Guest Editorial

Automated Writing Evaluation in Language Teaching: Theory, Development, and Application

Volker Hegelheimer¹, Ahmet Dursun¹ and Zhi Li²

The first automated writing evaluation (AWE) software for assessment purposes dates back to the 1960s when Page Ellis developed Project Essay Grade (PEG). Rapid advances in the fields of artificial intelligence and Natural Language Processing (NLP) in the last few decades have led to the development of more powerful commercial scoring engines, such as e-rater, developed by ETS, IntelliMetric by Vantage Learning, and Intelligent Essay Assessor (IEA) by Pearson Education. Recent years have seen the application of these scoring engines expanding to language learning and teaching contexts (i.e., Criterion, WriteToLearn, MyAccess!).

Opinions on the utility of AWE tools and their potential effects on educational practices vary, as shown by two frequently-cited books on AWE: Ericsson and Haswell (2006) and Shermis and Burstein (2013). While many AWE tools are promising with reportedly high reliability, the use of AWE in writing classrooms has provoked heated discussion and strong opposition, as articulated in the 2004 position statement of the Conference on College Composition and Communication (CCCC). Most recently, Chapelle, Cotos, and Lee (2015) have proposed a new, argument-based approach to evaluating the use of AWE-based assessments. In general, more studies are needed to evaluate the use of AWE tools in classrooms and beyond. This Special Issue intends to extend the discussion by bringing together a variety of studies related to

Affiliations

¹Iowa State University
²Paragon Testing Enterprises Inc.
email: volkerh@iastate.edu (corresponding author)
AWE in the context of Computer-Assisted Language Learning (CALL). The issue includes conceptual and empirical research on AWE tool development, AWE tool classroom implementation, and resulting pedagogical implications. Therefore, it is an informative resource for AWE designers and developers, applied-linguistics researchers, and language teachers and practitioners alike. With an emphasis on AWE development for classroom use and its implementation, this issue is a good complement to existing books on AWE. In total, six research articles that include a theoretical discussion and/or empirical research on the promise, challenges, and issues related to the development, implementation, or evaluation of AWE tools are presented as follows.

The first article by Ma and Slater titled ‘Connecting Criterion scores and classroom grading contexts: A systemic functional linguistic model for teaching and assessing causal language’ represents an effort at interpreting AWE scores from a theoretical perspective. In this study, Ma and Slater apply the model of developmental path of causal language in exploring the relationship between Criterion scores and ESL students’ use of causal language. Their findings reveal that Criterion scores reflect ESL students’ levels in the development of causal language. As its title suggests, this study exemplifies a theory-based interpretation of Criterion scores and provides recommendations regarding the potential use of Criterion in classroom contexts.

The next paper titled ‘How can writing tasks be characterized in a way serving pedagogical goals and automatic analysis needs?’ by Quixal and Meurers presents an argument of why pedagogical and computational requirements need to inform and complement each other for a successful implementation of automatically evaluated language learning tasks. More specifically, the paper proposes a task-based language teaching-based methodology and extends to include an explicit characterization of the NLP requirements to design, implement, and evaluate two foreign language writing tasks. To empirically validate the task design, Quixal and Meurers piloted these tasks with 21 students in two different universities in Europe. The results suggest that the design-based specifications for the interpretation, evaluation and annotation of learner responses were equally useful both from the pedagogical and language learning perspective as well as the computational perspective. This paper brings a unique approach to the design of automated language tasks in the sense that it links second language acquisition, foreign language teaching and learning as well as NLP.

What follows is the article ‘Automated error detection for developing grammar proficiency of ESL learners’, by Feng, Saricaoglu, and Chukharev-Hudilainen who introduce a newly developed open-source AWE system, CyWrite, and focus on its performance in the detection of four error types (quantifiers, subject-verb agreement, articles, and run-on sentences). Their
comparative study with reference to ETS’s *Criterion* showed that CyWrite demonstrated higher F-scores, a combined measure of detection accuracy, which uses both precision and recall. Through analyzing the cases of inaccurate error detection, Feng et al. also discuss potential approaches to improving the performance of CyWrite. Their article remains accessible despite the necessary discussion of more technical aspects of the product. Such papers, though challenging to write, are likely to become standard practice as authors seek to provide a more balanced treatment of the technological and pedagogical elements of the program. More importantly, this new AWE system appears promising in providing accurate detection of common grammatical errors made by English language learners.

Liu and Kunnan’s paper ‘Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn’ evaluates the scoring ability and the precision of feedback of a commercial AWE application, WriteToLearn, as compared to trained raters in an English-as-a-foreign language (EFL) context. In this study, Liu and Kunnan used many-facet Rasch measurement (MFRM) to investigate the relationship between automated scores and human rated scores as well as the accuracy of the feedback provided by WriteToLearn to those of trained human raters. In terms of scoring, the study found the AWE program to be consistent but quite severe in comparison to the human raters. It also failed to score a few cases. In terms of feedback accuracy, it had an overall precision rate of 49%. The study also found that WriteToLearn was not very effective at identifying certain types of errors including articles, prepositions, word choice, and expression, which are common among the study’s Chinese L1 target population.

This article is followed by Cotos and Pendar’s paper ‘Discourse classification into rhetorical functions for AWE feedback’ which presents a novel approach to the development of the automated analysis engine for the Research Writing Tutor (RWT). RWT is an AWE program designed to provide genre and discipline-specific feedback on the functional units of research articles in learners’ academic disciplines. More specifically, they report on the approach they employed to develop the automated analysis and feedback engine for research article introductions based on communicative moves and functional steps (Swales, 1990). To do this, they combined work in genre analysis and machine learning, which rely on linguistic cues to reveal the rhetorical functions of texts. Cotos and Pendar’s work displays an innovative use of NLP in the implementation of automated genre analysis. According to the authors, this approach also ‘paves the road for future AWE endeavors and forms of automated feedback that could facilitate construction of functional meaning in writing’.
The last paper is Burstein, Elliot, and Molloy’s study titled ‘Informing automated writing evaluation using the lens of genre: Two studies’ which aims to guide the development of future AWE systems through expanding the writing construct beyond academic essay genre. In the first study, they collected data through a national survey of K-12 and college educators, college students who completed internships in the workplace, as well as non-educators in the workplace. Then, they report on the results of a site survey study at a private four-year postsecondary institution and faculty focus group discussions targeting course writing requirements. Burstein et al. call for a fresh look into what AWE system developers should consider when designing next-generation AWE tools.

By bringing together a variety of researchers and practitioners who have employed qualitative, quantitative, or mixed-method methodologies in researching different AWE tools across different contexts and genres, this Special Issue aims to raise the awareness of researchers and practitioners regarding the use of AWE tools as part of classroom instruction putting the learners at the center of the design and development. This issue is timely as it illustrates the need for new-generation AWE systems. The papers in this issue offer both valuable guidance for implementation and practical suggestions for needed research on the use of AWE tools as potential language learning technologies. In a wider sense, we hope that this Special Issue helps to de-mystify how AWE tools are developed for pedagogical purposes and also to shed light on understanding best practices for teaching L2 writing with AWE tools.

With AWE tools becoming more powerful and readily available, researchers and practitioners should not ignore the potential benefits afforded by these tools. It is also our hope that this Special Issue will stimulate further discussion about: (1) the critical roles of linguistic theories in the development of AWE tools; (2) effective approaches to integrating AWE tools in a writing curriculum; and (3) the potential interaction between AWE tools and user characteristics such as learner motivation, self-regulated learning strategies, teachers’ beliefs and agency.

References


