English-learner perceptions of Automatic Writing Evaluators

Michael H. Brown

Kanda Institute of Foreign Languages

Abstract

The capacity for computers to evaluate human writing has moved far beyond simple spelling and grammar checkers. Computers are now able to analyze written data in several ways and take many factors into account, such as lexical density, frequency of simple and compound sentences, or difficulty of vocabulary. Furthermore, applications and programs known as Automatic Writing Evaluators (AWEs) that score or grade writing promise to help students and other learners improve their writing while reducing workloads for teachers. This essay describes and synthesizes recent research literature concerning AWEs, especially the research literature looking at English-learner perceptions of, and attitudes toward, AWEs. The attitudes and perceptions of students and other learners affect the way that they interact with AWEs, as well as their expectations of what AWEs can be used for. The essay concludes with a discussion of implications for teachers based on the learners' perceptions of AWEs.

Introduction

L2 writing tasks, and even L1 writing tasks, are often scenes of frustration. They can be time-consuming and anxiety-inducing. Ideally, perhaps, writing instructors would be able to dedicate significant time to providing feedback and helping individual learners edit their writing. Unfortunately, the time available to instructors is usually not sufficient for careful, precise, and individualized feedback on writing.

Automated Writing Evaluators (AWEs) are software that analyze various features of writing and provide editing suggestions, corrective feedback, and/or a score. AWEs may be intended for use by L1 writers, L2 writers, or both. Regardless of the writer's language background, AWEs are marketed as tools that will help writers to write clearly, efficiently, and effectively. They promise to save time and improve writing accuracy. The following sections will discuss how AWEs have been viewed by L2-learning writers, how my own students reacted to two different AWEs, and a discussion of benefits, drawbacks, and suggestions for using AWEs with L2 writers.

Examples of AWEs

In the preceding section I described how AWEs are marketed. The choice of the term marketing is intentional, and not metaphorical. Most AWEs are commercial products. For example, *Criterion* is an AWE developed by Educational Testing Service (ETS), the organization which administers standardized English proficiency tests such as the TOEFL and the TOEIC. It is available to schools and educational organizations for a fee. *Criterion* is an online writing tool that provides diagnostic feedback and a score/rating for submitted essays. ETS describes the benefits of using *Criterion* in the following ways, and these points may be generalized to other AWEs, too:

- Students: Have more opportunities to practice writing at their own pace, get immediate feedback and revise essays based on the feedback.
- Teachers: Can decrease their workload and free up time to concentrate on the content of students' work and teach higher level writing skills.

- Administrators: Can make data-driven decisions and easily monitor district, school and classroom writing performance.

(ETS, n.d)

Some AWEs are made available for free. Completely free AWEs are more likely to have been developed for local contexts; and by, or in collaboration with, instructors. An example of such an AWE is *Marking Mate* (Marking Mate, n.d.), which is described as a free academic writing support tool for English learners in specifically East Asian contexts (Jordan & Snyder, 2012).

A third model is the freemium model. Freemium AWEs are commercial AWEs that are partially free. Or, more accurately, some features are free to use, but access to all features requires purchasing or a subscription to the service. Two examples of freemium AWEs are *Grammarly* (Grammarly, n.d.) and *Write and Improve* (Cambridge English, n.d.).

Student Attitudes

Research into student attitudes toward AWEs has been inconclusive, although some common themes have started to emerge. For example, some research has looked at how students use AWEs, or what they use AWEs to do. In this vein, Reis & Huijser (2016) found that some students primarily viewed the *Marking Mate* AWE as only for checking for errors and not as a tool that can help them learn how to fix mistakes. In the same study, the researchers described some students as having “a desire to do more” (p. 532) with the tool, but that the features available constrained this ability.

Student appraisals of the helpfulness of AWEs is another research theme. In one study, for instance, feedback provided by *Criterion* was deemed very helpful by students (Li et al, 2015). However, it should be noted that the same study also found it was particular kinds of feedback, mostly phrase and sentence level grammar and vocabulary checking that was found to be most helpful. Feedback on other essay elements, such as content and organization, were not perceived to be helpful to the same degree.

Another related theme which has emerged is that of usability, a notion often conceived of in terms of the Technology Acceptance Model, or TAM (Davis et al., 1989). The TAM posits that there are two key factors that determine the likelihood of a potential user accepting and actually using a new technology. These factors are a) perceived usefulness and b) perceived ease of use. It is theorized that these two factors influence individual attitudes toward a technology and the intention to use the technology, which then influence whether the technology is actually used by an individual. Researchers in Australia, for example, found generally high student ratings of usability for *Grammarly*, along with positive student-perceived impact on their writing, suggesting that, according to the TAM, these students would be likely to continue using *Grammarly* in the future (Cavaleri & Dianati, 2016).

The notion of usability, which can also be described as how successfully a potential user can use a particular tool (Krug, 2014), will be a central component of the rest of this essay. Particularly for students writing in an L2, perceived ease of use and tool design that takes L2 learner perspectives into account are key to successfully using

a tool. It has been suggested that, when considering whether to recommend an AWE to students, one should first ask oneself “How useable is this tool?” (Paiz, 2017)

My class

I carried out a small-scale, quasi-action research project in a low-intermediate reading and writing class at a two-year vocational college in Tokyo, Japan. The class had 17 students, 16 of whom were female (one male), and the L1 of every student was Japanese. As part of a standardardized curriculum, the students were required to write three assessed (graded) pieces: two essays and a summary of a magazine article. A theoretical pre-supposition of such assignments is that although the students may “not have complete control over English vocabulary and grammar, their language proficiency is presumed to be strong enough so that the focus of assessment can be writing *per se*, not language proficiency as demonstrated through writing” (Weigle, 2013, p. 37).

For one essay, no particular instruction or recommendation regarding AWEs was given. For the magazine article summary, *Write and Improve* was introduced and used by the class for writing and revising drafts, before final submission for formal assessment. For the second essay, an opinion essay, *Grammarly* was introduced and used by the class for writing and revising drafts, before final submission for formal assessment. Both of these AWEs are available under the freemium model, and in both cases the free versions were utilized.

Write and Improve uses machine learning algorithms to ‘learn’ features of written English characteristic of non-native English writers, mark problematic grammar or

lexical items, and assign an estimated Common European Framework for Reference (CEFR) score to a piece of writing moments after submission. Students can take the feedback, revise their writing, and resubmit within the *Write and Submit* interface as many times as they like. I suggested students should aim for a particular CEFR score estimate before formally submitting their essays to me. Such goal-oriented revision processes can be motivating for students (Grimes & Warschauer, 2010). Furthermore, a major factor in selecting *Write and Improve* is that is designed specifically for learners of English as an L2 (Harrison, 2017).

As this was only a quasi-research project, and I was in fact a participant in the project, not only an observer, the primary data collected was based on in-class discussions about the AWE, the kinds of issues or problems that students reported, positive remarks, and my own interpretation of how successfully students were able to use *Write and Improve*. This data was collected in field notes; sometimes the notes were contemporaneous and recorded in a notebook during class, while at other times the notes were recorded after class as part of reflective practice.

Based on my interpretation of students' comments, discussions, and use of *Write and Improve*, the decision was made to experiment with a different AWE, *Grammarly*, for the second essay. The selection of *Grammarly* was based on the fact that prior research had indicated it was perceived as easy to use (Li et al., 2015) and it had, similar to *Write and Improve*, a free version available under a freemium model. In contrast to *Write and Improve*, however, *Grammarly* is a general purpose AWE with no special design intended to benefit non-native English writers.

The data collection methods for *Grammarly* were similar to those used for *Write and Improve*. That is, the data was based on in-class discussions, questions or comments from students, and my own perceptions of how well they were able to use the AWE. Recording methods were similar as well: contemporaneous in-class notes or notes recorded as part of reflective practice after class.

Findings, Interpretation, and Discussion

I had expected *Write and Improve* to be beneficial for student writing, and, in many ways, it was. Students were able to set goals, such as a particular CEFL level estimate, and they would submit and revise until they had met their goal. Then they would submit their draft. For me, this was very beneficial and time and effort-saving as many spelling and grammar errors were fixed before students submitted their drafts. However, many students had trouble understanding the feedback given by *Write and Improve*. While errors or problematic language was marked, students often could not figure out what they should do to fix the errors. Suggestions for fixes were not offered by the AWE for many errors, resulting in the students asking me for help. So, while after drafts were submitted there was a lighter workload for me, I was actually quite busy helping students make sense of the AWE's feedback prior to draft submission.

Grammarly, on the other hand, seemed to generate fewer issues for students. Feedback tended to include suggestions for fixing errors, and students asked fewer questions about how to make sense of the feedback. Students also indicated a general sense that *Grammarly* was easier to use. The reasons why it was perceived

as easier to use were not always clear, although a few ideas were repeated by several students.

First, the *Grammarly* writing interface is less cluttered than the interface for *Write and Improve*. Whereas the latter has a text writing box, an error box, a graph of CEFR estimates box, color and shape coded error markings, and a system of roll-over boxes for getting details about error markings, *Grammarly* has a simpler layout. *Grammarly* has a large writing space on the left side of the computer screen and items marked as errors are noted, and fixes suggested, in the margin to the right of the writing space. Students commented they were sometimes confused by the layout of *Write and Improve*, but that *Grammarly* was not as confusing.

Another point was made that to get feedback on *Write and Improve*, students had to finish their essay before they could receive feedback. This resulted at times in a heavily marked essay with a lot of feedback. The amount of feedback could feel overwhelming, or be discouraging. Feedback on *Grammarly* was more immediate, as potential errors are marked during the writing process. This is similar to how basic spelling and grammar checkers on word processing software work, but at a more advanced level. Several students remarked that they liked being able to fix errors as they arose, rather than trying to deal with many errors at once. In-class discussions noted that both AWEs were seen to be useful, but *Write and Improve* seemed more complicated to the students.

Regarding both AWEs, students commented that sometimes they did not know whether the items marked as errors were actually errors. I should note sometimes I

indicated to students that both AWEs occasionally make mistakes, and if the students were not sure whether some AWE feedback was accurate, they should check with me. There were several instances where students asked me to help them interpret some feedback, and I subsequently judged the feedback to be wrong, unnecessary, or overzealous (for example, *Grammarly* would mark nearly every instance of passive voice as questionable). Similarly, both AWEs at times failed to mark clear errors; so these errors were not caught until after students submitted their drafts to me. Especially for the *Grammarly* algorithms, it is conceivable that some errors are missed because “ESL errors are notoriously difficult to categorize” (Weigle, 2013, p. 44). This concern is less pronounced, though not absent, for *Write and Improve* because it is trained solely on English learner writing. As a result, students weren’t sure that they could always trust the AWEs. They indicated a preference for the AWE to supplement, but not replace, instructor feedback.

Although my students seemed to prefer *Grammarly* more than *Write and Improve*, I must note that the order in which these tools were introduced may have influenced how they were perceived. One study found that Japanese college students became more efficient with, or more comfortable using, *Criterion* as they became more familiar with it (Tsuda, 2014). Likewise, it is possible that *Grammarly* was perceived by my students as easier to use than *Write and Improve* because they were becoming more accustomed to using an AWE in general.

Finally, the preference for *Grammarly* seemed to stem from design features of its interface, not from a sense that it was more accurate or had better feedback. In other words, how feedback was presented seems to be what mattered a great deal.

Indeed, it is easy to imagine *Write and Improve* becoming more powerful, useful, and popular as its machine learning algorithms are exposed to more and more data from English learner writing. But if its interface continues to be perceived as relatively difficult, then it may be an underused tool in the future.

Conclusion

In this essay I have tried to explain student attitudes toward and perceptions of AWEs, and some of the factors influencing those attitudes. AWEs are becoming more popular and more prevalent in language education. They are also becoming more sophisticated. They are not seen as wholly accurate in their assessments. Nonetheless, they are perceived as beneficial by students when they are used for phrase and sentence level checks, but not all the time, and not for more complex aspects of writing such as cohesion or content appropriateness. AWEs are seen as editing support tools that can supplement instructor assessment, but cannot replace it. Perceived ease of use is a key factor driving student preferences regarding competing AWEs.

References

Cambridge English. (n.d.) *Write and Improve* [website]. https://writeandimprove.com

Cavaleri, M. & Dianati, S. (2016). You want me to check your grammar again? The perceived usefulness of an online grammar checker as perceived by students. *Journal of Academic Language and Learning*, (*10*)1, A223-A226.

Davis, F.D., Bagozzi, R.P., & Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, (*35*)8, 982-1003.

ETS. (n.d.). *Criterion* [website]. https://www.ets.org/criterion

Grammarly. (n.d.) *Grammarly* [website]. https://www.grammarly.com

Grimes, D. & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, (*8*)6. http://www.jtla.org

Harrison, L. (2017). Developing an ELT product based on machine learning: Write & Improve. In: *ELTjam* [website]. https://eltjam.com

Jordan, E. & Snyder, A. (2012). Marking Mate - A free web-based academic writing feedback tool for East Asian learners of English [presentation]. *JALTCALL 2012 Conference*. Kobe, Japan.

Krug, S. (2014). *Don't make me think, revisited: A common sense approach to web and mobile usability* (3rd ed). San Francisco: New Riders.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, (*27*), 1-18.

Marking Mate. (n.d.). *Marking Mate* [website]. http://writingtools.xjtlu.edu.cn:8080/mm/marking mate.html

Paiz, J.M. (2017). Taking Another Look at Student Perceptions of Automated Writing Evaluators (AWEs): The Case of a Sino-British Joint Venture. In: *ELT Research Bites* [website]. https://www.eltresearchbites.com

Reis, C. & Huijser, H. (2016). Correcting tool or learning tool? Student perceptions of an online essay writing support tool at Xi'an Jiaotong-Liverpool University. In: *Show Me The Learning*. S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds), pp. 529-533. Adelaide: ASCILITE.

Tsuda, N. (2014). Implementing Criterion (Automated Writing Evaluation) in Japanese College EFL Classes. *Language and Culture: The Journal of the Institute for Language and Culture*, (*18*), 25-45.

Weigle, S. (2013). English as a Second Language Writing and Automated Essay Evaluation. In: *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. M.D. Shermis & J. Burstein (Eds), pp. 36-54. Routledge.