Disiplinler Arası Dil Araştırmaları Dergisi [DADA]
Haziran 2022

International Journal of Interdisciplinary Language Studies [IJILS]
June 2022

Atıf (cite): Tömen, Musa (2022). Automated essay scoring feedback in foreign language writing: Does it coincide with instructor feedback?, *Disiplinler Arası Dil Araştırmaları Dergisi,* 4, 53-62.

Automated Essay Scoring Feedback in Foreign Language Writing: Does it Coincide with Instructor Feedback?

Musa TÖMEN[1]

Yabancı Dilde Yazmada Otomatik Kompozisyon Puanlama Geribildirimi Eğitmen Geribildirimi ile Uyumlu mu?

Özet

Bu çalışma, bir otomatik bir kompozisyon puanlama (OKP) sistemi olan Criterion tarafından verilen geri bildirim ile eğitmen geri bildirimini karşılaştırarak, öğrencilerin kağıtlarına mekanik geri bildirim vermede öğretmenlere zaman kazandırmak için kullanılıp kullanılamayacağını bulmayı amaçlamıştır. Bu çalışma Criterion tarafından verilen geri bildirimin eğitmen geri bildirimi ile ne ölçüde örtüştüğünü göstermeyi amaçlayan tanımlayıcı bir çalışmadır. Bu çalışmada, bir değerlendirme sistemi ve bir insan değerlendiriciden alınan geri bildirimler karşılaştırılmıştır. Bu çalışmada "OKP geribildirimi, İngilizce dilbilgisi ve mekanik açısından İngilizce yazmada eğitmen geribildirimi ile ne ölçüde örtüşmektedir?" sorusuna cevap aranmıştır. Sonuçlar, Criterion'ın öğrencilerin kağıtlarındaki dilbilgisi hatalarını ve mekanik hataları bulmada insan değerlendirici kadar tutarlı ve başarılı olduğunu göstermiştir.

Anahtar Sözcükler: İngilizce yazma, otomatik kompozisyon puanlama, dönüt, değerlendirme

Automated Essay Scoring Feedback in Foreign Language Writing: Does it Coincide with Instructor Feedback?

Abstract

This study aimed to find out whether Criterion, an automated essay scoring system (AES), can be used to save time for teachers in giving mechanical feedback to student papers by comparing the feedback given by Criterion with the instructor feedback. This is a descriptive study aiming to show to what extent the feedback given by the Criterion match with the instructor feedback. In this study, the feedback from Criterion and a human rater were compared. This study sought answer for the following research question: To what extent does AES feedback coincide with instructor feedback in English as a Foreign Language (EFL) writing in terms of grammar and mechanics? The results showed that Criterion was as accurate as the human rater in finding the grammar errors and mechanical errors in students' papers.

Key Words: EFL writing, Automated essay scoring, feedback, assessment

Makale Türü: Araştırma

Paper Type: Research

1. Introduction

Recently, there has been a growing amount of research concerning the automated essay scoring (AES) systems in foreign language writing. These systems were first introduced in first

[1] Research Assistant, Anadolu University, English Language Teaching, mtomen@anadolu.edu.tr, 0000-0002-7351-2440

Makale Geliş Tarihi (Recieved): 12.05.2022 Makale Kabul Tarihi (Accepted): 27.05.2022

language environment. They were first intended to reduce the time spent on giving feedback. However, later on these systems were used in detecting foreign language and second language learners' spelling errors, grammar mistakes, etc. The very idea that giving instant feedback to the students effortlessly has made these systems be thought as a help tool for language teachers and schools.

Despite differing points of view and contradictory study findings, feedback on form is critical in second language writing. Grammar accuracy in writing is valued highly by high-stakes tests and English teachers in general (Dikli and Bleyle, 2014). The instructors feel that they do not do their job properly unless they give feedback on grammar mistakes of students in their writing. At this point, it is claimed that human effort can be reduced thanks to technology. Ware (2011) states that AES systems can be used as support for writing instruction while there is controversy in using them in scoring the papers. AES systems can be used in giving mechanical feedback on students' papers. With the help of AES systems, students are expected to get greater autonomy and meet their need for feedback on sentence-level correctness (Ranalli, Link and Chukharev-Hudilanen, 2017).

Owing to the fact that AES systems are capable of evaluating an essay within seconds, they have taken part in writing evaluation since beginning. Educational Testing Service (ETS) has tried to validate its AES tool Criterion in various studies (Weigle, 2010; Weigle, 2011). Therefore, instructional use of AES systems has gained popularity in colleges and universities (Dikli and Bleyle, 2014). There are various AES systems available commercially and non-commercially. The most famous ones are Project Essay Grader by Page and Measurement Inc., Intelligent Essay Assessor and WriteToLearn by Pearson Assessments, Intellimetric and MY Access by IntelliMetric, and e-rater and Criterion by the ETS.

In this study, Criterion, which can provide feedback on five writing traits (grammar, usage, mechanics, style, and organization) was employed. Criterion was also designed for native speakers of English, but it is used widely by English as a Second/Foreign Language learners, as well (Warschauer and Ware, 2006). Criterion was developed and has been employed by Educational Testing Service (ETS) to enable instructors and administrators to focus more on writing by freeing up valuable class time. This study aimed to find out whether Criterion can be used to save time for teachers in giving mechanical feedback to student papers by comparing the feedback given by Criterion with the instructor feedback. This is a descriptive study aiming to show to what extent the feedback given by the Criterion match with the instructor feedback.

2. Literature Review

In foreign language writing, there is an ongoing controversy whether the feedback should focus on the form or on the content. Some argued that if form is emphasized, the students may neglect the content of the writing (Zamel, 1985); while others argued that focusing on grammar does not affect the content negatively (Fathman, 1990). Ferris (1995) also found out that pay-

ing attention to grammar was not ineffective as it was believed. However, it is recommended that teachers should be able to keep the balance between form and content while giving feedback (Ashwell, 2000).

On the other hand, grammar feedback was strongly criticized as it only deals with surface level correction (Truscott, 2007). Truscott (2007) also defined grammar feedback as 'harmful' and it should be refrained in second language writing. However, the popular trend in giving feedback in foreign language writing is grammar feedback because both teachers and students feel themselves safe if the essay is grammatically correct. Students review and improve their writings owing to mistake feedback, students develop accuracy over time because to error feedback, both students and teachers respect error feedback, and written correctness is vital in the real world, according to Ferris (2011).

AES systems were first introduced in 1960s and since then the research has been conducted to prove the validity and reliability of AES systems. AES systems use artificial intelligence to evaluate the essays and to provide feedback. In Asia, several higher education institutions have employed AES systems to reduce grading and instructional workloads (Chen and Cheng, 2008; Long, 2013; Otoshi, 2005). Many studies have reported reliably high agreement rates between computers and human raters (Nichols, 2004; Wang and Brown, 2007). Other studies of Attali (2007); Lee, Gentile and Kantor (2008), which were conducted by ETS, also showed that e-rater, the scoring engine of Criterion, is a reliable tool and can be used to score essays.

There are also studies focusing on the instructional use of AES systems. These classroom-based studies were mainly conducted by independent researchers (Dikli, 2010; Choi and Lee, 2010; Chen and Cheng, 2008) unlike the most of scoring engine studies. In Attali's study (2004), it was found that the students were able to use the feedback given by Criterion effectively and submitted their essays more than once by revising to submit more quality writings.

In their study conducted in EFL environment, Chen and Cheng (2008) examined the use of an AES system in three EFL classrooms and reported their perceptions. They discovered that students disliked the system, but that they liked it better when the AES system was used to assist students rewrite their papers. That is, they favoured the method if it was utilized to help them improve their papers and if the teacher provided comments thereafter. In her study, Dikli (2010) compared the feedback of an AES system and feedback of the teacher. This qualitative research was reported to prove that the AES feedback was generic, lengthy and redundant.

Choi and Lee (2010) looked at the usage of Criterion in a college ESL writing program, as well as the impact of Criterion feedback and teacher feedback on students' work. They discovered that providing these sorts of feedback jointly was most effective.

There were also studies indicating the certain limitations of AES systems. These limitations can be summarized as follows: AES systems require large corpus for training (Dikli, 2010); they are reliable in detecting local errors but not in global concerns (Attali, Lewis and

Steier, 2012); the current technology is not enough to detect local errors in L2 texts (Dikli and Bleyle, 2014); the language used in AES feedback can be complex and L2 learners may not be able to understand it, they may need additional modelling and guidance (Dikli, 2010); and finally AES feedback lacks human interaction and therefore the feedback it provides is potentially confusing for EFL learners (Wang et al., 2013).

Lack of human interaction is, on the other hand, regarded as a potential benefit for L2 writers (Liao, 2016). It relieves the anxiety of the writer as it triggers the sense of objectivity. Moreover, as AES systems provide instant feedback and grammatical explanation, immediate communication lets the students see and revise their errors within seconds of submitting their drafts. This instant feedback enables students to focus on specific linguistic features and subsequently improve their writing and gain writing confidence by being aware that they do not have grammatical mistakes in their writing (Chen and Cheng, 2008). As Dikli and Bleyle (2014) puts forward, even the most efficient human reader cannot outperform an AES system in providing instant feedback.

With this regard in mind, this present study aims to find out whether an AES system, Criterion, can be used in EFL writing classes. In this study, the feedback from Criterion and a human rater will be compared. This study seeks answer for the following research question:

1. To what extent does AES feedback coincide with instructor feedback in EFL writing in terms of grammar and mechanics?

3. Methodology

30 student essays from prior research were used in this investigation. Students in the prior study were asked to produce an argumentative essay using the prompt below (Figure 1). The Louvain Corpus of Native English Essays was used to choose the topic (LOCNESS). It is a collection of articles written in native English. Because LOCNESS was made up of argumentative writings in general, the argumentative essay and subject below were picked above other essay genres.

Figure 1. Writing Prompt

Write a well-developed argumentative essay on the topic below:

Technology and Imagination

Some people say that in our modern world, dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. Discuss your opinion about this statement.

The essays were analysed by the researcher and by Criterion in terms of grammar and mechanics. Among 144 essays, 30 were selected according to their length. The essays were at least 340 words. The essays were uploaded to Turnitin, a plagiarism check website which uses Criterion as a helper tool, in two Word documents. 15 essays were copied in each Word document because Criterion does not analyse texts more than 65,000 characters. That is why two Word documents were uploaded to the website.

Criterion analyses the essays for Grammar mistakes (subject-verb agreement, fragment, verb error, garbled, word error, pronoun, and possessive), mechanics (punctuation mistakes). It gives spelling errors and article errors separately. In this study these two types of errors were put under mechanics and grammar respectively. The number of the errors identified by the Criterion and the researcher were tabulated and compared in details.

4. Findings

In order to answer the research question, the essays were analysed for the errors in grammar and mechanics, mainly classified by the Criterion. Table 1 below highlights the numbers of errors across 30 essays in grammar category by both the instructor and the Criterion. The numbers show that the Criterion is as accurate as the instructor in identifying the mistakes unlike the study of Dikli and Bleyle (2014) who found that Criterion had problems in identifying L2 learners' errors. It may be because of the software update of ETS. There is some mismatch between the instructor and the Criterion feedback, however, when the general overview is taken into consideration, it can be said that the Criterion did well in finding grammar errors of the L2 learners.

Table 2 highlights the numbers of errors in 30 essays in mechanics category. The Criterion and the instructor feedbacks are again close to each other, which again shows that the Criterion did well in identifying spelling and punctuation errors of the L2 learners.

Table 1. Grammar Errors

Error type	*Instructor Total*	*Criterion Total*
Sentence conjunction	82	74
Wrong Word	24	10
Subject-verb agreement	32	32
Article Error	123	140
Garbled	27	19
Preposition error	14	17
Total	**302**	**292**

Table 2. Mechanics

Error type	*Instructor Total*	*Criterion Total*
Spelling	156	158
Punctuation Mistakes	50	60
Total	**206**	**218**

Sample feedback page of the Criterion is shown in Figure 2. The learners can see their errors in details, and by clicking on the error they can see a prompt for revising it (Figure 3).

Figure 2. Sample Feedback Page

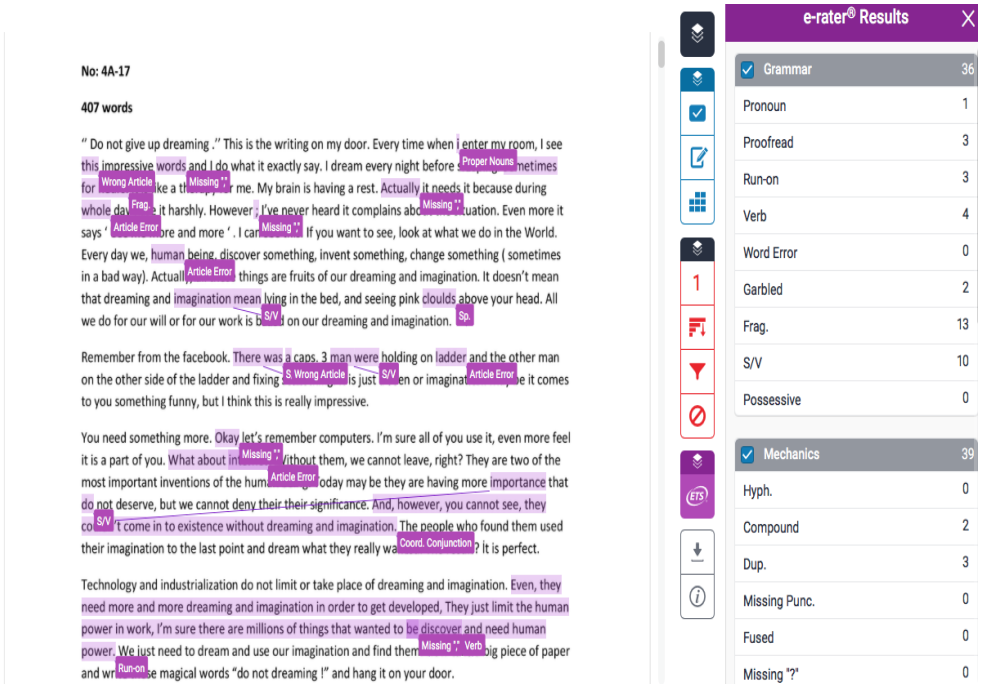
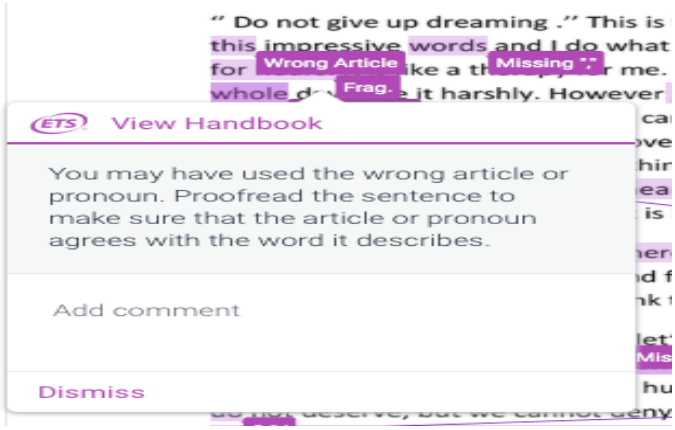


Figure 3. Sample Prompt



5. Discussion and Conclusion

In this study, the aim was to find out to what extent the feedback provided by an AES system, Criterion, coincides with the feedback given by a human rater in terms of grammar and mechanics category, mainly identified by ETS and used in Criterion. The human rater evaluated the essays according to the classifications identified in Criterion.

The results showed that Criterion was as accurate as the human rater in finding the grammar errors and mechanical errors in students' papers. This result is in conflict with the study of Dikli and Bleyle (2014), which has found that human raters are more accurate than Criterion in finding errors. This conflict can be attributed to the ETS continuous software update and since 2014, the Criterion may have been upgraded.

This result, however, in accordance with the claims that AES systems can be used to help teachers in writing classes by reducing the time spent on feedback (Attali, 2004; Chen and Cheng, 2008; Attali, Lewis and Steier, 2012; Choi and Lee, 2010; Wilson and Czik, 2016).

By looking at the findings it can be concluded that AES systems may be used in writing classes as a help tool in giving grammar feedback to students. They can also be used to promote autonomy of the learners and the learners can use AES systems to check their writings before submitting their final drafts. AES systems may provide learners multiple drafting and revising opportunity, which in turn may be useful in reducing their structural errors (Zhang, 2016).

The study can be expanded with interviews with the students and their perceptions may give some insights into using AES systems in EFL writing classes. A full-term classroom application of an AES system can also be a good study so as to observe the actual use of AES systems in revising and drafting. In addition, an experimental study can be conducted to find out the effect and benefits of AES systems.

Statement of Research and Publication Ethics

Ethics committee approval is not required for this article. While conducting the study, it was acted in accordance with research and publication ethics.

Author(s) Contributions to the Article

This study was conducted by one author, the researcher himself.

Funding Statement

The author received no specific funding for this work.

Conflicts of Interest

The author states that there is no conflict of interest.

References

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing, 9*(3), 227–257.

Attali, Y. (2004, April). Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education Annual Meeting, San Diego, CA.

Attali, Y.; Lewis, W.; Steier, M. (2012). Scoring with the computer: alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125-141.

Chen, C.F.; Cheng, W.Y. (2008). Beyond the design of automated writing evaluation: pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94-112.

Choi, J.; Lee, Y. (2010). The use of feedback in the ESL writing class integrating Automated Essay Scoring (AES). In D. Gibson, & B.Dodge (Eds.), Proceedings of society for information technology & teacher education international conference (pp. 3008–3012). Chesapeake, VA: AACE.

Cushing W.S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335-353.

Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal, 28*(1), 99-134.

Dikli, S.; Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback?. *Assessing writing, 22*, 1-17.

Fathman, A. (1990). *Teacher response to student writing: Focus on form versus content*. In B. Kroll (Ed.), Second Language Writing: Research insights for the classroom (pp. 178–190). Cambridge, UK: Cambridge University Press.

Ferris, D. (1995). Student reactions to teacher response in multiple draft composition 35 classrooms. *TESOL Quarterly, 29*(1), 33–50.

Ferris, D. (2011). Treatment of error in second language writing (2nd ed.). Ann Arbor, MI: University of Michigan Press.

Lee, Y.W.; Gentile, Claudia; Kantor, Robert (2008). Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater (RR 08-01). Princeton, NJ: Educational Testing Service (ETS).

Liao, H. C. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. *System, 62*, 77-92.

Long, R. (2013). A review of ETS's Criterion online writing program for student compositions. *The Language Teacher*, 37(3), 11-18.

Nichols, P. (2004). Evidence for the interpretation and use of scores from an Automated Essay Scorer. In Paper presented at the Annual Meeting of the American Educational Research Association (AERA) San Diego, CA.

Otoshi, J. (2005). An analysis of the use of Criterion in a writing classroom in Japan. *The JALT CALL Journal*, 1(1), 30-38.

Ranalli, J.; Link, S.; Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8-25.

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

Wang, J.; Brown, M. S.(2007). Automated Essay Scoring versus Human Scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, 6(2).

Wang, Y.J.; Shang, H.F.; Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, 26(3), 234-257.

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, 45, 769–774. doi:10.5054/tq.2011.272525

Warschauer, M.; Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 1–24.

Weigle, S. C. (2011). Validation of automated scores of TOEFL iBT® tasks against nontest indicators of writing ability. *ETS Research Report Series*, 2011(2).

Wilson, J.; Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94-109.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly*, 19(1), 79–97.