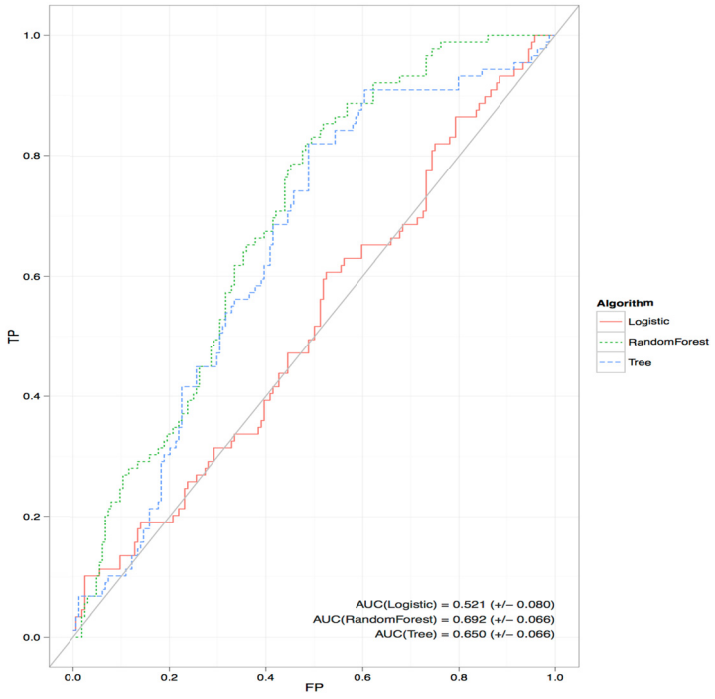
Appendix A

Example of track change feature in word to monitor students' changes in their text for each draft.

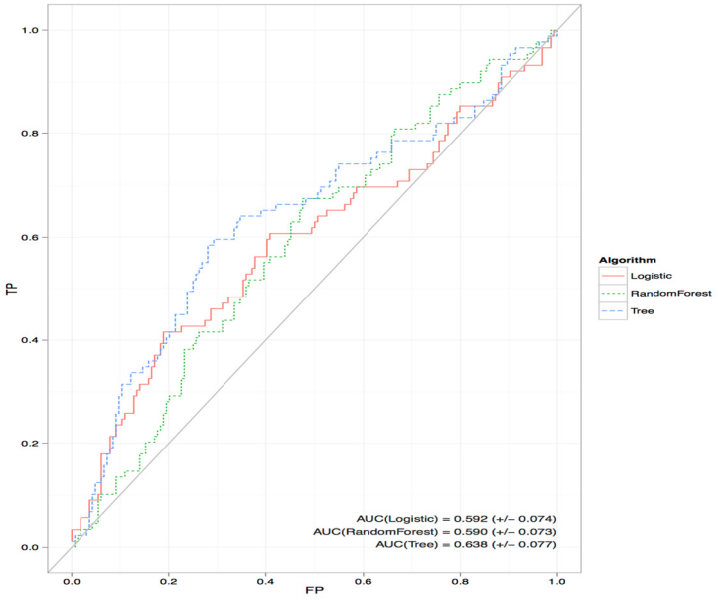


Appendix B

1. ROC curve Linguistic features



2. ROC curve Review features



3. ROC curve All features