Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education. New Zealand: Asia-Pacific Society for Computers in Education

Automatic Classification of Teacher Feedback and Its Potential Applications for EFL Writing

Gary CHENG[a*], Shu-Mei Gloria CHWO[b], Julia CHEN[c], Dennis FOUNG[c], Vincent LAM[c], & Michael TOM[c]

[a]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*

[b]*Department of Applied English, HungKuang University, Taiwan*

[c]*English Language Centre, The Hong Kong Polytechnic University, Hong Kong*

*chengks@eduhk.hk

Abstract: This paper presents and discusses the initial data from a project that aims to develop a system for automatic tracking of student responses to teacher feedback in draft revision. One main purpose of the project is to design and implement a method for automatic classification of teacher feedback on students' draft essays in the EFL context. In this paper, we propose the automatic classification method and evaluate its performance in terms of accuracy. Our findings show that an accuracy of over 96% was achieved when classifying teacher feedback using the proposed method. They also show that the classification results could be analysed with other sets of data such as assessment grades to help teachers reflect on their use of feedback types and refine their feedback practice. This study can provide a basis for future research into automatic analysis of the impact of various feedback types on student revision.

Keywords: Automatic classification, teacher feedback, student essay, EFL writing

1. Background

Feedback is regarded as one of the most significant influences on student learning and achievement (Hattie & Timperly, 2007). In the English as Foreign Language (EFL) context, feedback has a great potential for improving the quality of student writing (Hyland, 2003; Cheng, 2017). A large body of research focused on the role corrective feedback (also known as grammar correction) plays during the process of writing. However, evidence from prior research was inconclusive as to the effectiveness of corrective feedback on student writing. Some studies (Truscott, 2004; Truscott, 2007) argued that this feedback type has little value for EFL development and it can even have an adverse effect on students' ability to write accurately, while some others highlighted the importance of grammatical correctness in academic writing and suggested that it can benefit students' writing (Ferris, 2004, Lee, 2008).

In addition to grammar correction, research was undertaken to identify what other characteristics of teacher feedback are important to substantial and successful revisions of student writing. It was found that longer and text-specific feedback (e.g. criticism on an issue) was more effective in encouraging student revisions than shorter and general feedback (e.g. positive comments) (Ferris, 1997). It was also found that the success of student revisions was associated more strongly with the types of problems identified by the feedback (e.g. incorrect lexical choice, lack of explanation and insufficient details) than the syntactic forms of the feedback (e.g. declarative, question and imperative) (Conrad & Goldstein, 1999).

However, the findings about teacher feedback were primarily drawn from a limited number of studies and student cases in English-speaking countries (Chiu & Savignon, 2006; Ferris, 2003). They may not be generalised to other populations where EFL students are living in their non-English speaking hometown such as Hong Kong and Taiwan. Besides this, the methodology of previous studies was based on manual classification of teacher feedback. It is not appropriate for analysing a large data set of written comments and may give rise to a problem of consistency in classification.

Given the limitations of prior research, this study sought to propose and evaluate an automatic approach to classify teacher feedback on student essays from an EFL writing course at a Hong Kong

university. It also aimed to explore the relationship between feedback types and assessments. Results of this study can offer insights into the effectiveness of using an automatic approach to identify different types of teacher feedback. They can also demonstrate the potential of using the automatic approach to examine the correlation between the number of feedback in different types and the grades given by teachers on student essays. With this initiative, teachers can be provided with some evidence on their use of feedback types on which feedback types they use are weakly linked with their assessments.

This paper is organised as follows. Section 2 proposes the classification framework for teacher feedback. Section 3 provides details of the automatic classification method for teacher feedback. Section 4 describes the research methodology of the study. Section 5 presents and discusses the results obtained from analysing the data collected during the study. Finally, conclusion and future work are given in Section 6.

2. Classification Framework for Teacher Feedback

To characterise the ways that teachers frame their feedback, Straub (1997) identified six categories of teacher feedback. The categories include (1) praise, (2) criticism, (3) imperative, (4) advice, (5) open question, and (6) closed question. Chen and Hamp-Lyons (1999) noted that Straub's (1997) taxonomy of feedback categories was derived from the L1 (first language) context and it mainly focused on the syntactic forms of the feedback, so they added two more categories to fit into the EFL context: mechanics (i.e. the feedback that deals with grammar and punctuation) and '?' (i.e. the feedback that conveys a meaning of 'do not understand'). The current study adapts and extends the models of Straub (1997) and Chen and Hamp-Lyons (1999) by taking into account both text-specific features (e.g. content and language use) and syntactic structures (e.g. imperative and question forms) of teacher feedback on student essays. The proposed classification framework is shown in Table 1.

Table 1: Classification of feedback types.

Code	Category	Description	Example
T1	Praise	Positive comments, non-controlling	Well written
T2	Criticism	Negative comments or evaluations, authoritative	Confusing
T3	Imperative	Comments that tell the student writer to do or change something, usually starting with a verb in the imperative form	Be consistent
T4	Advice	Suggestive comments often in conditional mode	Maybe you could add some details here
T5	Closed question	Questions that either get a 'yes' or 'no' as answer, or else a simple one-word answer	Do you think you have given an adequate evaluation?
T6	Open question	Questions that require more than a 'yes' or 'no' answer, often starting with 'what' 'where', 'why', 'who', when', and 'how'	What does this mean?
T7	Content	Comments that often deal with the clarity and meaning of content, ideas and views	Some ideas need further elaboration
T8	Language use	Comments that often deal with the grammar, punctuation, spelling and word choice	Reword this
T9	Organisation	Comments that often deal with the organisation of ideas and linkage between sentences or paragraphs	In general, the ideas flow well
T10	Referencing and formatting	Comments that often deal with citations, quotations and references	Non-academic sources

3. Automatic Classification Method

The automatic classification method is based on matching teacher feedback against syntactic rules and semantic words extracted from a set of manually annotated data. In this study, a total of 3412 teachers' written comments on 90 students' draft essays were collected to build the data set. According to the proposed classification framework, each teacher comment was manually annotated by two researchers. Discrepancies between the researchers were resolved by discussion to reach consensus on the annotation standard.

The basic unit of annotation was a single sentence. Every sentence in a comment was marked up with two feedback types. One type was concerned about the form (T1 to T6) and the other was about the aspect (T7 to T10). Figure 1 shows a sample student text with annotated teacher comments.

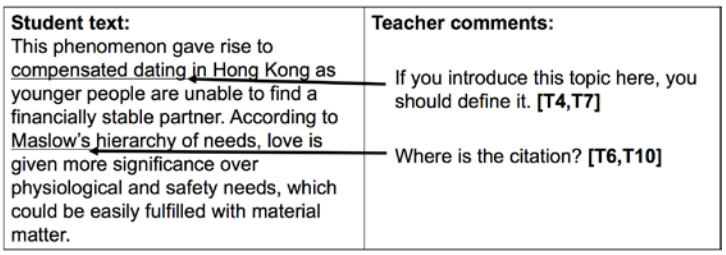

Figure 1. A Sample Student Text with Annotated Teacher Comments.

On the syntactic side, every sentence was processed with a part-of-speech (POS) tagger to assign parts of speech to each word (Bird, Klein, & Loper, 2009). The most common sequences of POS were identified and extracted to form a set of distinctive syntactic rules for each feedback type. On the semantic side, a word-by-type matrix was constructed based on Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998). Each feedback type was characterised by a vector of word weights.

To classify a new comment, we first apply the POS tagger to it at the sentence level. Every tagged sentence will be matched against existing syntactic rules to determine its feedback types. If there are no matches, the sentence will be semantically transformed into a vector of word weights. The vector will subsequently be classified into a feedback type where their cosine similarity is the highest among other types.

4. Research Methodology

4.1 Study Context

This study is part of the research project namely "Towards automatic tracking of student responses to teacher feedback in draft revision". It was undertaken at the English Language Centre of a Hong Kong university in the first semester of the academic year 2016/17. Participants of this study were students attending the Advanced English for University Studies (AEUS) course. AEUS was a 13-week, credit-bearing course that required students to research for, write, plan and revise an academic position argument essay, and to defend their views and engage with those of others clearly and logically in an mini oral defence. As part of the course requirement, students had to submit two academic position argument essays. The first was a 600-word draft, and the second was a polished, final essay on the same topic of 1200 words.

4.2 Participants

Ninety-two undergraduate students (30 males and 62 females) enrolled on AEUS gave their written consent to participate in this study. Their ages ranged from 17 to 21 years (M=18.15 and SD=0.94). They came from 6 class groups and 5 academic disciplines: Advertising Design (1 group), Accounting and Finance (1 group), Mental Health Nursing (1 group), Nursing (2 groups), and Physiotherapy (1

group). Three English language instructors (IA, IB and IC) who taught the participating class groups were also involved in this study. Details of the participating class groups can be found in Table 2.

Table 2: Details of participating groups.

Group Code	Programme of Study	No. of Participants	Instructor Code
A&F	Accounting & Finance	20	IA
AD	Advertising Design	11	IC
MHN	Mental Health Nursing	11	IB
N1	Nursing	16	IA
N2	Nursing	15	IB
P	Physiotherapy	19	IB

4.3 Data Collection and Analysis

Students in the AEUS course were required to submit a 600-word draft essay on a topic of their choice by drawing on academic sources such as peer-reviewed journal articles at Week 7. They received letter grades and written feedback on their draft before preparing and submitting their final essay by Week 11. The assessment criteria and weighting for the draft essay were: content (30%), organisation (20%), language (30%) and referencing (20%). Primary data of this study were the feedback on the draft and the assessment results. The feedback was categorised manually by researchers and automatically by a classification tool implemented in Python (Bird el al., 2009), respectively. Accuracy of the automatic classification method was measured by the proportion of machine classifications that agree with manual classifications. Five-fold cross-validation was performed, with four-fifth of the annotated feedback extracted as the training data and the remaining as the testing data. This procedure was repeated five times until all feedback in the original data set was tested once. The classification accuracy was calculated as the average over the five iterations. Additionally, letter grades in assessments were converted to numerical scores following university guidelines (A+ = 4.5, A = 4, B+ = 3.5, B = 3, C+ = 2.5, C = 2, D+ = 1.5, D = 1, F = 0). The results of classifications along with assessment scores were used to calculate the correlation between feedback types and assessments given by teachers.

5. Results and Discussion

5.1 Accuracy of the Automatic Classification Method

To contrast the actual and predicted classifications of teacher feedback and report the accuracy of the automatic classification method, a two-way contingency table known as a confusion matrix is shown in Table 3. In the confusion matrix, each row (or column) refers to the count of a feedback type identified by human (or machine). The numbers of correct classifications (i.e. the results of machine classification are identical to those of the manual classification) are represented by diagonal cells, while the numbers of mis-classifications are represented by off-diagonal cells.

A confusion matrix is a good way to illustrate the accuracy of the automatic classification method. As can be seen in Table 3, the proposed method could identify different types of teacher feedback with high degrees of accuracy ranging from 96% upwards. The results are very encouraging and suggest that it is feasible to classify teacher feedback in an automatic way.

5.2 Correlation between Feedback Types and Assessments

Pearson correlation coefficients were calculated between the number of feedback in different types and the assessment grades. Table 4 provides a summary of the correlation coefficients. It shows that the number of feedback in several types are significantly correlated with the assessment grades in their corresponding areas. These can be found between most types of feedback (except advice) on content and the assessment in content ($|r|$ = .248 to .418, $p < .01$), two types of feedback (advice and praise) on organisation and the assessment in organisation ($|r|$ = .223 to .450, $p < .01$ or .05), three types of

feedback (praise, criticism and imperative) on language and the assessment in language ($|r|$ = .361 to .489, $p < .01$), and three types of feedback (praise, criticism and imperative) on referencing and the assessment in referencing ($|r|$ = .419 to .564, $p < .01$). The results indicate that the assessment grades are correlated more with praise, criticism and imperative but less with advice, closed question and open question. They imply that teachers tended to give comments in a strong authoritative mode rather than in a less controlling mode (e.g. suggestions or hints) if a draft essay needed more changes. Given this kind of statistical summary, it would be helpful for teachers to reflect on their use of feedback types and refine their feedback practice to benefit students' writing. Further, it would be conducive to the investigation of the impact of various feedback types on student revision in the final essay.

Table 3: Confusion matrix.

		Machine Classification											Accuracy
		T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	**T10**	**Unclassified**	
Manual Classification	**T1**	197	0	0	0	0	0	0	0	0	0	0	100%
	T2	46	1722	4	3	0	0	0	0	0	0	7	96.6%
	T3	8	2	533	0	0	0	0	0	0	0	0	98.2%
	T4	5	2	1	255	0	0	0	0	0	0	0	97.0%
	T5	0	0	0	0	246	4	0	0	0	0	0	98.4%
	T6	0	0	0	0	9	368	0	0	0	0	0	97.6%
	T7	0	0	0	0	0	0	1266	5	4	5	0	98.9%
	T8	0	0	0	0	0	0	32	1503	18	3	0	96.6%
	T9	0	0	0	0	0	0	1	0	164	1	0	98.8%
	T10	0	0	0	0	0	0	13	0	1	402	0	96.6%

Table 4: Correlation between the numbers of different feedback types and the assessment results.

Feedback Types	**Assessment Results**			
	Content	Organisation	Language	Referencing
Content (Praise)	.418**			
Content (Criticism)	-.293**			
Content (Imperative)	-.315**			
Content (Advice)	.045			
Content (Closed Question)	-.314**			
Content (Open Question)	-.248**			
Organisation (Praise)		.450**		
Organisation (Criticism)		-.126		
Organisation (Imperative)		.014		
Organisation (Advice)		.223*		
Organisation (Closed Question)		.197		
Organisation (Open Question)		.049		
Language (Praise)			.489**	
Language (Criticism)			-.395**	
Language (Imperative)			-.361**	
Language (Advice)			-.071	
Language (Closed Question)			-.001	
Language (Open Question)			-.033	
Referencing (Praise)				.419**
Referencing (Criticism)				-.564**
Referencing (Imperative)				-.535**
Referencing (Advice)				-.123
Referencing (Closed Question)				-.064
Referencing (Open Question)				-.194

*$p < .05$ (2-tailed), **$p < .01$ (2-tailed)

6. Conclusion and Future Work

This study proposes a method for automatic classification of teacher feedback on students' draft essays in the EFL context. Drawing on a ten-category classification framework, the proposed method identifies types of teacher feedback by matching against syntactic and semantic features extracted from a set of manually annotated data. The findings of this study show that the proposed method could achieve very good performance in terms of classification accuracy (over 96%). They also demonstrate the potential of using the classification results as a source of reflection to enhance teachers' feedback practice.

Future work involves designing and implementing methods for automatic identification and classification of student revision in the final essay. In addition, the association between different types of teacher feedback on student revision will be examined. This would give insights into what kind of feedback is most effective in facilitating students to make substantial and successful revisions to EFL writing.

Acknowledgements

This research was financially supported in part by the Hong Kong SAR Government under General Research Fund (GRF no. 18608816).

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Chen, J., & Hamp-Lyons, L. (1999). Effective feedback on student writing. In J. James (Ed.), *Quality in teaching and learning in higher education: A collection of referred papers from the first conference* (pp. 113-120). Hong Kong: Hong Kong Polytechnic University.

Cheng, G. (2017). The impact of online automated feedback on students' reflective journal writing in an EFL course. *The Internet and Higher Education*, *34*, 18-27.

Chiu, C.-Y., & Savignon, S. J. (2006). Writing to mean: Computer-mediated feedback in online tutoring of multidraft compositions. *CALICO Journal*, *24*(1), 97-114.

Conrad, S. M., & Goldstein, L. M. (1999). ESL student revision after teacher-written comments: Text, contexts, and individuals. *Journal of Second Language Writing, 8*(2), 147-179.

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *TESOL Quarterly*, *31*(2), 315-339.

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Lawrence Erlbaum.

Ferris, D. R. (2004). The "grammar correction" debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime...?). *Journal of Second Language Writing*, *13*(1), 49-62.

Hattie, J., & Timperly, H. (2007). The power of feedback. Review of Educational Research, *77*(1), 81-112.

Hyland, K. (2003). *Second language writing*. New York: Cambridge University Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284.

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing*, *17*(2), 69-85.

Straub, R. (1997). Students' reactions to teacher comments: An exploratory study. *Research in the Teaching of English*, *31*(1), 91-119.

Truscott, J. (2004). Evidence and conjecture on the effects of correction: A response to Chandler. *Journal of Second Language Writing*, *13*(4), 337-343.

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, *16*(4), 255-272.