International Journal of TESOL Studies (2021)
Vol. 3 (2) 71-87 https://doi.org/10.46451/ijts.2021.06.06

Article

Automated Writing Evaluation: The Accuracy of Grammarly's Feedback on Form

Marina Dodigovic*
University of La Rioja, Spain

Artak Tovmasyan
American University of Armenia, Armenia

Abstract
Automated Writing Evaluation (AWE) is gaining more and more presence, with Grammarly being a prominent example of this kind of software, which provides automated feedback on essay writing. While the Grammarly advertisement is full of praise regarding its ability to give meaningful feedback at various levels, including grammar or form, plagiarism, vocabulary and style, there has been little third party research to support these claims. Since Grammarly's vocabulary feedback accuracy has already been examined to some extent (Dodigovic, Mlynarski & Wei, 2016), this study was designed to evaluate the accuracy of Grammarly's grammar checker feedback. It did so by investigating the reports generated by Grammarly as a response to a small corpus of EFL writing, compiled from essays written by Armenian undergraduate students. The results were compared with those of human raters. In addition to Grammarly's error detection power, the study also investigated the kind of remediation that Grammarly suggested. The comparisons revealed that on detected errors of form, Grammarly mostly provides accurate feedback, with occasional inconsistencies. It was also found to provide mainly adequate remediation for the correctly detected errors of form. However, it also left a number of errors undetected, which led to the omission of remediation. The study concluded that, despite its considerable usefulness, when using Grammarly, English learners should avoid relying solely on its feedback.

Keywords
AWE, Grammar Checkers, Grammarly, ICALL, AI

1 Introduction

Automated Writing Evaluation (AWE), of which Grammarly is an example, has become widespread in English as an Additional Language (EAL) classrooms. Currently, one of the models of AWE use includes teachers as part of the evaluation. Within this model, the teacher will need to check the feedback

***Corresponding Author**
Address: Department of Modern Filologies, C/ San José de Calasanz, 33, 26004 Logroño, La Rioja, Spain
Email: mdodigov@gmail.com

provided by such software and compose the final feedback (Williamson, 2008). However, an emerging model eliminates the participation of the teacher and provides instant feedback to the learners through the software. While some criticize this model because of its allegedly inaccurate feedback (Mitkov, 2005), there are very few studies that evaluate the performance of such software. This might facilitate potentially inadequate use of the software.

The present study evaluates the accuracy of feedback provided on form through the second model, where teachers are eliminated from the automated feedback provision. What is meant by form is mostly grammatical structure of individual utterances. While Grammarly may be present in some students' daily life, it is not clear to what extent they are able to critically evaluate the feedback received, especially in the absence of teacher mediation. This study is likely to assist students, teachers and educational administrators in making executive decisions regarding the purchase and use of this software.

2 Literature Review

2.1 Errors in general

Errors are the deviations from the standard, accepted rules of a language committed by the learners of a second language (Ellis, 1992). They are generally viewed as one of the most important aspects in language learning by many EFL teachers (Polio, 2012). Errors may point to a variety of causes, including negative transfer from L1 (Zobl, 1980), ineffective teaching methodology or the inadequacies of learning strategies (Dodigovic, 2005; Huang, 2002).

Theoretically and practically, errors have been viewed and addressed in different ways. The behaviorist approach views errors as the inability of learners to change the old habits and create new ones. These occur in the learner output when the learners fail to adequately respond to a stimulus in L2 (Walsh & Lado, 1957). The behaviorists place emphasis on the attempt to avoid errors at all cost, lest the wrong verbal behaviour be internalised. However, the view changes in the 1960s (Corder, 1967), leading to the idea of regularities in learner errors which represent a language of its own, i.e. interlanguage. The cognitivists approach error correction in a number of ways, since cognitivism views errors as signs that learners are testing hypotheses (Ortega, 2014). Thus James (1998) argues that error correction as a curative way of instruction is more effective than preventive instruction based on error avoidance. Not everyone though favours explicit correction. Interactionists for example value indirect correction through repetition and recasts (Sarem & Shirzadi, 2014). It seems that Grammarly combines a variety of the above approaches.

2.2 Artificial intelligence and natural language processing

Although not much is communicated about the makeup of Grammarly by its developers, given its capacity to detect and correct learner errors, one must assume that the software owes this capacity to Artificial Intelligence (AI). AI includes Natural Language Processing (NLP), which arises from Computational Linguistics (Chapelle & Sauro, 2017).

NLP is used to create intelligent tutoring systems (ITS) (Aleven, Roll, McLaren, & Koedinger, 2016; Amaral & Meurers, 2011; Crossley, Varner, Roscoe, & McNamara, 2013; Ferreira, Moore, & Mellish, 2007; Lee, Lee, Noh, Lee, & Lee, 2011). The application of ITS, automated writing evaluation (AWE), automated essay scoring (AES) etc. is also known as Intelligent Computer Assisted Language Learning (ICALL) (Schulze, 2008).

The studies of language tutors that incorporate NLP have taken a twofold approach: (1) focus on communication and (2) focus on form. However, some provide feedback on both dimensions (Holland

& Kaplan, 1995). For instance, one such aid, called the Intelligent Tutor has the ability to identify and correct the learner errors, using multiple sources of data in order to acquire knowledge needed to diagnose the errors (Dodigovic, 2007).

2.3 Grammar checkers and error analysis

Grammar checkers, such as those found in word processors and Grammarly, are deemed to use Error Analysis (EA). This is a branch of Applied Linguistics, initiated to satisfy the needs of teachers and researchers that had to analyze big chunks of input for errors, their diagnosis and correction. This field of study quickly started changing the understanding of Second Language Acquisition (SLA) during the 1970s (Heift & Schulze, 2007). While EA provides many benefits to understanding and supporting learning, it was also criticized for its limitations, including its inability to address all errors and the lack of focus on what the learner can produce (Dagneaux, Denness, & Granger, 1998).

An article to the topic (Wilson, Olinghouse, & Andrada, 2014) argues that a computer providing feedback on writing of learners without the intervention of the teacher has no significant effect. It investigated 4-8 graders' writings for a statewide assessment to determine whether the revisions based on automated feedback would show improvement. While during the first wave of testing the gains in student writing suggested associations with gains in automated writing feedback (AWF) under certain circumstances, the overall improvement growth declined over time. Another study (Debuse, Lawley, & Shibl, 2009) pinpoints the educators' perspectives on using automated writing evaluation software as positive. While the research revolves around the educators' and learners' perspectives of such software packages and their effects on second language acquisition along with AWE serving as a reliable tool for language assessment, there is a noticeable dearth of articles that evaluate the accuracy of these applications and online services (Xi, 2010).

Amongst the few exceptions (e.g. Chukharev-Hudilainen & Saricaoglu, 2016; Cotos, 2011; El-Ebyary & Windeatt, 2010; Lai, 2010) is a study investigating the accuracy of Criterion, an advanced software package with similar functionality to Grammarly that is available to schools and universities (Hoang, 2019). It evaluates the feedback as a tool that has a positive impact on further drafts and writings of students that are exposed to the feedback of the application during early and developing stages of writing. Warschauer and Ware, in their 2006 study, compiled a collection of reports on software packages that were commercially available at that time. All three of these packages were compared in a table, where the products named "MY Access!" and "Holt Online Essay Scoring" provided holistic and component scoring, while "Criterion", which is developed by ETS, provided a single holistic score. The latter was said to provide a "wide range of individualized feedback", while the other two packages observed provided "limited individualized feedback" (Warschauer & Ware, 2006, p. 3).

A study conducted in China used Grammarly and the Vocabulary Size Test (VST) to detect plagiarism (Dodigovic, Mlynarski, & Wei, 2016). The study used Grammarly to analyze the data and divide the output from the software into corresponding categories to track frequencies. Although this study only assessed Grammarly's output on lexical errors it is methodologically similar to the present study, with the only difference being the focus on form in the present study.

Another way of assessing the value of AWE software is by comparing the feedback of software packages to those of human raters (Dikli & Bleyle, 2014; Shintani & Aubrey, 2016). Finally comparing the instructor's feedback with that of an AES software Criterion (Li, Link, Ma, Yang & Hegelheimer, 2014; Pahl & Kenny, 2009), based on the perception of students revealed questionable efficacy of the software in ESL contexts.

A study investigating the use of AWE as a pedagogical tool (Sevcikova, 2017) used Grammarly and Marking Mate, along with 10 teachers, to evaluate four texts. When comparing the feedback reports, only one text demonstrated major discrepancies between AWE and human rater reports. This study compared

the final report grades provided by human raters and AWE. However, this could result in inaccuracy of the comparison, as the reports might display the same grade for each essay, for very different reasons. For example, human raters and AWE could identify different errors in the text, but award the same grade.

A similar study (Dembsey, 2017) looked at the accuracy of correction (comment cards) provided by Grammarly. While methodologically not too different from the rest of the literature in the field (Dale, 2016; Schraudner, 2007), this article makes strong claims about Grammarly's comment cards that occur when the application detects errors in text. The author recommends that writing centers make their own decisions when it comes to compensating for the limitations of their services. Instead of offering students Grammarly, Dembsey (2017) is in favour of training human consultants to give more effective feedback. While this paper did marginally comment on the accuracy of Grammarly's feedback, this was far from being in its focus. In this situation, it seems purposeful to examine the accuracy of Grammarly's feedback directly, using its correction of actual student writing.

3 Methodology

3.1 Research design

This study employs a mixed-method approach, including a corpus of student writing and its computational and manual processing in order to answer the following research questions:

1. How accurate is the detection of errors of form by Grammarly?
2. How accurate are the corrections of errors of form provided by Grammarly?

3.2.1 The corpus

In order to evaluate the performance of Grammarly when it comes to checking text for errors of form and providing correction, a 36000-word corpus was compiled using 56 essays (Fig. 1). These essays were written by first year undergraduate students of the American University of Armenia (AUA), most in their late teens, all of whom studied English as their major. The first language of the students was Armenian. In compliance with the institutional code, the students gave consent for their anonymized essays to be used in research.

3.2.2 Grammarly

Grammarly is an application developed with a preliminary aim of enhancing English language writing in both general and academic settings through providing feedback on the text input, supporting a variety of platforms, such as Microsoft Windows and Mac OS with a desktop software interface as well as various web browsers using its web interface. Grammarly can also be installed as a web browser extension, where it checks texts within different websites. Recently Grammarly added yet another application to its features, this time, on smartphones. Users can now add Grammarly as a virtual keyboard that brings up grammatically correct "next-word" suggestions based on what the users have already typed in. The version of Grammarly's desktop app used in this study is 1.5.36.

The online and desktop platforms have been through many updates and improvements in detecting errors in language, currently operationalizing up to 400 grammatical structures, an option to enhance vocabulary and selection of a specific style of writing to follow while providing feedback on the writing of its customers. This number was mentioned in the literature to be drastically lower "250 error types" only a few years before (Dodigovic, Mlynarski, & Wei ., 2016, p. 227).

3.3 Data collection

To be able to gather data on the effectiveness of grammatical error detection by Grammarly, the corpus described above was first read for errors. A native speaker of English, 3 graduate students majoring in TEFL and a professor of English read the papers for errors. The sets of identified errors were reviewed by the researcher in order to find some common ground. After consolidating the individual error lists into one, the common list was compared with the feedback gained from Grammarly. The results of the comparison were verified by the rest of the team.

The same corpus was divided into sections and uploaded into Grammarly in order to generate automated feedback on these errors. The corpus had to be divided into sections as Grammarly only supports a limited amount of text per each check.

Grammarly Premium was used to generate the feedback on the corpus. The premium package supports many additional services, such as plagiarism detection and vocabulary enhancement. While those two features were not used, because this study is investigating form errors rather than semantics and style ones, another premium feature, which allows the selection of a document type to check the text accordingly, was enabled. For the subcorpora used in this study, the document type of "Essay" was used from the "Academic" subcategory. The other features of Grammarly, such as plagiarism detection and vocabulary enhancement were disabled. In order for Grammarly to produce form error feedback only, the contextual spelling, punctuation and style error checkers were also disabled (see Figure 1).

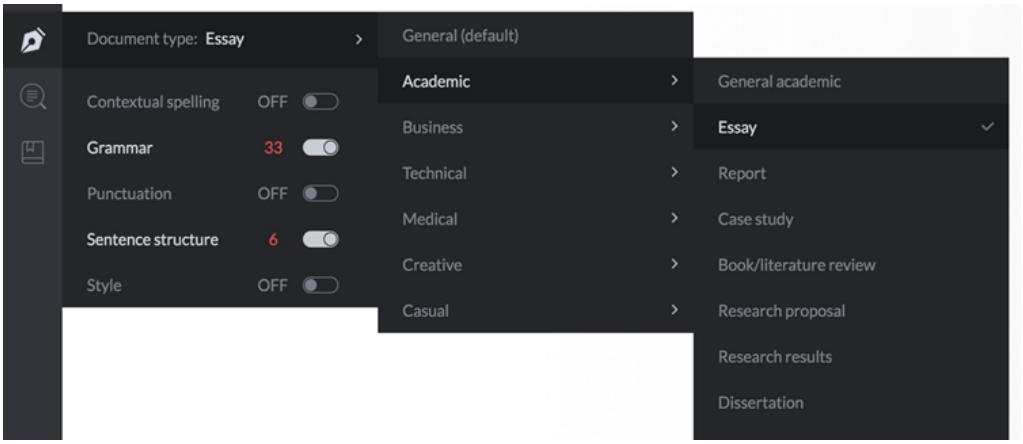

Figure 1. The choice of checkers and documents styles used to generate feedback

Grammarly creates reports marking the errors with various colors, each identifying the weight of the error. For example, red highlights in Grammarly represent critical issues, such as wrong preposition use, while the yellow highlights represent advanced issues, such as redundant determiners in text. However, the vocabulary Grammarly uses to address the typology of form errors discovered in text is also simplified for the purpose of being more user-friendly. This is not the case with the human raters that read the same corpus for errors. Thus, in order to successfully contrast the output of Grammarly with that of the raters, the typology for both needed to be the same, which was arranged for the purpose of this study.

While collecting the feedback on form errors provided by human raters and the report on the same errors provided by Grammarly, two quite different typologies of form errors emerged. The taxonomy used to describe the human raters' notes on each error in corpus was adapted from a study on form errors (Wu & Garza, 2014). Grammarly, on the other hand, uses its own terms. Grammarly's error taxonomy is not publicly available, although some parts of it become visible when Grammarly creates reports, which is the exact method that was employed to reconstruct Grammarly's tacit taxonomy used to report on form

errors in this study (see Figure 2). The advantage of the Wu & Garza (2014) taxonomy is that it seems to rest on the premises quite similar to those that must have guided the authors of Grammarly, namely that language is multilayered, including lexis, grammar, semantics and other aspects.

Most of the typology that emerged in Grammarly's reports is easily identifiable when it comes to matching it to the taxonomy used by the human raters. However, some of the error types used by Grammarly can be found to be connected to multiple categories of the converted taxonomy. For example, Grammarly's "incomplete sentences" category includes errors of verb/subject omission and fragment errors, while modifier and quantifier misuse can apply to verb, sentence structure and singular and plural categories in the human rater's error taxonomy (see Figure 2).

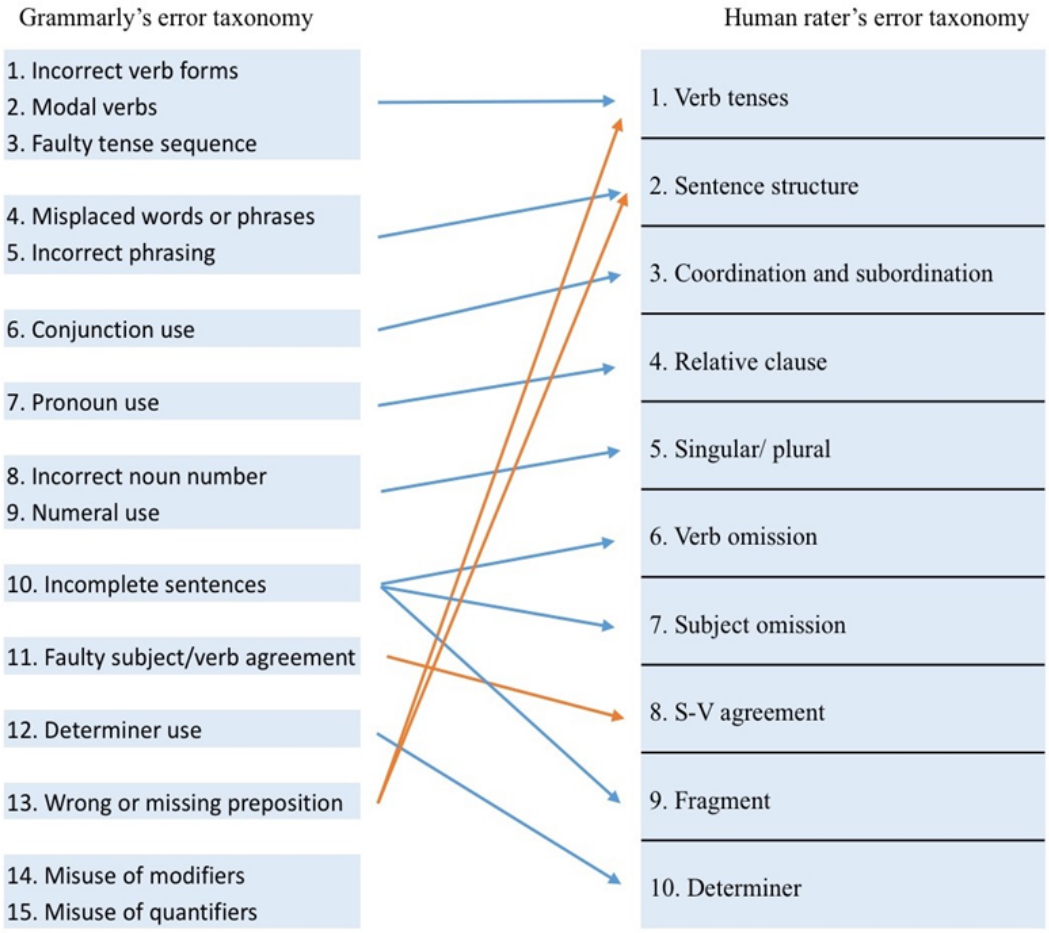

Figure 2. Form error taxonomy used by Grammarly versus human raters including connections between the two taxonomies of form errors

It should be noted that in Figure 2, some lines are red to make it easier to see their connection to the other taxonomy. The last two classes of error, 14 and 15, can connect to multiple error types in the other classification, depending on the context. Their connection was intentionally discarded in order to avoid confusion.

3.4 Data analysis

After the documents with error markings from human raters and the reports from Grammarly were collected, an important step was to convert all qualitative data into quantitative (see Table 1).

Table 1

The Description of Data Sources, Analysis and Quantitative Results

Source	Data type (QUAL)	Procedure	Result (QUAN)
1.Human raters	Documents containing errors that were marked.	The documents were first compared, then the errors found in each were counted and classified according to the human rater's taxonomy.	A dataset of errors that demonstrates the frequency of each form error.
2.Grammarly report	A pdf report of the text, with errors marked.	The reports were scanned to count the error frequencies. The errors were counted according to the taxonomy Grammarly used.	A dataset of errors that demonstrates the frequency of each form error.
3.Grammarly UI	A pdf report of the text with solutions/hints and the digital user interface.	The interface was scanned for correction suggestions and hints. A new taxonomy emerged when counting the correction instances.	A dataset of corrective messages provided by Grammarly.

The errors were counted in two different ways:

1. Using the emerging taxonomy (see Figure 2).
2. Using the pre-determined error detection evaluation flow chart (see Figure 3).

In this flow chart, detected errors represent errors flagged by the software for the correct reason, meaning when compared to those made by the human raters, these errors were found to be correctly highlighted and commented on. On the other hand, missed errors represent the errors that were undetected by Grammarly but detected by the human raters.

The general category is for cases of somewhat correct error detection, where the software correctly detects the error but fails to pinpoint a certain phrase or sentence in the paragraph. The false positive (otherwise known as "over-flagging") classification is for cases when the software flags a section of the text that contains no error.

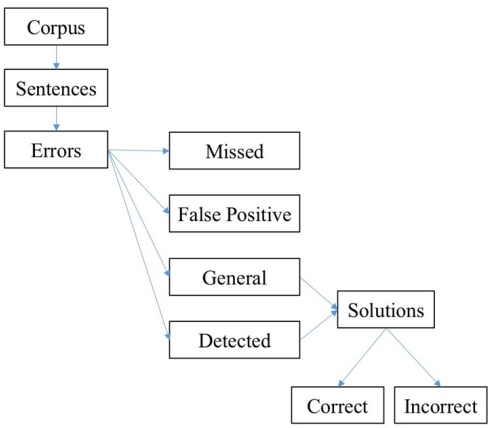

Figure 3. The emergent typology of Grammarly's feedback accuracy

Note: There are no corrections or hints for the missed errors. The corrections and hints for the errors flagged as a false positive were discarded.

As for the "Solutions" category, there are only two subcategories to choose from. The "correct", that demonstrates a correct detection of the problem and, more significantly for this subcategory, the correct hint or solution for the user to find the solution to fix the error. The other category is "incorrect", where the correction or hint for the detected error does not match the underlying problem.

This procedure was repeated with all 7 subcorpora that were analyzed by Grammarly and descriptive statistics were used to report the data. The most important section of the data is undoubtedly the percentage of matches found during the comparison of the first two sources. The quantitative data below reports the rate of matches found in the error detection of the two previously mentioned sources, this way, providing information on the accuracy of form error detection by Grammarly. The percentage coverage of errors was calculated by dividing the number of Grammarly errors that matched those of the Human Raters divided by the total number of errors detected by Human Raters.

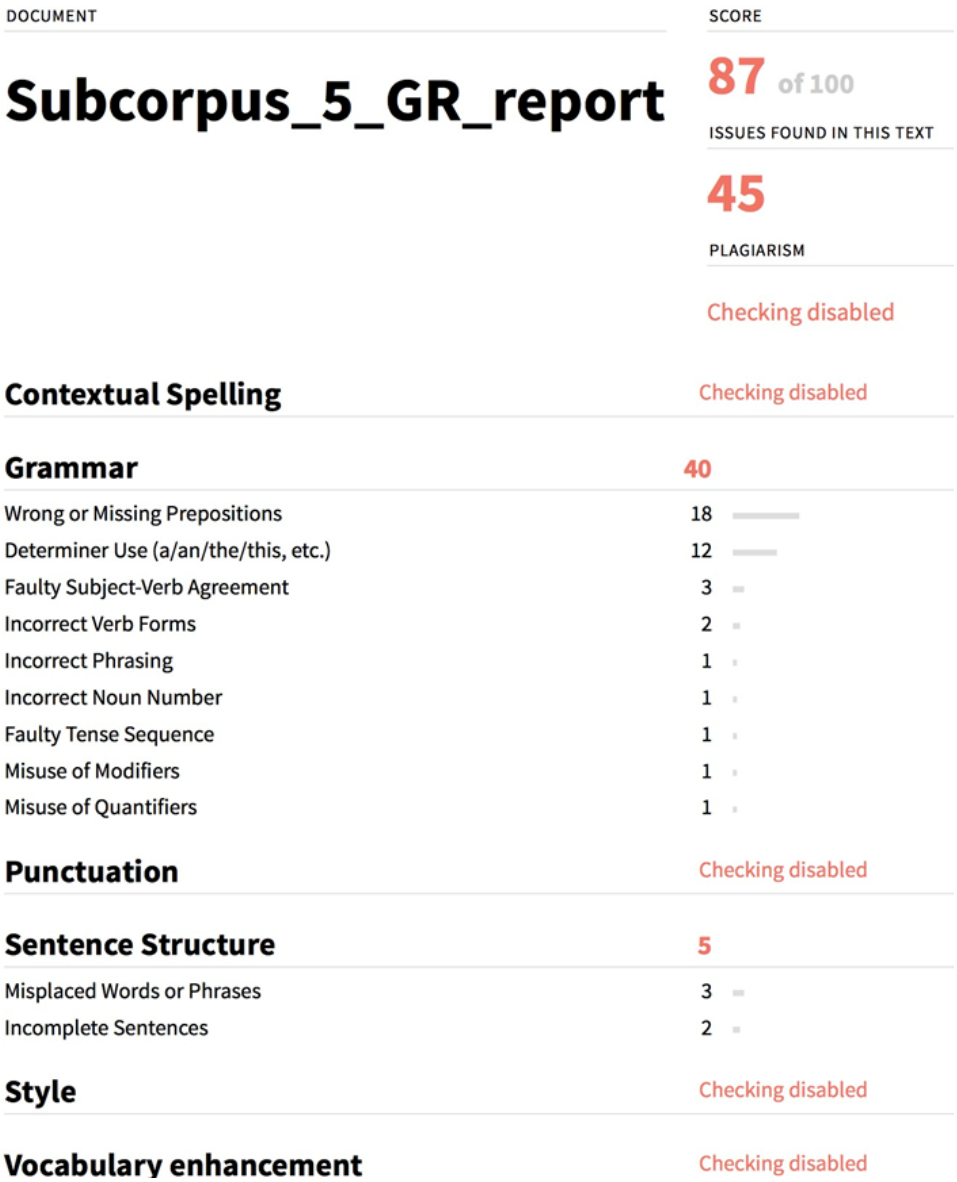

Figure 4. Grammarly report example

Figure 4 shows the first page of the .pdf report generated by Grammarly for one of the subcorpora. The reports also contain the complete subcorpus with all errors highlighted. However, the reports do not include the correction or hint cards.

4 Results

4.1 Grammarly reports on form

The following Grammarly reports were derived from the 7 sub-corpora and consolidated manually. (see Table 2).

Table 2

Reports Generated by Grammarly on the Seven Subcorpora

		Subcorpora							
	Error type	1	2	3	4	5	6	7	Total
1	Determiner use	14	7	8	18	12	9	5	73
2	Wrong or missing preposition	10	5	7	8	13	7	8	58
3	Faulty subject/verb agreement	3	3	0	5	4	2	0	17
4	Modal verbs	1	0	0	0	0	0	0	1
5	Incorrect noun number	1	0	0	1	1	0	0	3
6	Misplaced words or phrases	5	2	2	6	3	3	1	22
7	Incomplete sentences	1	1	2	1	2	0	0	7
8	Incorrect verb forms	4	2	1	1	2	1	0	11
9	Incorrect phrasing	0	1	0	0	1	0	0	2
10	Misuse of modifiers	0	0	1	0	1	0	0	2
11	Misuse of quantifiers	0	0	1	0	0	0	0	1
12	Numeral use	0	0	1	0	0	0	0	1
13	Pronoun use	0	0	0	1	0	0	0	1
14	Conjunction use	0	0	0	0	1	0	0	1
15	Faulty tense sequence	0	1	0	0	0	1	0	2
	Total for subcorpora	39	22	23	41	41	23	14	203

When investigating Table 2, the first thing that stands out is the total count of various form error classes detected by Grammarly. For example, the most frequent error types are determiner use, wrong or missed prepositions and misplaced words or phrases.

The taxonomy used in the consolidated report represents a conversion to the common scale applied also by human raters. On average, the classes of form errors that had the highest frequency occurred 51 (SD = 26.21) times. The error classes with the lowest frequency were modal verb, misuse of quantifiers and pronoun and conjunction use errors, followed by misuse of modifiers, incorrect phrasing and faulty sequence of tenses across subcorpora produced (M = 1.2, SD = 0.44).

4.2 The accuracy analysis

In order to address the first question of this study, Table 3 was compiled. It represents errors found by Grammarly per subcorpus and their evaluation by the research team. Grammarly was able to detect 203 form errors.

On average, Grammarly flagged 29 (SD = 11.06) form errors per subcorpus. Of those, an average of 20 (M = 19.57, SD = 5.82) for each subcorpus were found to be correctly identified. These are the

statistics of the "Detected" category in the error accuracy flow chart. "Detected", in this scenario, means that the error was accurately highlighted and flagged as an error and subsequently classified as correctly identified by the research team. The accuracy of correction given is reported below (see Table 5).

"General" is the class for errors detected, but not correctly. For example, Grammarly highlights a section in a sentence and detects an error. However, the error is in a different section of that sentence. The average error detection per subcorpus that was classified as "General" is lower that the "Detected" (M = 2.5, SD = 1.3), while the total count of such instances (18) is considerably lower as well.

The third column in Table 4 represents the total number of cases that were classified as "detected" or "general". These classes represent Grammarly's accurately detected errors. A total of 155 out of 203 error flags by Grammarly were accurate (76.35% of 203), not including the missed errors. The percentage column demonstrates the accuracy of the total number of form errors detected by Grammarly. The last row of the table is calculated horizontally, while the rest of the table is calculated vertically.

Table 3

The Classification of Form Error Flagging by Grammarly

		Errors							
SC	Total *f*	Detected	General	Total correct	%	False positive	Missed	Total incorrect	%
1	39(7)	23	4	27	58.6	12	7	19	41.3
2	22(2)	18	1	19	79.1	3	2	5	20.8
3	23(2)	19	2	21	84	2	2	4	16
4	41(7)	26	4	30	62.5	11	7	18	37.5
5	41(6)	25	4	29	61.7	12	6	18	38.2
6	23(3)	17	2	19	73	4	3	7	26.9
7	14(3)	9	1	10	58.8	4	3	7	41.1
SUM	203(30)	137	18	155	66%	48	30	78	33%

Note: SC = subcorpora, 203 = Grammarly-detected only, 233 = missed (30) combined

The next two columns, false positives and missed are very informative as well. False positive (otherwise known as over-flagging) and missed (otherwise known as false negative) demonstrate Grammarly's inability to flag errors accurately, or to detect them at all. It is also important to understand which of the two are more detrimental to the user. However, these two classes of undetected errors on average are less common than the positive detection of errors by Grammarly (33% of all errors present in the corpus). The table shows that Grammarly was approximately 66% accurate in its flagging, or, approximately 33% inaccurate, not including the 30 errors that Grammarly missed.

The range of the percentages for accurate (from 56.2% in the first subcorpus to 84% in the third subcorpus) and inaccurate (from 16% in the third subcorpus to 41.1% in the seventh subcorpus) error flagging is within a 30% span, which is not negligible.

4.3 Analysis through contrast

The comparison between the human rater and Grammarly reports reveals some differences in the frequency of errors per category in the converted and combined taxonomy of errors. Table 5 demonstrates the frequency of errors found in the corpora analyzed by the human raters vs. that of Grammarly. It is essential to mention, that the frequency of errors found in the texts by different human raters was

averaged out to get a more accurate frequency of errors. As stated in the data analysis section, the errors detected by both human raters and Grammarly were further manually evaluated by the researcher. The errors that were found by one human rater only were discarded. There were altogether 32 cases of disagreement among human raters.

Table 4

The Comparison of Frequencies of Error Detection by Human Raters and Grammarly

	Error Categories	Human raters (*f*)	Grammarly (*f*)	Match (%)
1	Verb tenses	44	39	88.6
2	Sentence structure	47	35	74.4
3	Coordination and subordination	13	1	7.6
4	Relative clause	9	1	11.1
5	Singular/ plural	11	4	36.3
6	Verb omission	17	3	17.6
7	Subject omission	15	2	13.3
8	S-V agreement	24	17	70.8
9	Fragment	8	2	25
10	Determiner	74	51	68.9

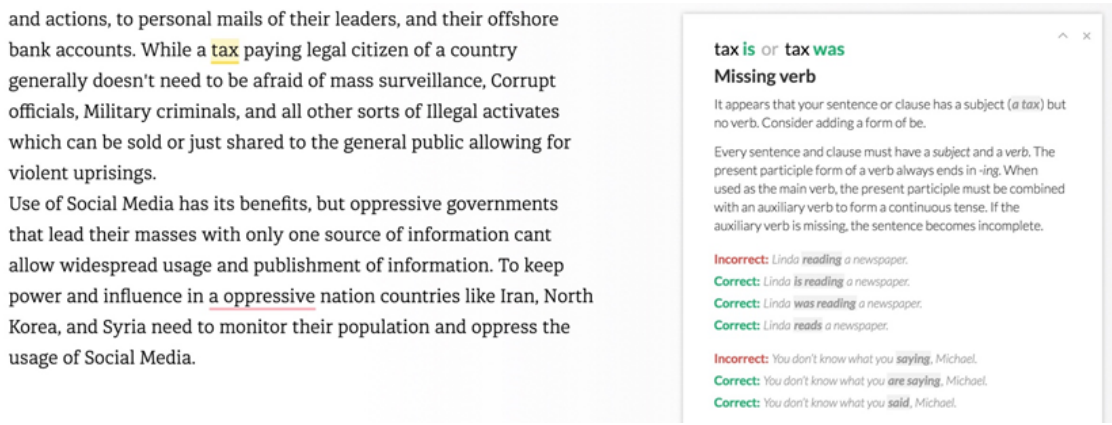

Figure 5. Example of an incorrect detection of an error by Grammarly

Note. This error detection case was classified as "over-flagging".

Table 5 represents the level of convergence between the human raters and Grammarly. It applies to the entire corpus, based on the adopted taxonomy, described in chapters two and three. The difference between the two provides a clue as to the effectiveness of Grammarly at the detection of various error types.

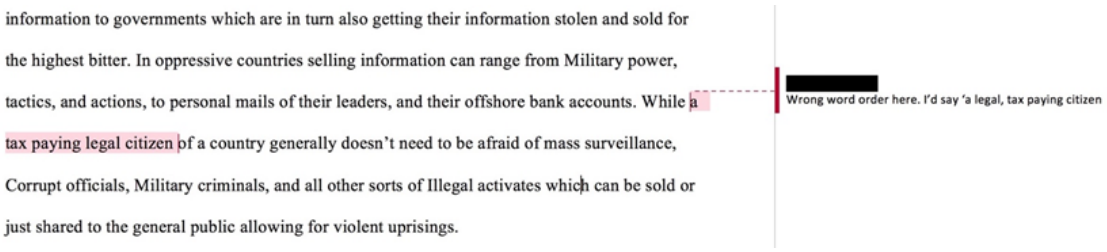

Figure 6. Example of human rater feedback discrepancy

Figure 6 above exemplifies a case which was identified by one, but not all human raters. This case, alongside another 31 cases only identified by one marker, was not included in the total counts.

4.4 Correction

The purpose of this study was also to determine to what extent the corrective feedback on form provided by Grammarly is accurate in the given setting and sample.

Table 5
The Classification of Grammarly Solutions to Form Errors

	Errors (*f*)		Solutions (*f*)	
SC	Detected	Correct	Correct	Incorrect
1	39	27	19	8
2	22	19	13	6
3	23	21	14	7
4	41	30	21	9
5	41	29	21	8
6	23	19	11	8
7	14	10	6	4

Table 5 demonstrates that Grammarly is to some extent effective when providing correction for the errors that it correctly detects. It is essential to mention that in order to count the error correction instances, first, the missed and over-flagged errors were excluded as it is expected that if Grammarly incorrectly detects an error, the correction of the error will be incorrect as well.

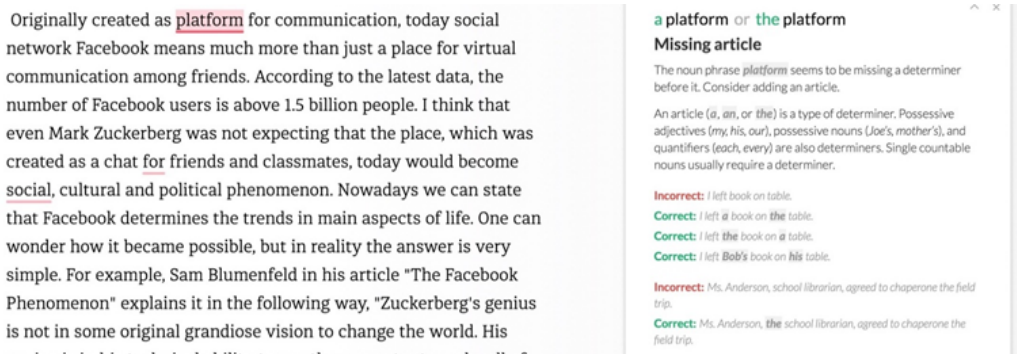

Figure 7. Example of an accurate error detection and correction by Grammarly.

5 Discussion of Findings

The objectives of this study were to assess the accuracy of firstly form error detection and secondly their correction by Grammarly. The findings demonstrate that Grammarly has approximately one third of inaccuracies while preforming either task, in addition to errors it ignores. This differs to some extent from both previously reviewed papers that deal specifically with the evaluation of Grammarly's performance, not just in outcomes, but also in methodology. Unlike the present study, the first paper by Sevcikova (2017) finds very little discrepancy between Grammarly's reports and those produced by human raters, while the second paper (Dembsey, 2017) gives a definite preference to human raters.

Dembsey (2017) also introduced a list of positive and negative reviews compiled from various sources, including blogs and websites. Some of the positive reviews mentioned in the list match the findings of this study. An example of this are Grammarly's categories of errors and the clear explanations provided for the correctly detected errors. On the other hand, its confidence in Grammarly's ability to handle large texts was not consistent with this study. While Grammarly could handle large essays, it was not able to analyze a relatively small corpus (cca 36000 words)."

Dembsey (2017) lends a voice, amongst other things, to those reviewers who find fault with Grammarly. The largest number of such reviewers reported false positives (over-flagged errors) and false negatives (missed errors) (Dembsey, 2017). This is similar to the findings of the present study. Other reviewers report unclear or technical explanations in the corrective feedback on each detected error. While it was not within the scope of this study to evaluate the ease of taking advantage of the correction or the comprehension of Grammarly's error explanations, it could still be noted that Grammarly occasionally did present too technical and at times unclear explanations. The false-positives in Grammarly's feedback on form could be connected to the hypercorrection as an error cause (Touchie, 1986). Touchie describes this cause of errors as the repetitive correction of the same error that causes errors elsewhere in structures that would otherwise be correct.

Another, more general, article argues that automated feedback on learner writing has no significant effect when the teacher is not involved (Wilson, Olinghouse, & Andrada, 2014). The present study seems to support such a conclusion, given the relatively high inaccuracy rate of Grammarly. Thus, it would seem that the involvement of a teacher in the process of feedback provision and review could eliminate the chance of hypercorrection through over-flagging, thus making the feedback more useful to the learner.

In the reviewed literature, the perspectives on error explanations and their effects become more positive over time, which suggests that Grammarly tries to improve the solution/hint provision service. However, while Grammarly is trying to make the terminology user-friendly and easily comprehensible for EAL learners and others, they are also running the risk of oversimplifying the error categories and classes. This was the main reason a taxonomy conversion system was introduced in the present study to manually evaluate the Grammarly error reports more effectively.

While reporting the convergence level (in Table 5) that answers the first research question of the study, some other patterns were found that are not within the scope of this study to investigate. For example, the average convergence level for the highly common error types of verb tenses, determiner and sentence structure is 77.3% (SD = 10.1) while the less frequent error type convergence percentages, for coordination/subordination, relative clause and fragment errors, is 14.5% (SD = 9.2). This is a concern, since patterns of this kind are frequent in ICALL accuracy evaluation studies (e.g. Darus & Ching (2009)) as they obviously abound in EAL learner output. However, investigating this anomaly is not the purpose of this study, although it raises interesting questions for further research.

6 Limitations

One limitation with the current study is that Grammarly's feature set developed over the duration of the study and is likely to continue to develop. The detection of errors and the suitability of the correction feedback could have improved.

Another limitation is the human raters' partial disagreement on error identification. The reports that the human raters generate are intertwined with a number of factors that are different for each human rater. The human raters that evaluated the texts for this study have different backgrounds. Thus, it is actually surprising that no more than 32 cases of disagreement occurred.

The last limitation is the relative uniformity of the essays included in the corpus. The essays came from Armenian undergraduate students, all majoring in English, which could be an influencing factor for the results of this study.

7 Conclusions

This paper investigated Grammarly in order to find out how accurate it is in its analysis of and feedback on form, i.e. grammar errors in text. Of the errors detected by Grammarly, approximately 66% were found to be accurate error identifications, while the rest were false positives. Moreover, a number of errors were ignored by Grammarly, approximately 7% of the total number of identified errors. These numbers are to some extent discouraging with regard to EAL students as immediate users, as it is not clear how such students would be able to take on board the correct feedback and ignore the erroneous suggestions. In addition, Grammarly's inability to identify some errors might lead such students to believe that the structures used are correct, which is a concern.

The findings of this study therefore suggest that Grammarly could be used by EAL students as an additional tool of form-related feedback but under the close supervision of a teacher. Currently, in light of the findings of this study, the usage of Grammarly as a stand-alone product may not be optimal.

While Grammarly, as shown in this study, has aspects in its feedback that some learners of English could potentially benefit from, it also has issues that could potentially have a negative effect on the learning. This aspect of Grammarly, within a broader context of AWE, lacks research. Thus, a recommendation here is to study the positive and negative effects of both accurate and inaccurate feedback by Grammarly on EAL learners. It would also be useful to investigate how learners interpret the feedback and what they do in the immediate follow-up. This type of study could potentially illuminate the underlying problems of AWE systems and help researchers and developers from multiple disciplines uncover the full potential that AWE systems may have for language learners.

References

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, but only so Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education, 26*(1), 205–223. https://doi.org/10.1007/s40593-015-0089-1

Chapelle, C. A., & Sauro, S. (2017). *The Handbook of Technology and Second Language Teaching and Learning. Wiley.* Retrieved from https://books.google.am/books?id=Fg8rDwAAQBAJ

Chukharev-Hudilainen, E., & Saricaoglu, A. (2016). Causal discourse analyzer: improving automated feedback on academic ESL writing. *Computer Assisted Language Learning, 29*(3), 494–516. https://doi.org/10.1080/09588221.2014.991795

Corder, S. P. (1967). The significance of learner's errors. *IRAL-International Review of Applied Linguistics in Language Teaching, 5*(1–4), 161–170.

Cotos, E. (2011). Potential of Automated Writing Evaluation Feedback. *CALICO Journal, 28*(2), 420–459. https://doi.org/10.11139/cj.28.2.420-459

Crossley, S. A., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7926 LNAI, pp. 269–278). https://doi.org/10.1007/978-3-642-39112-5-28

Dagneaux, E., Denness, S., & Granger, S. (1998). *Computer-aided error analysis. System, 26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

Dale, R. (2016). Checking in on grammar checking. *Natural Language Engineering, 22*(3), 491–495. https://doi.org/10.1017/S1351324916000061

Darus, S., & Ching, K. H. (2009). Common errors in written English essays of form one Chinese

students: A case study. *European Journal of Social Sciences, 10*(2), 242–253.

Debuse, J. C. W., Lawley, M., & Shibl, R. (2009). AJET - Educators' perceptions of automated feedback systems. *Australasian Journal of Educational Technology, 24*(4), 1–12.

Dembsey, J. M. (2017). Closing the Grammarly Gaps: A Study of Claims and Feedback from an Online Grammar Program. *The Writing Center Journal, 36*(1), 63–100. Retrieved from http://www.jstor.org/stable/44252638

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. https://doi.org/10.1016/j.asw.2014.03.006

Dodigovic, M. (2005). *Artificial Intelligence in Second Language Learning:Raising Error Awareness*. Multilingual Matters. Retrieved from http://www.multilingual-matters.com/display.asp?K=9781853598296

Dodigovic, M. (2007). Artificial Intelligence and Second Language Learning: An Efficient Approach to Error Remediation. *Language Awareness, 16*(2), 99–113. https://doi.org/10.2167/la416.0

Dodigovic, M., Mlynarski, J., & Wei, R. (2016). Vocabulary Size Assessment as a Predictor of Plagiarism. *Current Issues in Language Evaluation, Assessment and Testing: Research and Practice,* 222.

El-Ebyary, K., & Windeatt, S. (2010). The Impact of Computer-Based Feedback on Students' Written Work. *International Journal of English Studies, 10*(2), 121–142. Retrieved from http://ezproxy.usherbrooke.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ936915&site=ehost-live

Ellis, R. (1992). *Second language acquisition & language pedagogy*. Clevedon: Multilingual Matters.

Ferreira, A., Moore, J. D., & Mellish, C. (2007). A Study of Feedback Strategies in Foreign Language Classrooms and Tutorials with Implications for Intelligent Computer-Assisted Language Learning Systems. *International Journal of Artificial Intelligence in Education, 17*, 389–422.

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in Computer-assisted Language Learning : Parsers and Pedagogues Routledge Studies in Computer- assisted Language Learning* (C. CAROL, Ed.). Taylor & Francis Routledge.

Hoang, T. L. G. (2019). Examining Automated Corrective Feedback in EFL Writing Classrooms: A Case Study of Criterion . PhD Thesis, University of Melbourne.

Holland, V. M., & Kaplan, J. D. (1995). Natural language processing techniques in computer-assisted language learning: Status and instructional issues. *Instructional Science, 23*(5–6), 351–380. https://doi.org/10.1007/BF00896878

Huang, J. (2002). Error analysis in English teaching: A review of studies, 19–34.

Jacobs, G., & Rodgers, C. (1999). Treacherous Allies: Foreign Language Grammar Checkers. *CALICO Journal, 16*(4), 509–530.

James, C. (1998). *Errors in Language Learning and Use*. *Exploring Error Analysis.* London: Longman.

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41*(3), 432–454. https://doi.org/10.1111/j.1467-8535.2009.00959.x

Lee, S., Lee, J., Noh, H., Lee, K., & Lee, G. G. (2011). Grammatical error simulation for computer-assisted language learning. *Knowledge-Based Systems, 24*(6), 868–876. https://doi.org/10.1016/j.knosys.2011.03.008

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*(1), 66–78. https://doi.org/10.1016/

j.system.2014.02.007

Mitkov, R. (2005). *The Oxford Handbook of Computational Linguistics*. OUP Oxford. Retrieved from https://books.google.am/books?id=yl6AnaKtVAkC

Ortega, L. (2014). *Understanding second language acquisition*. Routledge.

Pahl, C., & Kenny, C. (2009). Interactive correction and recommendation for computer language learning and training. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 854–866. https://doi.org/10.1109/TKDE.2008.144

Polio, C. (2012). The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing, 21*(4), 375–389. https://doi.org/10.1016/j.jslw.2012.09.004

Sarem, S. N., & Shirzadi, Y. (2014). A Critical Review of the Interactionist Approach to Second Language Acquisition. *Journal of Applied Linguistics and Language Research, 1*(1), 62–74.

Schraudner, M. (2007). The Online Teacher' s Assistant : Using Automated Correction Programs to Supplement Learning and Lesson Planning, 124–136.

Schulze, M. (2008). AI in CALL--Artificially Inflated or Almost Imminent? *Calico Journal, 25*(3), 510–527. Retrieved from http://journals.sfu.ca/CALICO/index.php/calico/article/view/793

Sevcikova, B. L. (2017). Automated Essay Scoring as a Pedagogical Tool, *2017*(5), 1–19.

Shintani, N., & Aubrey, S. (2016). The Effectiveness of Synchronous and Asynchronous Written Corrective Feedback on Grammatical Accuracy in a Computer-Mediated Environment. *Modern Language Journal, 100*(1), 296–319. https://doi.org/10.1111/modl.12317

Touchie, H. (1986). Second language learning errors: Their types, causes, and treatment. *JALT Journal, 8*(1), 75–80.

Walsh, D. D., & Lado, R. (1957). Linguistics across Cultures. Applied Linguistics for Language Teachers. *Hispania, 40*(3), 390. https://doi.org/10.2307/335383

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 2*(2), 1–24. https://doi.org/10.1191/1362168806lr190oa

Williamson, D. (2008). A Framework for Implementing Automated Scoring. Language, (February), 1–23.

automated and human scoring. *Journal of Educational Measurement, 36*(2), 158–184.

Wilson, J., Olinghouse, N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal, 12*(1), 93–118.

Wu, H., & Garza, E. V. (2014). Types and Attributes of English Writing Errors in the EFL Context—A Study of Error Analysis. *Journal of Language Teaching and Research, 5*(6), 1256–1262. https://doi.org/10.4304/jltr.5.6.1256-1262

Zobl, H. (1980). Developmental and Transfer Errors: Their Common Bases and (Possibly) Differential Effects on Subsequent Learning. *TESOL Quarterly, 14*(4), 469–479. https://doi.org/10.2307/3586235

Artak Tovmasyan is an English and Science teacher who received his MA from the American University of Armenia. He previously guided the development of an E-Learning platform and a bot for language assessment. He is interested in the automation of language learning and assessment, as well as innovation in Educational Technology.

Marina Dodigovic, PhD, is an honorary professor of English and TESOL at University of la Rioja. She has taught in MA TESOL programs internationally and conducted relevant research, which is documented in a number of books and peer reviewed journal articles to her name. Her most recent research is directed toward vocabulary learning.