

ARTICLE



## Automated written corrective feedback: Error-correction performance and timing of delivery

*Jim Ranalli, Iowa State University*

*Taichi Yamashita, University of Toledo*

### Abstract

*To the extent automated written corrective feedback (AWCF) tools such as Grammarly are based on sophisticated error-correction technologies, such as machine-learning techniques, they have the potential to find and correct more common L2 error types than simpler spelling and grammar checkers such as the one included in Microsoft Word (technically known as MS-NLP). Moreover, AWCF tools can deliver feedback synchronously, although not instantaneously, as often appears to be the case with MS-NLP. Cognitive theory and recent L2 research suggest that synchronous corrective feedback may aid L2 development, but also that error-flagging at suboptimal times could cause disfluencies in L2 students' writing processes. To contribute to the knowledge needed for appropriate application of this new genre of writing-support technology, we evaluated Grammarly's capacity to address common L2 problem areas, as well as issues with its feedback-delivery timing, using MS-NLP as a benchmark. Grammarly was found to flag 10 times as many common L2 error types as MS-NLP in the same corpus of student texts while also displaying an average 17.5-second delay in feedback delivery, exceeding the distraction-potential threshold defined for the L2 student writers in our sample. Implications for the use of AWCF tools in L2 settings are discussed.*

**Keywords:** *Syntax/Grammar, Writing, Human-Computer Interaction*

**Language(s) Learned in This Study:** *English*

**APA Citation:** Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology*, 26(1), 1–25. <http://hdl.handle.net/10125/73465>

### Introduction

At the same time that much of L2 writing in English is taking place across a variety of digital spaces, powerful and sophisticated error-correction tools have become available across these spaces. Learners have come to expect that some form of automated help, at least with spelling, will be provided in email programs, learning management systems, and mobile device keyboards. The reach of sophisticated error-correction methods addressing not only spelling but complex areas of grammar has increased with the advent of tools such as Grammarly, which integrates into web browsers, office productivity software, mobile devices, and even Google Docs. Importantly, these tools can operate synchronously, providing feedback as writers write.

Recent published work on automated help for writing has focused on so-called AWE (automated writing evaluation) tools such as Criterion (e.g., Lavolette et al., 2015; Ranalli et al., 2017) and MY Access! (Chen & Cheng, 2008; Dikli, 2010). These tools deliver feedback asynchronously, allow access only through standalone web interfaces, and attempt to address both grammatical errors and higher-level issues (e.g., organization) with mixed results. This has left under-investigated another type of tool for automated feedback on writing—one which delivers feedback synchronously, is accessed in a convenient diversity of ways, and harnesses state-of-the-art technologies in focusing on lower-level concerns, including error types common to L2 writers—that we refer to as the automated written corrective feedback (AWCF) tool.

Viewed from cognitive theoretical perspectives on L2 writing and L2 learning, this new genre presents both opportunities and risks. Synchronous corrective feedback (CF) provided by teachers or text-chat interlocutors has been found to lead to increased gains in grammatical accuracy compared to CF that is delayed (Arroyo & Yilmaz, 2018; Shintani & Aubrey, 2016). AWCF tools could provide a more practicable and frequent source of such feedback. Yet synchronous AWCF may also constitute a potential source of distraction for writers if the timing of its delivery is misaligned with the cognitive processes involved in text production.

To contribute to the knowledge needed for appropriate applications of this new technological genre, we undertook evaluations of Grammarly from both system-centric and user-centric perspectives (Chodorow et al., 2010). For the former, we assessed Grammarly's error-correction performance vis-à-vis the unique needs of L2 writers, and for the latter, the consequences of Grammarly's enhanced error-correction capabilities on the timing of its feedback.

### Automated Corrective Feedback for L2 Student Writers

Researchers in the field of grammatical error correction (GEC) distinguish between methods and systems aimed at L1 users versus those designed for L2 users of a language, or, more recently both L1 and L2 users (Napoles et al., 2019) because the different groups are characterized by different common error types. L1 student writers' most common errors after spelling errors include lack of a comma after an introductory element and vague pronoun reference (Connors & Lunsford, 1988). By contrast, an error-annotated version of the Cambridge Learner Corpus (CLC), which represents a wide range of L1s and English proficiency levels, shows errors involving word choice, prepositions, and determiners to have the highest proportions after spelling errors (Leacock et al., 2014). [Table 1](#) lists the 10 highest-ranking errors in the CLC.

**Table 1**

*Top-ranked L2 Written Errors in the Cambridge Learners Corpus* (adapted from Leacock et al., 2014, p. 20)

Rank	Error Type	Example
1	Content Word Choice Error	<i>We need to deliver the merchandise on a daily <b>*base/basis</b>.</i>
2	Preposition Error	<i>Our society is developing <b>*in/at</b> high speed.</i>
3	Determiner Error	<i>We must try our best to avoid <b>*the/a</b> shortage of fresh water.</i>
4	Comma Error	<i>However, <b>*/</b>, I'll meet you later.</i>
5	Inflectional Morphology	<i>The women <b>*weared/wore</b> long dresses.</i>
6	Wrong Verb Tense	<i>I look forward to <b>*see/seeing</b> you.</i>
7	Derivational Morphology	<i>It has already been <b>*arrangement/arranged</b>.</i>
8	Pronoun	<i>I want to make <b>*me/myself</b> fit.</i>
9	Agreement Error	<i>I <b>*were/was</b> in my house.</i>
10	Run-on Sentence	<i>They deliver documents to them <b>they</b> provide fast service.</i>

The different frequencies of L1/L2 written errors necessitate different error-correction approaches. MS-NLP, which has been included with MS Word since 1997, is a system designed for detecting L1 errors. It comprises a parser and dictionary of morphological information small enough to be stored and operated on the user's local machine. According to Leacock et al. (2014), MS-NLP is based on a formal grammar called *augmented phrase structure*, which requires the expertise of trained linguists to create rules addressing, for example, subject-verb agreement and fragment errors. These rules are implemented when users initiate the spelling-grammar check in MS Word. Following an initial parse, the system creates a parse tree, which is

then converted to a semantic representation called a logical form. Critique rules are then applied to the logical form to check for rule violations, which, if found, initiate error-correction algorithms. While it remains “arguably the world’s most heavily used linguistic analysis system” (Leacock et al., 2014, p. 10), informal reports suggest the English version has been modified over successive releases so as to detect fewer and fewer error types (Kies, 2008), possibly in response to user complaints about false positives (Bishop, 2005).

Some of the most common error types in L2 English student writing, however, present challenges for GEC because the choice of the appropriate form rests not on syntactic rules but contextual dependencies. Gamon et al. (2009) describe how preposition and article errors require a great deal of contextual information to detect and correct. Choice of the correct preposition, for example, may depend on the noun that follows it, the verb that precedes it, the noun that precedes it, or some combination of these (Chodorow et al., 2010). Statistical or machine learning-based techniques have thus been used to tackle such errors because they obviate the need for intensive, manual effort and grammatical expertise in devising a large set of rules. Such systems avoid the need for exact matches by assigning higher probabilities to particular words that are more frequent and lower probabilities to those that are not through the use of statistical classifiers (e.g., maximum entropy<sup>1</sup> classifiers and support vector machines) in combination with different information types such as token-context, syntactic-context, part-of-speech (POS), and semantic information (Leacock et al., 2014).

Context also factors into solutions for addressing what is the most frequent error type in both L1 and L2 writing: misspellings. Spelling errors are a special concern in GEC not only for their frequency but because they can degrade the performance of NLP systems (Nagata et al., 2017). Most spell-checking systems have been based on research into L1 spelling errors (Heift & Schulze, 2007; Rimrott & Heift, 2005, 2008), and so they are less effective in addressing L2 spelling errors. This is because the underlying research shows L1 spelling errors to typically involve only the omission, inclusion, transposition, or substitution of a single letter (Rimrott & Heift, 2005). L2 misspellings tend to be more complex, originating with a number of different causes such as misapplication of morphological rules, L1 transfer, and lack of L2 morphological and phonological knowledge (Heift & Schulze, 2007). As such, they often involve edit distances greater than one—that is, two or more operations are needed for transformation into the corrected form—so they are difficult for L1-based spell checkers to handle. In response, researchers have developed systems that can address L2 spelling errors using contextual information (Flor & Futagi, 2012) or error-correction models derived from learner corpora (Nagata et al., 2017).

There is a general consensus that addressing the needs of L2 student writers for automated corrective feedback requires a hybrid approach, including hand-crafted rules for simpler errors, machine-learning techniques for those that are more context-dependent, and parser-based analyses for errors involving long-distance dependencies such as subject-verb agreement (Leacock et al., 2014). In addition, the need for pre- and post-processing routines (e.g., splitting text into sentences and applying exclusion rules to minimize false positives) means that analysis of L2 student text necessitates complex suites of procedures and, as a consequence, the computing power of remote servers rather than the user’s local CPU. An AWC tool like Grammarly is thus likely to display longer error-flagging delays than MS-NLP because of (a) the requirement that text be transmitted to and from remote servers for processing; (b) the large and complex array of computational processes involved; and (c) according to a Grammarly technical lead (M. Romanyshyn, personal communication, January 6, 2019), the requirement, in the case of some types of checks, for text to be bounded by sentence-final punctuation in order to facilitate parsing and part-of-speech tagging. Because our purpose was to evaluate both the error-detection performance of Grammarly as well as the effects of error-correction performance on the timing of feedback delivery, a brief review of GEC evaluation techniques is necessary.

## Evaluation of GEC Techniques and Systems

GEC evaluations have typically involved two measures: precision and recall, which originate in the field of information retrieval. Precision is concerned with false positives; that is, the proportion of flagged items that are not, in fact, errors. A precision rate of .73 for missing article errors, for example, would mean that

73% of a system's missing-article flaggings had been confirmed by human annotators to indeed be such errors. By contrast, recall involves false negatives; specifically, the proportion of actual errors that have been flagged. Recall of .35 for fragment errors would mean a system identified 35% of the total number of fragment errors in a corpus as attested by human annotators. There is a trade-off such that increasing precision leads to lower rates of recall and vice versa. GEC developers prioritize precision over recall because false positives are thought to be more detrimental to learners than false negatives, a position for which there is some empirical support (Nagata & Nakatani, 2010).

GEC researchers typically measure precision and recall with reference to a particular error-correction technique or method. Han et al. (2006), for example, reported precision and recall of .90 and .40 for an article-error detection approach based on a maximum entropy classifier and the use of token and POS contexts. Tetreault and Chodorow (2008) reported precision and recall of .84 and .19, respectively, for a preposition-error detection approach based on a maximum entropy classifier, the use of token and POS contexts, and heuristic rules. Developers may also specify a baseline of performance needed before a new method can be added to a tool. According to Quinlan et al., (2009), Criterion's developers require precision of .80 or above in testing. Actual performance in operational systems may vary considerably, however, since individual error-correction methods perform differently when combined with other methods and when applied to different types of text than those with which they were trained or tested. Classroom-based studies of Criterion have found precision rates as low as .51 for *Extra comma* errors (Ranalli et al., 2017) and .43 for *Wrong article* errors (Lavolette et al., 2015).

For what currently may be the two most widely used error-correction systems,<sup>2</sup> performance data are hard to come by; peer-reviewed evaluations for either Grammarly or MS-NLP could not be located for this study, with the exception of an analysis of MS Word's spell checker. As part of an investigation of spelling errors as predictors of L2 proficiency, Bestgen and Granger (2011) spell-checked a corpus of L2 English student writing using MS Word 2007 and found precision and recall rates of .80 and .82, respectively; a total correction rate was not reported. For comparison's sake, Rimrott and Heift (2008) found recall of .94 but correction of only .62 for the German-language version of MS-NLP analyzing the writing of L2 learners of German. Noting that most of the errors in their corpus were multiple-edit misspellings, Rimrott and Heift concluded that there is "a need to design spell checkers that specifically target L2 misspellings" (2008, p. 86).

### **The Timing of Written Corrective Feedback**

Cognitive theory provides a basis for consideration of feedback-timing issues as they relate to both L2 development and L2 writing. In a review connecting the pedagogical practice of focus on form to cognitive models, Doughty (2001) situated an important putative trigger of acquisitional processes—the cognitive comparison of student-produced and target-like forms—in working memory, postulating a 40-second window for optimal focus on form based on how long the forms to be compared can be held in short-term memory. Similarly, Long (2007) underscored the importance of providing CF within the time span in which learners are using linguistic forms to convey meaning. Recent empirical studies that adopted Doughty's view (Arroyo & Yilmaz, 2018) and Long's view (Shintani & Aubrey, 2016) both showed students who received immediate written CF outperforming those in delayed CF conditions on tests of accurate use of the target feature. However, Doughty's and Long's claims were made with reference to oral CF research, which has tended to operationalize CF timing as immediate if provided during a task and delayed if provided after a task. Therefore, this work has limited potential to inform thinking about the optimal timing of AWCF during writing so as to support L2 development.

With regard to the effects of feedback timing on writing processes, a key theme in cognitive models of writing is that working-memory resources are limited; conflicting demands on these resources can therefore prevent writers from accomplishing their goals effectively (Galbraith & Vedder, 2019). Two key processes for our purposes here are translation and transcription. Translation is the process whereby proposed ideas in non-linguistic form are converted into linguistic strings, and transcription is the process whereby linguistic strings are converted into text (Chenoweth & Hayes, 2001, 2003). Translation and transcription

are dependent on *verbal* working memory in particular—the subvocal, articulatory rehearsal process that can counteract the otherwise rapid decay of verbal information in short-term memory, and which writers experience as speaking to themselves while composing (Chenoweth & Hayes, 2003).

An interesting and important feature of the output of these processes is that it takes the form not of complete sentences but rather sentence parts, which are generated rapidly and usually terminated by a pause, so Chenoweth and Hayes (2001, 2003) coined the term *p-burst*, or pause burst. Research has shown that p-bursts are shorter when writers work in an L2 versus an L1, and when produced by students with less linguistic experience in the L2 (Chenoweth & Hayes, 2001). Other research demonstrates the potential for unrelated verbal information to interfere with translation and transcription. Chenoweth and Hayes (2003) showed that articulatory suppression (i.e., having participants repeat a syllable to themselves while composing sentences) slowed sentence production and reduced p-burst length by 40%. Ransdell et al., (2002) found that when participants had to write while attending to background speech, fluency, sentence length, and writing quality deteriorated. In Van Waes et al. (2010), participants were given a written sentence stem to complete using an auditory prompt while also identifying and correcting an error in the sentence stem. Given the choice of which task to perform first, the participants chose to complete the sentence before performing the correction in about 90% of cases, suggesting a felt need to avoid losing the proposed text in verbal working memory by suppressing revision.

Applying these ideas to the realm of synchronous AWCF, we propose that flaggings will represent potential distractions to the extent they interfere with working memory during translation and transcription. Thus, for synchronous AWCF to avoid such interference, it should address only errors that relate to the current contents of verbal working memory. We propose two indices for measuring the potential for synchronous AWCF to create distractions: error-flagging delay and p-burst duration. Error-flagging delay is the temporal difference between the commitment of an error and its flagging by an automated tool. P-burst duration is based on the p-burst, the unit of text production that represents the current contents of verbal working memory. In earlier studies (e.g., Chenoweth & Hayes, 2001), p-bursts were measured in number of words, but now, with keystroke logging software such as Inputlog (Leijten & Van Waes, 2013), p-burst duration can also be measured in milliseconds, facilitating comparison with error-feedback delay. If the delay is short enough that flagging addresses an error encompassed in the current p-burst, we assume the feedback can be incorporated into translation and transcription at comparatively little cost to attentional resources. On the other hand, if error-flagging delay exceeds p-burst duration, the flagging may draw the writers' attention to text produced in a previous p-burst, thereby placing conflicting demands on verbal working memory and thus constituting a potential distraction.

## The Present Investigation

We decided to investigate Grammarly's timing issues and its error-correction performance by comparing it to MS-NLP because this legacy GEC system provided useful benchmarks in two ways. First, as a system designed to correct L1 writing errors, it allowed a measure of how far automated error-correction techniques have come in addressing the unique needs of L2 learners. Second, it facilitated a comparison of Grammarly's feedback timing to another synchronous CF system whose operation is both familiar and generally perceived as unproblematic for users. We thus formulated the following research questions.

1. How do Grammarly and MS-NLP compare in terms of their capacity to address error types common to L2 student writers?
2. How do Grammarly and MS-NLP compare in terms of the timing of error flagging?

The research questions and the nature of the technologies involved required us to conduct two separate studies. These were part of a larger project that also investigated the effects of synchronous AWCF on revision behavior and text quality as well as the cognitive, behavioral, and affective dimensions of L2 students' engagement with AWCF.

## Study 1

Study 1 addressed the first research question regarding the extent to which Grammarly and MS-NLP are able to address errors common to L2 learners.

The study was based on a corpus of 68 essays written by incoming international students at a large Midwestern research university as part of an English placement test. The essay was one of two integrated reading-writing tasks. Working in a computer testing center, the students were presented with two short readings on the topic of violent video games. In the longer of the two tasks, students had 30 minutes to write an essay in which they argued for or against banning violent video games, using ideas from the readings and their own views on the topic. The essays were written in the Canvas learning management system in a text-entry tool with no automated checking enabled. The total number of words in the corpus was 20,187. Average text length was 296.87 words ( $SD = 74.63$ ).

The sample consisted of 25 females and 42 males whose most common L1s were Chinese ( $n = 21$ ), Arabic ( $n = 10$ ), Korean ( $n = 6$ ), and Portuguese ( $n = 4$ ); the 21 other L1s included Bengali, Kikuyu, Kurdish, and Vietnamese (gender and L1 data were not available for one test-taker). Fourteen test-takers were graduate students, and the rest were undergraduates. The placement test is given to students who score below 100 on the TOEFL iBT but whose scores meet the minimum threshold for admission to the university (72 for undergraduates and 78 for graduates).

To obtain the AWCF for analysis, we opened each text in MS Word and submitted it for checking to both Grammarly Premium (the paid version of Grammarly, which includes all checks and other features) and MS-NLP. In the former case, checking was achieved using Grammarly's plug-in for MS Word, which, when activated, automatically turns off MS-NLP. All flaggings produced by each tool were screen-recorded using the software TechSmith Morae. In both tools, checks in the style category, and in Grammarly, checks in the vocabulary enhancement and plagiarism categories, were deactivated because the nature of the advice in these categories makes it difficult to characterize flaggings as accurate or inaccurate; for example, blanket recommendations against use of the passive voice or so-called overused words such as "obviously."

Using the annotation functionality of TechSmith Morae, we coded each flagging first for error-type using the labeling provided by the tools themselves. In cases where MS-NLP provided no error-type label, we used labels for similar error types observed in Grammarly so as to facilitate comparison; these included *Misspelled word*, *Capitalization*, "A" vs. "an," and *Verb form*.

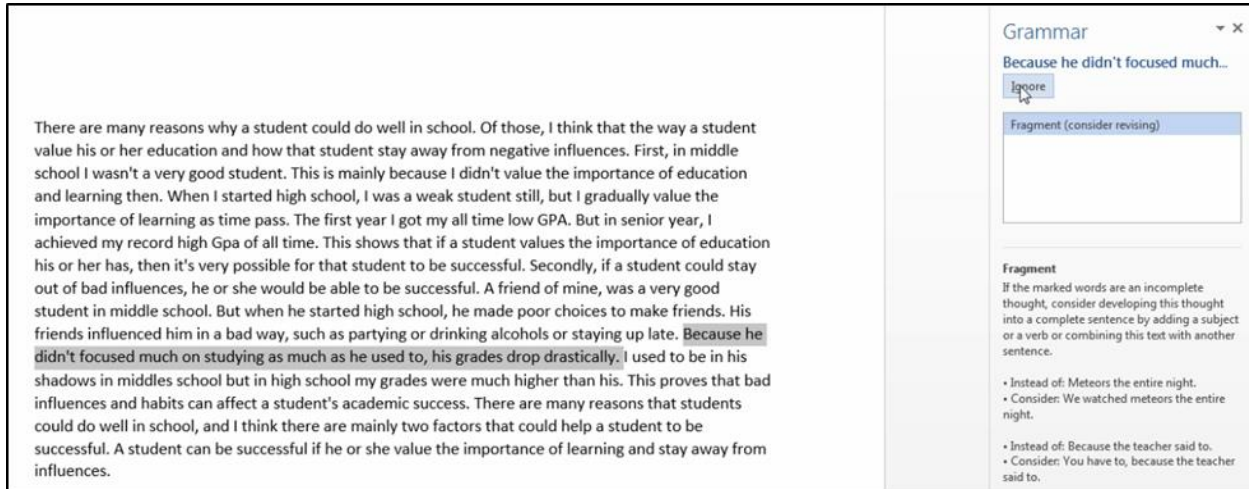
To analyze precision, we coded for two dimensions of accuracy: *accuracy of flagging* (i.e., does the flagged feature indeed represent an error of the specified type?) and *accuracy of correction* (i.e., is the suggested correction appropriate for the specific context?).<sup>3</sup> The following codes were used for both accuracy dimensions: *accurate*, *inaccurate*, and *indeterminate*, with the latter used in cases where the writer's intention was indiscernible. For accuracy of correction, an additional code, *unknown*, was used for cases where the tool provided no specific suggestion. For example, in [Figure 1](#) below, MS-NLP's flagging was coded as *inaccurate* because the highlighted feature does not represent a fragment (although it does contain errors involving verb form and tense), and its correction is coded as *unknown* because of the generic recommendation to "consider revising." [Figure 2](#) shows a case where Grammarly's flagging was coded as *accurate*, but the correction was coded as *inaccurate* because the suggested verb "equip" is not syntactically appropriate in the given context.

The first author and a trained research assistant independently coded all 2,310 error flaggings (1,515 for Grammarly and 795 for MS-NLP) with 91.7% agreement for accuracy of flagging and 89.4% for accuracy of correction. We also calculated interrater reliability (Cohen's kappa) at .674 for accuracy of flagging and .695 for accuracy of correction, both of which represent substantial agreement (Landis & Koch, 1977). Discrepancies were resolved through discussion so that final agreement on both accuracy dimensions was 100%. After exporting the coded data from TechSmith Morae, we calculated precision by first removing 118 items that had been coded *indeterminate* for accuracy of flagging,<sup>4</sup> leaving 1,412 Grammarly flaggings

and 780 MS-NLP flaggings, and then dividing the number of flaggings and corrections coded as accurate by the total number of flaggings and corrections.

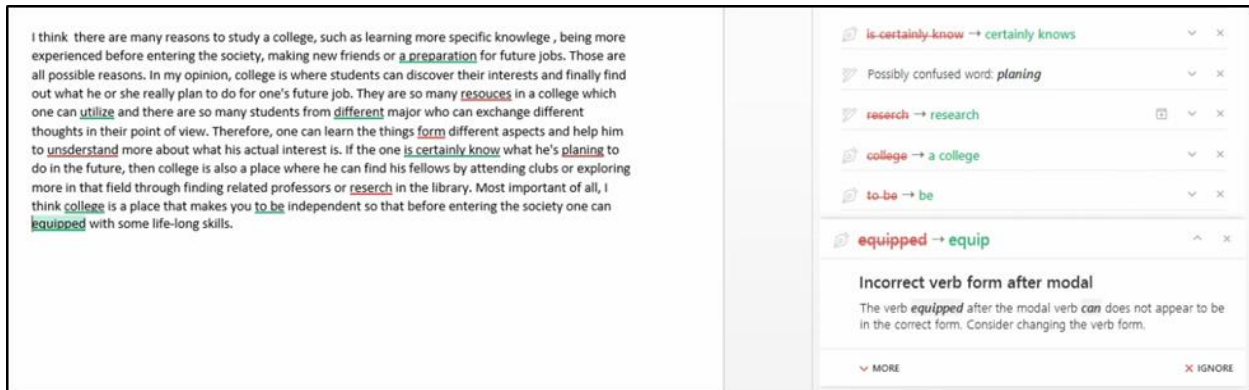
**Figure 1**

*MS-NLP Flagging Coded as Inaccurate and Suggestion Coded as Unknown*



**Figure 2**

*Grammarly Flagging Coded as Accurate and Suggestion Coded as Inaccurate*



We also measured recall of spelling, preposition, article, and subject-verb agreement errors.<sup>5</sup> For each of these analyses, the first author worked with the second author or a native-speaking research assistant to independently review each of the 68 texts manually to identify all errors in each category using a set of annotation guidelines and to propose at least one acceptable correction for each error. Initial agreement for flagging and correction was 88.2% and 88% respectively for spelling errors, 68.9% and 62.2% respectively for preposition errors, 87.2% and 86.9% respectively for article-related errors, and 87.5% for both flagging and correction in the case of Subject-verb agreement errors. Discrepancies were resolved through discussion so that final agreement for both flagging and correction of all four error types was 100%. The final counts were 556 attested spelling errors, 473 attested article errors, 271 attested preposition errors, and 138 attested subject-verb agreement errors. The first author then used the screen-capture annotation data and an MS Excel spreadsheet to record the extent to which the two tools had flagged each attested error and provided at least one of the corrections deemed acceptable. Recall was then calculated by dividing

the number of attested errors flagged by each system by the total number of attested errors, and correction was calculated in a similar manner.

## Results and Discussion

Overall, Grammarly flagged 1,412 items comprising 114 error types (see the complete list in [Appendix A](#)), with precision and correction rates of .88 and .83 respectively. MS-NLP flagged 780 items comprising 22 error types ([Appendix B](#)), with overall precision and correction rates of .92 and .81 respectively. The large discrepancy in the number of error types is partly accounted for by Grammarly's identification of a wider variety of error types but also by the way it differentiates among variants of some types; for example, in addition to a generic *Incorrect verb form*, nine subtypes were specified, including *Incorrect verb form after modal* and *Incorrect verb form after "Do" or "does."* MS-NLP was observed to do this only with respect to errors involving missing or extra spaces, which were divided into *Space between words*, *Space before punctuation*, and *Space after punctuation*.

To facilitate comparisons with the top-ranked L2 problem areas in the Cambridge Learner Corpus ([Table 1](#) above), we decided to modify two CLC categories, *Inflectional morphology* and *Derivational morphology*, because such errors were spread among a number of error types that also included other unrelated errors, particularly in the case of Grammarly. In their place, we substituted a different category, *Verb form*, because this captured morphological errors found by both tools and because verb-form errors have been found to be common in other analyses of errors in L2 English writing (e.g., Chan, 2010). Both tools' coverage of the resulting nine L2 problem areas is reported in [Table 2](#), along with precision rates, correction rates, and the number of error types corresponding to each L2 problem area (see [Appendices A](#) and [B](#) for complete information about these correspondences).

Results showed Grammarly generating more than 10 times the number of flaggings (856) related to the most common L2 problem areas than did MS-NLP (81). Over 60% of Grammarly's total flaggings addressed these L2 problem areas compared to just over 10% of MS-NLP's total flaggings. Overall precision rates were higher for MS-NLP than Grammarly—.88 versus .84, respectively, although the larger and more diverse set of items flagged by Grammarly must be taken into account—while the two tools performed similarly in correction, .81 for Grammarly versus .79 for MS-NLP. No flaggings were recorded for either tool in the category *Wrong verb tense*.<sup>6</sup>

Among Grammarly's flaggings, determiner-related flaggings were the most frequent, accounting for one in every five flaggings. All nine error-types included in this category addressed articles in particular. Collectively, these demonstrated the lowest precision among Grammarly's flaggings (.08), which is attributable to the tool's assumption that a noun phrase not marked as singular (by means of an article) or plural (by adding *-s*) was intended to be singular in form despite it being clear from the context that the plural form was called for (e.g., "These situations happen because \***adolescent** failed to recognize difference between video game and their real life." > **the adolescent**).<sup>7</sup> Regarding preposition errors, Grammarly flagged 79 items across five error types, with both precision and correction rates of .94.

For its part, MS-NLP displayed the highest number of flaggings in the *Agreement error* category. Whereas all of its 61 flaggings constituted subject-verb agreement errors, Grammarly's 160 flaggings included both subject-verb and determiner-noun agreement errors. MS-NLP flagged only eight determiner-related items, all involving the use of "A" versus "an", and two preposition-related items, both involving the need to combine nominal elements following "between" with the word "and" (e.g., "... kids the ages \***between 12-14** > **between 12 and 14**"). These findings highlight the limitations of MS-NLP's rules-based approach insofar as they show that most article and preposition errors are not detectable via parsing and pattern-matching techniques.

Results of the recall analyses are reported in [Table 3](#). For the two types of errors addressed by both tools, recall of spelling errors was where Grammarly and MS-NLP came closest to parity, .86 and .74, respectively. Regarding subject-verb agreement errors, Grammarly identified and corrected about two-thirds of those in the corpus while MS-NLP identified and corrected slightly less than one-third. For the

two error types that required context-analytic techniques to detect (and for which MS-NLP therefore provides no feedback), Grammarly identified one-third of the preposition errors in the corpus and supplied the appropriate correction in slightly less than a third of cases. Grammarly's recall for article-related errors was higher at .47, with a correction rate of .45. Edit-distance data for spelling errors ([Table 4](#)) showed Grammarly finding 10% more multiple-edit spelling errors and providing appropriate corrections for such errors in 25% more cases.

**Table 2***Performance of Grammarly and MS-NLP in Addressing Common L2 Problem Areas*

	Grammarly					MS-NLP				
	Flaggings	% of Total Flaggings	Error-types	Precision	Correction	Flaggings	% of Total Flaggings	Error-types	Precision	Correction
Agreement Error	160	11.33%	22	0.84	0.79	61	7.80%	1	0.87	0.82
Comma Error	141	9.99%	12	0.89	0.88	1	0.10%	1	0	0
Content Word Choice Error	60	4.25%	5	0.83	0.8	2	0.3%	1	1	1
Determiner Error	289	20.47%	9	0.8	0.78	8	1.03%	2	0.88	0.75
Preposition Error	79	5.59%	5	0.94	0.94	2	0.30%	1	1	0
Pronoun Error	30	2.12%	6	0.9	0.83	1	0.10%	1	1	1
Run-on Sentence	2	0.14%	1	1	1	-	-	-	-	-
Verb-form Error	95	6.73%	19	0.84	0.73	6	0.80%	1	1	0.83
Wrong verb Tense	-	-	-	-	-	-	-	-	-	-
Total	856	60.62%	79	0.84	0.81	81	10.43%	8	0.88	0.79

**Table 3***Comparison of Recall and Correction Rates for Selected Error Types*

<b>Error Type</b>	<b><i>n</i></b>	<b>Grammarly</b>		<b>MS-NLP</b>	
		<b>Recall</b>	<b>Correction</b>	<b>Recall</b>	<b>Correction</b>
Spelling	556	0.86	0.84	0.74	0.67
Article	437	0.47	0.45	-	-
Preposition	271	0.33	0.31	-	-
Subject-Verb Agreement	138	0.67	0.65	0.35	0.34

**Table 4***Recall and Correction Rates for Spelling Errors According to Edit Distance*

<b>Edit distance</b>	<b>Number of Errors</b>	<b>Grammarly</b>		<b>MS-NLP</b>	
		<b>Recall</b>	<b>Correction</b>	<b>Recall</b>	<b>Correction</b>
Single-edit	431	0.93	0.92	0.80	0.78
Multiple-edit	125	0.68	0.58	0.58	0.33

Having established that Grammarly is able to find and correct more common L2 error types than MS-NLP does with greater accuracy, we move on to the consequences of Grammarly's enhanced error-correction capacities regarding feedback timing, which was the focus of the second study.

## Study 2

Study 2 addressed the second research question regarding how Grammarly and MS-NLP differ in terms of the timing of error-flagging.

We recruited 20 students from an ESL writing course for undergraduates at the same university as Study 1. The group consisted of 10 females and 10 males whose average age was 19.6 years ( $SD = 1.31$ ). Ten of the students spoke Chinese as their first language; other L1s included Malay, Pali, Japanese, Indonesian, and Brazilian Portuguese. Having been placed into the ESL writing program by the placement test described in Study 1, their TOEFL iBT scores would have been between 72 and 100. Each participant wrote two short essays based on different prompts, both addressing educational themes (*What makes a successful student?* and *Why do people attend college or university?*). The time limit for each essay task was 30 minutes. Average text length was 299.8 words ( $SD = 72.95$ ). The 40 essays together comprised 11,733 words.

Participants composed their essays individually in the research team's office on a desktop computer connected to the university's network via an Ethernet cable to ensure the fastest possible internet connection. One essay was written with MS-NLP operating synchronously and the other with the Grammarly plug-in for MS Word operating synchronously.

Because our goal in Study 2 was to compare typical feedback timing across the two tools, we accepted each system's default checks for synchronous operation on the assumption that most L2 students do not modify these settings. For MS-NLP, this meant only spelling and "format consistency" errors (e.g., missing or extra

spaces) would be identified; in the version of Grammarly Premium available at the time, this meant only the spelling, punctuation, grammar, and sentence structure checks were activated. Tasks and tools were counterbalanced to avoid ordering effects. TechSmith Morae was used for recording, coding, and measuring on-screen activity. [Inputlog](#) was used to record and analyze participants' keystroke timing.

To measure differences in the timing of error flagging across tools, we first categorized each flagging according to whether it occurred at the point of inscription (i.e., where the writer was currently typing) or at a point earlier in the text, and then tallied up all instances of each timing category. Next, to estimate an average time lag between error production and flagging, we focused on cases where flagging occurred earlier in the text; the timeline markers in TechSmith Morae were used to measure the elapsed time between (a) completion of the text that would be flagged and (b) the flagging itself. These measures were then used to calculate the average feedback delay for each participant.

Finally, to relate the timing of feedback delivery to the timing of p-bursts, we used the automated analysis tools in Inputlog to collect average p-burst lengths for each participant based on a pause threshold of two seconds between two typing events.<sup>8</sup> Two seconds is a standard benchmark (Chanquoy et al., 1996; Spelman Miller, 2000; Sullivan & Lindgren, 2006) for separating pauses indicative of cognitive activity, such as planning or evaluating, from pauses indicating transitions between keystrokes, which are motoric in nature and generally average less than one second among even the slowest writers (Sullivan & Lindgren, 2006).

## Results and Discussion

In total, Grammarly flagged 535 items as errors compared to MS-NLP's 536 flaggings. Frequencies for the two temporal location categories ([Table 5](#)) showed that only about one in five of Grammarly's flaggings occurred at the point of inscription, with the rest (78.9%) occurring earlier in the text. By contrast, nearly nine out of ten of MS-NLP's flaggings occurred at the point of inscription, with only 11.4% occurring earlier in the text. This can be attributed to the fact that MS-NLP's default checks for spelling and format consistency relied on simple pattern matching routines that could be performed nearly instantaneously.

**Table 5**

*Frequency of Flaggings by Temporal Location*

	Grammarly	MS-NLP
Point of inscription	113 (21.1%)	475 (88.6%)
Earlier in the text	422 (78.9%)	61 (11.4%)
Total	535	536

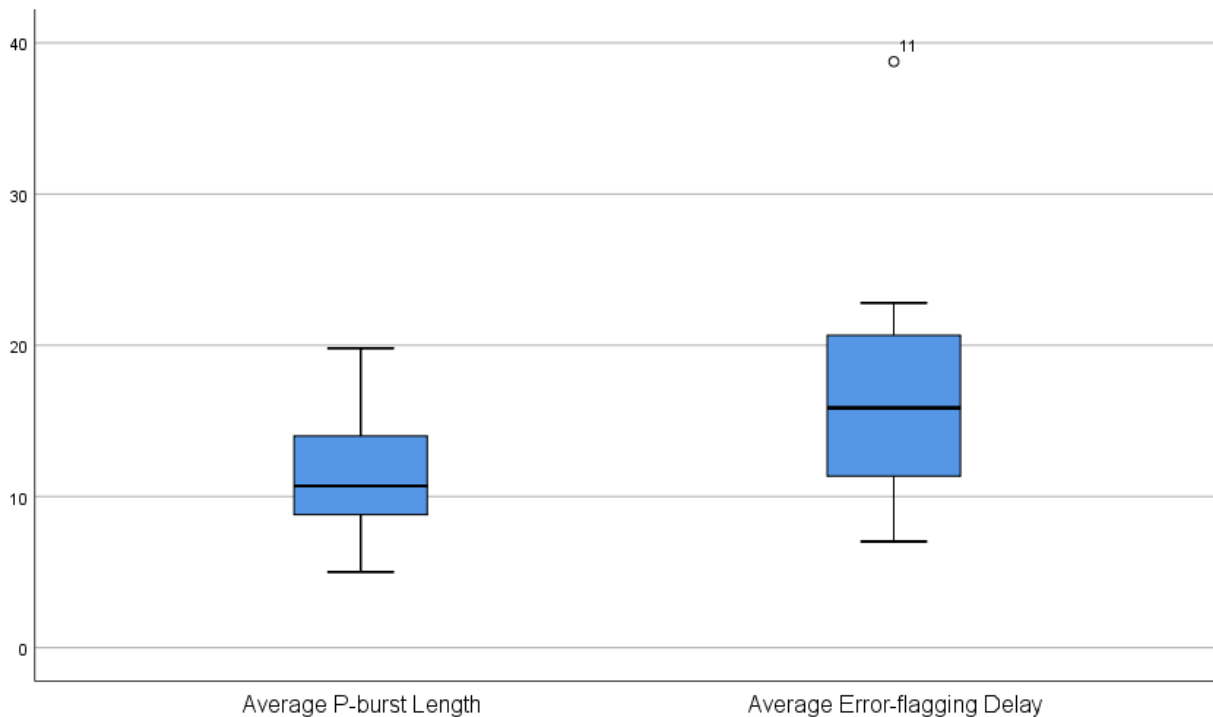
Based on these findings, we decided to restrict our analysis of feedback timing relative to p-burst duration to the Grammarly data only. Average p-burst duration for the group while using Grammarly was 11.63 seconds ( $SD = 4.38$ ) while the average feedback delay experienced among the participants while using Grammarly was 17.45 seconds ( $SD = 35.52$ ). Boxplots of these data ([Figure 3](#)) show the comparatively greater variance in feedback delay, which is probably attributable to the considerable variation in terms of when writers would type sentence-final punctuation (a variable, as noted above, that can delay initialization of some of Grammarly's checks).

To test for significance, we used a non-parametric technique because of the unequal variances. A Wilcoxon Signed Ranks test comparing the two distributions showed that the feedback-delay ranks (median = 14.45 seconds) were statistically higher than the p-burst ranks (median = 10.7 seconds),  $Z = -2.80$ ,  $p = .005$ ,  $r =$

-.442 (a medium to large effect according to the scale in Fritz et al., 2012). This means that when Grammarly was operating synchronously, new flaggings usually appeared in segments of previously written text as opposed to the segment being transcribed contemporaneously.

### Figure 3

*Average P-burst Length and Error-flagging Delay in the Grammarly Condition (N = 20)*



## General Discussion

To briefly summarize the results, Grammarly outperformed MS-NLP by flagging a much larger and more diverse set of error types representing both common problems for L2 writers of English as well as complex computational challenges. Despite the differences in the frequency and range of error types detected, Grammarly's precision was only slightly lower than MS-NLP's and its correction rate higher. For comparison's sake, Grammarly's overall precision rate and individual precision rates for most error types ([Appendix A](#)) met the 80% benchmark used by developers of Criterion (Quinlan et al., 2009).

In terms of recall, the two tools approached parity with respect to spelling errors, but edit distance data showed Grammarly better addressing the more complex spelling errors characteristic of L2 learners. Furthermore, Grammarly flagged and corrected nearly twice the number of subject-verb agreement errors as MS-NLP. Grammarly also flagged and corrected about half of all attested article errors and a third of all preposition errors, both common L2 error types that MS-NLP could not address in any substantive way. Evidently, progress has been made regarding the call in Rimrott and Heift (2008) for spell checkers that better attend to the needs of L2 writers. Progress may also be reflected in the recall and correction rates for article and preposition errors found here insofar as they exceeded those reported for the experimental systems in both Han et al. (2006) and Tetreault and Chodorow (2008).

Regarding feedback timing, the average delay in Grammarly's flaggings exceeded the average p-burst

length of the L2 student writers in our sample, suggesting that these writers were generally receiving feedback from Grammarly about linguistic units that no longer represented the current contents of verbal working memory, and which thus constituted potential distractions. By contrast, the majority of MS-NLP's flaggings occurred contemporaneously with the completion of the word that was flagged, which meant this feedback likely coincided with the concurrent contents of verbal working memory.

### Implications

L2 student writers stand to benefit from the enhanced error-correction capabilities of AWCF tools. However, research has shown that such writers already prioritize sentence-level grammatical concerns over higher-level issues in evaluating and revising their work (Barkaoui, 2007). Regardless of feedback delays, there may be a risk of reinforcing students' low-level focus by facilitating continuous access to CF without considering the nature of the task at hand. At the same time, there may be writing tasks for which slightly delayed synchronous AWCF proves not only harmless to task completion (e.g., a simple email message) but supportive of learning (e.g., an L2 practice task addressing definite/indefinite/zero articles), particularly when one considers the 40-second cognitive window for ideal focus on form proposed by Doughty (2001), within which Grammarly's synchronous feedback would generally appear. In this regard, Manchón's (2011) contrast between *learning to write* and *writing to learn* may be helpful for differentiating among writing-task types in pedagogical contexts. For *learning to write* tasks, where the focus is on development of writing skills, students can be taught to evaluate tasks as to the amount of cognitive demand that they will require, and on this basis, to decide whether to use a tool like Grammarly synchronously or asynchronously. In addition to avoiding potential distractions, restricting the use of AWCF to appropriate junctures in the writing process may create space for students to consider the feedback more carefully, which can help both with identifying inaccurate flaggings and possibly facilitating L2 learning.

A second implication concerns the development and marketing of AWCF tools. Companies such as Grammarly should make users aware of the potential for problems with synchronous operation. In their promotional materials and user tutorials, they should inform users of the need to consider the impact of constant access to CF, and encourage them to make strategic decisions about their engagement with Grammarly based on these considerations. AWCF tools should also allow users to toggle the program's checks on and off. In the version of Grammarly used in the present study, this was easily accomplished using buttons for each category of check (e.g., grammar, punctuation, style). In the new web-interface released to paid subscribers in 2019, it does not appear possible to completely deactivate checking.

The final implication is that the results affirm our view that Grammarly and other tools that (a) use sophisticated, hybrid GEC approaches to target unique problem areas for L2 writers, and (b) operate synchronously across multiple applications, platforms, and devices—the other major commercially available exemplar being Ginger (see Swier, 2016, for a description of the latter)—represent a distinct genre of writing-support technology that must be recognized and understood on its own terms. This means differentiating it from MS-NLP as well as asynchronous AWE tools such as Criterion and MY Access!, which may incorporate similar error-correction technologies but which users interact with differently. The case for a new genre is strengthened when one considers that the term *AWE* has also been applied to systems that do not address sentence-level concerns, such as Writing Pal (Roscoe & McNamara, 2013) and the Research Writing Tutor (Cotos, 2015), which target writing strategies and functional discourse, respectively. Thus, we have proposed the term *AWCF tool* (see also Ranalli, 2018), to try to capture the unique qualities of this new genre. Although the term *GEC* is already in currency among developers, we note that *AWCF* connects these tools to existing research in both the SLA and L2 writing domains regarding teacher-provided WCF, with which we see potential synergy. The main features differentiating AWCF tools from AWE tools and MS-NLP are summarized in [Table 6](#).

**Table 6**

*Comparison of Automated Written Corrective Feedback (AWCF) Tools, Automated Writing Evaluation (AWE) Tools, and Microsoft Natural Language Processing (MS-NLP)*

	<b>AWCF Tools</b>	<b>AWE Tools</b>	<b>MS-NLP</b>
Examples	<i>Grammarly, Ginger, Grammar Suggestions</i>	<i>Criterion, MY Access!, Research Writing Tutor, Writing Pal</i>	The spelling and grammar checker in MS Word, Outlook, and Office 365
Access	Multiple ways (e.g., web apps, browser extensions, productivity software plugins, mobile device keyboards)	Standalone web-based interfaces	Office productivity software
Delivery Mode	Synchronous and asynchronous	Asynchronous only	Synchronous and asynchronous
Analysis	Combinations of complex techniques performed on remote servers	Combinations of complex techniques performed on remote servers	Simpler techniques performed on user's local machine
Focus	Lower-level concerns (e.g., spelling, grammar, and punctuation)	Lower- and higher-level, or only higher-level, concerns (e.g., organization, discourse, writing strategies)	Lower-level concerns

*Note.* *Grammar Suggestions* is a machine-learning based synchronous AWCF tool that was included in Google Docs starting in 2019.

### Limitations and Future Research

The small, relatively homogenous sample, the use of a single writing task type, and the fact that only four error types were addressed in the recall analysis together mean that caution should be exercised in generalizing the findings regarding both error-correction performance and feedback timing. In addition, we note that Grammarly is continuously being developed and improved; in the two years between the start of the project and the writing of this report, Grammarly's claims about the number of features it could identify increased from 250 to 400 (Grammarly, n.d.). The most significant limitation of the study, however, is that no perception or learning data were collected. Research is needed to confirm whether discrepancies in feedback timing relative to p-burst duration are indeed perceived as distracting by users and, if so, what individual, contextual, and task conditions influence such perceptions. Verbal reports, eye-tracking, or both could be used in such studies to triangulate keystroke logging data and increase veridicality. Similarly, research should investigate the potential for L2 development as a result of exposure to synchronous AWCF, with writing goal (i.e., *writing to learn* vs. *learning to write*) and feedback timing (e.g., point of inscription versus earlier in the text) as independent variables.

In addition, studies evaluating other AWCF tools, including comparison studies, are needed to help potential users in deciding which, if any, could suit their needs while also revealing meaningful variation within the genre. Because these tools are commercial products, the for-profit companies that produce them are less forthcoming with information about their performance (Perelman, 2016), so such studies will provide a valuable public service. Such evaluations could also address error-correction performance regarding collocation errors, which is another area of GEC innovation (Leacock et al., 2014). Finally, research is needed to understand how continuous exposure to synchronous AWCF tools influences L2

students' own self-initiated revisions, given that the ability to identify, diagnose, and correct errors in one's own writing is a vital skill. As mentioned, this study was part of a larger project that also investigated this latter concern.

## Conclusion

Much work has gone into the development of automated error-correction technologies for L2 writers of English. These innovations are spreading widely as Grammarly and other similar tools attract new users. Grammarly's capacity to correct frequent and complex L2 written errors based on these technologies comes at a cost in terms of the speed with which it can deliver its feedback, and this has consequences for the way users interact with the tool. This study sought to advance basic understanding of these issues so as to inform applications of Grammarly and similar tools to L2 writing and L2 learning in ways that can support the work of researchers, educators, and students.

## Acknowledgements

The authors gratefully acknowledge the diligent work of research assistants Elizabeth Lee, Cay Bappe, and Gabi Mitchell.

## Notes

1. In information science, *entropy* refers to a measure of average uncertainty with respect to a given variable's possible outcomes.
2. In 2016, Microsoft claimed to have 1.2 billion users of its Office suite along with 60 million active users of its cloud-based Office 365 service, which also includes MS Word. At the time of writing, Grammarly claims a daily user base of 20 million.
3. In the case of MS-NLP, which often provides multiple suggestions, feedback was coded as *accurate* if an appropriate correction appeared anywhere in the list of suggestions.
4. Our rationale for removing items coded as *indeterminate* for accuracy of flagging was that if human annotators could not identify what form was appropriate for a writer's intended meaning, an automated checker could not reasonably be expected to do so either, so such items should not detract from the tool's precision and recall statistics.
5. A comprehensive recall analysis of all the errors in our corpus was not attempted because such analyses are notoriously time-consuming and difficult, as has been noted in both the GEC (e.g., Leacock et al., 2014) and L2 writing (e.g., Polio, 1997) literatures. The number and type of errors found will depend on how particular errors are corrected, and because many L2 errors can be corrected in multiple ways, a comprehensive analysis would need to produce a correction not only for every error but for every possible interpretation of every error, and intercoder reliability would be low. Instead, we followed the recent practice of GEC developers annotating corpora for specific error types relevant to a given purpose (e.g., Tetreault & Chodorow, 2008). We chose spelling and subject-verb agreement errors because we knew these were detected by both Grammarly and MS-NLP and thus they could facilitate informative comparisons. Additionally, we chose article and preposition errors because they represent key L2 problem areas that GEC developers have addressed recently through innovative techniques, as discussed in the literature review.
6. In Grammarly, one error-type involving a tense label, *Incorrect verb form in perfect tense*, ended up being categorized under the L2 problem area *Verb-form error*. Grammarly generated eight flaggings for two other tense-related error types, *Incorrect verb tense* (6) and *Incorrect use of progressive tense* (2), but because all of these flaggings were coded *indeterminate* for accuracy of flagging, they were not included in our analyses. Grammarly clearly attempts to address verb-tense

- errors but without much success with respect to our corpus.
7. Grammarly could take a cue here from MS-NLP, which in correcting Subject-Verb Agreement errors supplies both a singular and plural option (e.g., “If **\*a person use** more time ... > **a person uses | people use**”).
  8. A second analysis was conducted using an individual pause threshold defined for each participant using inter-keystroke interval (IKI) measures from a typing task that had been administered as part of the research protocol. This pause threshold was determined by multiplying each individual’s average IKI by three, as recommended in Wengelin (2006). These individual pause thresholds resulted in an average p-burst duration of 3.11 seconds ( $SD = 1.15$ ) for the sample ( $N = 20$ ). We decided to report the analysis incorporating the larger p-burst duration in order to maintain comparability with previous research and to err on the side of conservatism. Needless to say, adoption of the individualized pause threshold would not only support but bolster our interpretations in Study 2.

## References

- Arroyo, D. C., & Yilmaz, Y. (2018). An open for replication study: The role of feedback timing in synchronous computer-mediated communication. *Language Learning*, 68(4), 942–972. <https://doi.org/10.1111/lang.12300>
- Barkaoui, K. (2007). Revision in second language writing: What teachers need to know. *TESL Canada Journal*, 25(1), 81–92. <https://doi.org/10.18806/tesl.v25i1.109>
- Bestgen, Y., & Granger, S. (2011). Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2–3), 235–252.
- Bishop, T. (2005, March 27). A Word to the unwise—program’s grammar check isn’t so smart. *Seattle Post-Intelligencer*. <https://www.seattlepi.com/business/article/A-Word-to-the-unwise-program-s-grammar-check-1169572.php>
- Chan, A. Y. W. (2010). Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners. *TESOL Quarterly*, 44(2), 295–319. <https://doi.org/10.5054/tq.2010.219941>
- Chanquoy, L., Foulin, J.-N., & Fayol, M. (1996). Writing in adults: A real-time approach. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 36–43). Amsterdam University Press.
- Chen, C.-F. E., & Cheng, W.-Y. E. (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. <https://www.lltjournal.org/item/2631>
- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80–98. <https://doi.org/10.1177/0741088301018001004>
- Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication*, 20(1), 99–118. <https://doi.org/10.1177/0741088303253572>
- Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3), 419–436. <https://doi.org/10.1177/0265532210364391>
- Connors, R. J., & Lunsford, A. A. (1988). Frequency of formal errors in current college writing, or Ma and Pa Kettle do research. *College Composition and Communication*, 39(4), 395–409. <https://doi.org/10.2307/357695>

- Cotos, E. (2015). Automated Writing Analysis for writing pedagogy: From healthy tension to tangible prospects. *Writing and Pedagogy*, 7(2–3), 197–231. <https://core.ac.uk/download/pdf/38937327.pdf>
- Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal*, 28(1), 99–134.
- Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and Second Language Acquisition* (pp. 206–257). Cambridge University Press.
- Flor, M., & Futagi, Y. (2012). *On using context for automatic correction of non-word misspellings in student essays* [Paper presentation]. The Seventh Workshop on Building Educational Applications Using NLP, Montréal, Canada. <https://aclanthology.org/W12-2012.pdf>
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1), 2–18. <https://doi.org/10.1037/a0024338>
- Galbraith, D., & Vedder, I. (2019). Methodological advances in investigating L2 writing processes: Challenges and perspectives. *Studies in Second Language Acquisition*, 41(3), 633–645. <https://doi.org/10.1017/S0272263119000366>
- Gamon, M., Leacock, C., Brockett, C., Dolan, W. B., Gao, J., Belenko, D., & Klementiev, A. (2009). Using statistical techniques and web search to correct ESL errors. *CALICO Journal*, 26(3), 491–511.
- Grammarly. (n.d.). What is Grammarly Premium, and how is it different from the free version? *Grammarly Support*. <https://support.grammarly.com/hc/en-us/articles/115000090812-What-is-Grammarly-Premium-and-how-is-it-different-from-the-free-version->
- Han, N.-R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12(2), 115–129. <https://doi.org/10.1017/S1351324906004190>
- Heift, T., & Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. Routledge.
- Kies, D. (2008). *Evaluating grammar checkers: A comparative ten-year study* [Paper presentation]. The 6th International Conference on Education and Information Systems, Technologies and Applications: EISTA, Orlando, Florida, USA.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, 19(2), 50–68. <https://www.lltjournal.org/item/2903>
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). Automated grammatical error detection for language learners, second edition. *Synthesis Lectures on Human Language Technologies*, 7(1), 1–154. <https://doi.org/10.2200/S00562ED1V01Y201401HLT025>
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3), 358–392. <https://doi.org/10.1177/0741088313491692>
- Long, M. H. (2007). *Problems in SLA*. Lawrence Erlbaum.
- Manchón, R. M. (Ed.) (2011). *Learning-to-write and writing-to-learn in an additional language*. John Benjamins.

- Nagata, R., & Nakatani, K. (2010). *Evaluating performance of grammatical error detection to maximize learning effect* [Paper presentation]. Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China. <https://aclanthology.org/C10-2103/>
- Nagata, R., Takamura, H., & Neubig, G. (2017). Adaptive spelling error correction models for learner English. *Procedia Computer Science*, 112, 474–483. <https://doi.org/10.1016/j.procs.2017.08.065>
- Napoles, C., Nădejde, M., & Tetreault, J. (2019). Enabling robust grammatical error correction in new domains: Data sets, metrics, and analyses. *Transactions of the Association for Computational Linguistics*, 7, 551–566. [https://doi.org/10.1162/tacl\\_a\\_00282](https://doi.org/10.1162/tacl_a_00282)
- Perelman, L. (2016). Grammar checkers do not work. *WLN: A Journal of Writing Center Scholarship*, 40(7–8), 11–20.
- Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct coverage of the e-rater scoring engine. *Educational Testing Service*, 2009(1), i–35. <https://doi.org/10.1002/j.2333-8504.2009.tb02158.x>
- Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, 31(7), 653–674. <https://doi.org/10.1080/09588221.2018.1428994>
- 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. <https://doi.org/10.1080/01443410.2015.1136407>
- Ransdell, S., Levy, C. M., & Kellogg, R. T. (2002). The structure of writing processes as revealed by secondary task demands. *L1-Educational Studies in Language and Literature*, 2(2), 141–163. <https://doi.org/10.1023/A:1020851300668>
- Rimrott, A., & Heift, T. (2005). Language learners and generic spell checkers in CALL. *CALICO Journal*, 23(1), 17–48.
- Rimrott, A., & Heift, T. (2008). Evaluating automatic detection of misspellings in German. *Language Learning & Technology*, 12(3), 73–92. <https://www.lltjournal.org/item/2642>
- Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010–1025. <https://doi.org/10.1037/a0032340>
- Shintani, N., & Aubrey, S. (2016). The effectiveness of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. *The Modern Language Journal*, 100(1), 296–319. <https://doi.org/10.1111/modl.12317>
- Spelman Miller, K. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research*, 4(2), 123–148. <https://doi.org/10.1177/136216880000400203>
- Sullivan, K. P. H., & Lindgren, E. (2006). Analysing online revision. In G. Rijlaarsdam (Series Ed.) and K. P. H. Sullivan, & E. Lindgren (Volume Eds.), *Computer key-stroke logging and writing: Methods and applications (Studies in Writing, Vol. 18)* (pp. 157–18). Elsevier. [https://doi.org/10.1163/9780080460932\\_008](https://doi.org/10.1163/9780080460932_008)
- Swier, R. (2016). Ginger software suite of writing services & apps. *CALICO Journal*, 33(2), 282–290.
- Tetreault, J. R., & Chodorow, M. (2008). *Native judgments of non-native usage: Experiments in preposition error detection*. [Workshop]. Human Judgements in Computational Linguistics, Manchester, UK. <https://aclanthology.org/W08-1205.pdf>

- Van Waes, L., Leijten, M., & Quinlan, T. (2010). Reading during sentence composing and error correction: A multilevel analysis of the influences of task complexity. *Reading and Writing*, 23(7), 803–834. <https://doi.org/10.1007/s11145-009-9190-x>
- Wengelin, Å. (2006). Examining pauses in writing: Theory, methods, and empirical data. In G. Rijlaarsdam (Series Ed.) and K. P. H. Sullivan, & E. Lindgren (Volume Eds.), *Computer key-stroke logging and writing: Methods and applications (Studies in Writing, Vol. 18)* (pp. 107–130). Elsevier. [https://doi.org/10.1163/9780080460932\\_008](https://doi.org/10.1163/9780080460932_008)

## Appendix A

### Grammarly Error-flagging Data Including Precision, Correction, and Correspondence with L2 Problem Areas

Error Type	L2 Problem Area	Total Flaggings	Percentage of Total	Precision	Correction
Misspelled word		350	24.79%	0.99	0.95
Missing article	DET	137	9.70%	0.61	0.58
Incorrect article use	DET	122	8.64%	0.97	0.96
Possibly confused word	CWC*	101	7.15%	0.85	0.81
Missing comma in compound sentence	COM	50	3.54%	0.88	0.88
Confused preposition	PRP	45	3.19%	0.93	0.93
Incorrect verb form	VFE	38	2.69%	0.79	0.55
Incorrect verb form with plural subject	AGR	32	2.27%	0.88	0.88
Incorrect verb form with singular subject	AGR	32	2.27%	0.72	0.63
Sentence fragment		26	1.84%	0.54	0.12
Missing comma after introductory phrase	COM	24	1.70%	0.92	0.92
Possibly miswritten word		21	1.49%	0.86	0.81
Comma splice	COM	19	1.35%	0.79	0.79
Incorrect verb form with personal pronoun	AGR	15	1.06%	0.80	0.73
Lowercase pronoun “I”		14	0.99%	1.00	1.00
Missing preposition	PRP	14	0.99%	0.93	0.93
Missing pronoun	PRN	14	0.99%	1.00	0.86
Incorrect noun form	AGR	13	0.92%	0.85	0.77
Missing hyphen		12	0.85%	0.92	1.00
Redundant preposition	PRP	12	0.85%	1.00	1.00
Wrong verb form	VFE	12	0.85%	0.83	0.75
Incorrect verb form after modal	VFE	11	0.78%	1.00	1.00
Singular noun after plural quantifier	AGR	11	0.78%	1.00	1.00
“These” with singular noun	AGR	11	0.78%	1.00	1.00

Missing word		10	0.71%	0.90	0.90
Redundant indefinite article	DET	9	0.64%	1.00	1.00
Unnecessary comma in complex sentence	COM	9	0.64%	0.89	0.89
Incorrect use of comma	COM	8	0.57%	0.75	0.75
Indefinite article with plural noun	AGR	8	0.57%	1.00	1.00
Missing comma after introductory clause	COM	8	0.57%	1.00	1.00
Missing comma(s) with interrupter	COM	8	0.57%	1.00	0.88
Inconsistent spelling		7	0.50%	1.00	1.00
Missing verb	VFE	7	0.50%	0.86	0.86
Possibly confused “affect” and “effect”	CWC	7	0.50%	1.00	0.86
Wrong article with set expression	DET	7	0.50%	1.00	1.00
Confused pronoun	PRN	6	0.42%	1.00	1.00
Incorrect preposition after adjective	PRP	6	0.42%	0.83	0.83
Incorrect punctuation		6	0.42%	1.00	1.00
Missing comma in a series	COM	6	0.42%	1.00	1.00
The use of “a” versus “an”	DET	6	0.42%	0.83	0.83
Unknown word		6	0.42%	0.67	0.00
Unnecessary pronoun	PRN	6	0.42%	1.00	1.00
Incorrect verb form in perfect tense	VFE	5	0.35%	1.00	1.00
Misused determiner	AGR	5	0.35%	0.80	0.40
Redundant use of article	DET	5	0.35%	1.00	1.00
Singular noun with plural number	AGR	5	0.35%	0.80	0.80
Infinitive instead of gerund	VFE	4	0.28%	0.50	0.50
Possibly confused “who” and “whom”		4	0.28%	0.75	0.75
To-infinitive instead of bare form	VFE	4	0.28%	1.00	1.00
Confused “everytime” and “every time”		3	0.21%	1.00	1.00
Confused possessive and contraction		3	0.21%	0.33	0.33
Faulty parallelism		3	0.21%	1.00	0.67
Improper comma between subject and verb	COM	3	0.21%	0.67	0.67
Incorrect punctuation with abbreviation		3	0.21%	1.00	1.00
Incorrect verb form with indefinite pronoun	AGR	3	0.21%	1.00	1.00
Missing comma	COM	3	0.21%	1.00	1.00
“Other” with singular noun	AGR	3	0.21%	0.67	0.67
Possibly confused “lets” and “let's”		3	0.21%	1.00	1.00

Possibly confused “specially” and “especially”	CWC	3	0.21%	1.00	1.00
Singular quantifier with plural noun	AGR	3	0.21%	1.00	1.00
“There” with singular noun	AGR	3	0.21%	0.67	0.67
“This” with plural noun	AGR	3	0.21%	0.67	0.67
“Those” with singular noun	AGR	3	0.21%	1.00	0.67
Unnecessary ellipsis		3	0.21%	1.00	0.67
Adjective instead of adverb	CWC	2	0.14%	1.00	1.00
Adverb instead of adjective	CWC	2	0.14%	1.00	1.00
Capitalization		2	0.14%	1.00	1.00
Confused “which” and “who”	PRN	2	0.14%	0.00	0.00
Dangling modifier		2	0.14%	0.50	0.50
Incorrect comma	COM	2	0.14%	1.00	1.00
Incorrect form for to-infinitive	VFE	2	0.14%	1.00	1.00
Incorrect plural verb with collective noun	AGR	2	0.14%	0.50	0.50
Incorrect punctuation with quotation mark		2	0.14%	1.00	1.00
Incorrect quantifier	AGR	2	0.14%	1.00	1.00
Incorrect verb		2	0.14%	0.00	0.00
Incorrect verb form with compound subject	AGR	2	0.14%	1.00	1.00
Incorrect verb form with conditional	VFE	2	0.14%	1.00	1.00
Inversion		2	0.14%	1.00	0.00
Missing subject		2	0.14%	0.50	0.00
Possibly confused “other” and “others”		2	0.14%	1.00	1.00
Redundancy	PRP	2	0.14%	1.00	1.00
Run-on sentence	ROS	2	0.14%	1.00	1.00
Squinting modifier		2	0.14%	0.00	0.00
Compound instead of comparative		1	0.07%	1.00	1.00
“Do” with modal verb	VFE	1	0.07%	1.00	1.00
Double comparative		1	0.07%	1.00	1.00
Gerund instead of to-infinitive	VFE	1	0.07%	0.00	0.00
Incorrect modifier	AGR	1	0.07%	1.00	1.00
Incorrect negative verb form	VFE	1	0.07%	1.00	1.00
Incorrect quantifier with uncountable noun	AGR	1	0.07%	0.00	0.00
Incorrect verb form after “do” or “does”	VFE	1	0.07%	1.00	1.00
Intransitive verb in passive voice	VFE	1	0.07%	1.00	1.00

Missing article before noun	DET	1	0.07%	1.00	1.00
Missing hyphen in a number		1	0.07%	1.00	1.00
Non-infinitive after “to”	VFE	1	0.07%	1.00	1.00
Past participle without auxiliary verb	VFE	1	0.07%	0.00	0.00
Personal instead of possessive pronoun	PRN	1	0.07%	0.00	0.00
Plural noun with singular verb	AGR	1	0.07%	1.00	1.00
Possibly confused “everyday” and “every day”		1	0.07%	1.00	1.00
Possibly confused “sometime” and “some time”		1	0.07%	1.00	1.00
Possibly confused “than” and “then”		1	0.07%	0.00	0.00
Possibly confused “there” and “their”		1	0.07%	1.00	1.00
Redundant article	DET	1	0.07%	1.00	1.00
Redundant comma before “not”	COM	1	0.07%	1.00	1.00
Redundant determiner	DET	1	0.07%	1.00	1.00
Redundant reflexive pronoun	PRN	1	0.07%	1.00	1.00
Redundant word		1	0.07%	1.00	1.00
Repeated word		1	0.07%	1.00	1.00
Simple or compound adjective		1	0.07%	0.00	0.00
“That” with plural noun	AGR	1	0.07%	1.00	1.00
“To” after modal verb	VFE	1	0.07%	1.00	1.00
To-infinitive instead of prepositional phrase	VFE	1	0.07%	1.00	0.00
Wrong verb form after “be\get used to”	VFE	1	0.07%	1.00	1.00
Wrong word choice	*	1	0.07%	0.00	0.00
Total		1412	100%	0.88	0.83

## Appendix B

### MS-NLP Error-flagging Data Including Precision, Correction, and Correspondence with L2 Problem Areas

Error Type	L2 Problem Area	Total Flaggings	Percentage Overall	Precision	Correction
Misspelled word		464	59.5%	0.92	0.80
Space after punctuation		75	9.6%	1.00	0.97
Subject-verb agreement	AGR	61	7.8%	0.87	0.82
Capitalization		56	7.2%	0.95	0.95
Space between words		28	3.6%	1.00	0.96
Fragment		25	3.2%	0.68	0.00
Space before punctuation		24	3.1%	1.00	1.00
Possessive use		9	1.2%	0.89	0.89
“A” versus “an”	DET	7	0.9%	0.86	0.71
Verb form	VFE	6	0.8%	1.00	0.83
Order of words		5	0.6%	0.80	0.40
Commonly confused words	CWC*	4	0.5%	0.75	0.75
Comparative use		4	0.5%	1.00	1.00
Punctuation		3	0.4%	1.00	0.33
Use of “between”	PRP	2	0.3%	1.00	0.00
Pronoun use	PRN	1	0.1%	1.00	1.00
Comma use	COM	1	0.1%	0.00	0.00
Conjunction use		1	0.1%	0.00	0.00
Extra word	DET	1	0.1%	1.00	1.00
Possible question		1	0.1%	0.00	0.00
Question mark use		1	0.1%	1.00	1.00
Repeated word		1	0.1%	1.00	1.00
Total		780	100%	0.92	0.81

## About the Authors

Jim Ranalli, PhD, is an Assistant Professor in the TESL/Applied Linguistics Program at Iowa State University. His research addresses the intersection of L2 writing, technology, and self-regulated learning. He is particularly interested in innovative uses of computers for scaffolding and assessing the development of EAP writing skills.

**E-mail:** [jranalli@iastate.edu](mailto:jranalli@iastate.edu)

Taichi Yamashita, PhD, is a Visiting Assistant Professor in the Department of World Languages & Cultures at the University of Toledo. His research interests include instructed second language acquisition and computer-assisted language learning. He is particularly interested in how uses of computers in second language classrooms affect the way learners learn their second language skills.

**E-mail:** [taichi.yamashita@utoledo.edu](mailto:taichi.yamashita@utoledo.edu)