Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

Using Grammar Checkers in an ESL Context: An Investigation of Automatic Corrective Feedback

Paul John[1] and Nina Woll[2]

Abstract

Our study examines written corrective feedback generated by two online grammar checkers (GCs), Grammarly and Virtual Writing Tutor, and by the grammar checking function of Microsoft Word. We tested the technology on a wide range of grammatical error types from two sources: a set of authentic ESL compositions and a series of simple sentences we generated ourselves. The GCs were evaluated in terms of (1) coverage (number of errors flagged), (2) appropriacy of proposed replacement forms, and (3) rates of "false alarms" (forms mistakenly flagged as incorrect). Although Grammarly and Virtual Writing Tutor outperformed Microsoft Word, neither of the online GCs had high rates of overall coverage (<50%). Consequently, they cannot be relied on to supply comprehensive feedback on student compositions. The finding of higher identification rates for errors from simple rather than authentic sentences reinforces this conclusion. Nonetheless, since few inaccurate replacement forms and false alarms were observed, only rarely is the feedback actively misleading. In addition, the GCs were better at handling some error types than others. Ultimately, we suggest that teachers use GCs with specially designed classroom activities that target selected error types before learners apply the technology to their own writing.

KEYWORDS: GRAMMAR CHECKERS; WRITTEN CORRECTIVE FEEDBACK; FOCUS ON FORM; ENGLISH SECOND LANGUAGE LEARNING.

Affiliations

[1]Université du Québec à Trois-Rivières, Canada.
email: paul.john@uqtr.ca

[2]Université du Québec à Trois-Rivières, Canada.
email: nina.woll@uqtr.ca

CALICO JOURNAL VOL 37.2 2020 169–192
https://doi.org/10.1558/cj.36523

EQUINOX

equinoxonline

1. Introduction

The study reported on here investigates automatic corrective feedback generated by grammar checkers (GCs) in order to assess the appropriacy of using this technology for English second language (ESL) learning. Alongside communicative competence, an important element of learning a second/foreign language (L2) is the development of grammatical accuracy, notably via written corrective feedback. Through such feedback, teachers can effectively incorporate a focus on form into the communicative classroom, thus promoting accuracy and preventing fossilization (Bitchener, 2008; Bitchener & Knoch, 2010; Chandler, 2003; Ellis, 2009; Ellis, Sheen, Murakami, & Takashima, 2008; Ferris, 2006, 2011; Ferris, Liu, Sinha, & Senna, 2013; Shintani, Ellis & Suzuki, 2014; cf. Truscott, 2007). All the same, written feedback has the drawback of being time-consuming and impractical for teachers, especially when applied to multiple drafts. The potential of grammar checking software to reduce teachers' workloads by generating automatic feedback is thus appealing.[1] Moreover, it can be frustrating for teachers to note that learners often ignore the feedback they have so carefully prepared (Guénette, 2007). Part of the problem is that learners receive teacher feedback too long after the composition process. The feedback from GCs, on the other hand, has the advantage of immediacy: it can be accessed right away, while learners are still engaged in the task (for the benefits of immediate feedback, see Jurma & Deidre, 1984; Samuels & Wu, 2003; see also Lavolette, Polio, & Kahng, 2014 for an opposing view). A further advantage is that learners can receive the feedback from GCs independently, beyond the confines of the classroom. The technology can thus be used for autonomous and ubiquitous (anywhere, anytime) learning. In essence, GCs look like an invaluable tool for use in an ESL context.

Nonetheless, important questions remain regarding the accuracy of automatic corrective feedback. Our study assessed the feedback generated by two online GCs (Grammarly and Virtual Writing Tutor) and the grammar checking function associated with Microsoft Word. Our reasoning was that if Word, an omnipresent word-processing program, performs sufficiently well, there is no need to turn to online tools. Among online GCs, Grammarly is probably the most highly recommended, hence its inclusion. Virtual Writing Tutor merits selection as a rare GC designed specifically for ESL learners. Virtual Writing Tutor also has the advantage of providing feedback entirely free of charge, whereas Grammarly at times flags errors as "advanced issues" for which feedback is available only to subscribers.

We assessed the GCs according to: (1) the degree of coverage (how many errors the GCs identify vs. overlook); (2) the appropriacy of proposed

replacement forms; and (3) the extent to which they misidentify correct forms as errors (so-called "false alarms"). The accuracy of feedback is especially important in an L2 context. First, low coverage, with the GC failing to identify certain errors, gives L2 learners a wrong impression of the accuracy of their writing. Next, and more critically, inaccurate suggestions for replacement forms and false alarms can lead learners seriously astray. While L1 writers can rely on their native speaker intuitions to override such inaccurate feedback, L2 writers are more or less at the mercy of the software. Arguably, they are thus more susceptible to being misled, which may result in confusion, frustration and, paradoxically, more error.

We evaluated automatic corrective feedback on a broad range of grammatical error types and compared the GCs' performance on errors from, on the one hand, authentic compositions produced by francophone ESL learners and, on the other, simple sentences instantiating the typical errors of francophones. Our particular concern was with the consequences of our results for ESL teachers and learners. To the extent that GCs generate accurate and comprehensive corrective feedback, particularly on authentic L2 writing, they can relieve teachers of (part of) the feedback burden and promote learner autonomy. The next section establishes a backdrop for our study via an overview of previous research.

2. Background

Previous studies have examined both GCs and the grammar verification function incorporated into automatic writing evaluation (AWE) systems, which are designed to provide not only corrective feedback but also essay scoring and other types of feedback (e.g., on organization and development). Regarding the former, research on GCs has generally adopted a narrow focus, evaluating the software on a restricted set of error types, primarily articles/determiners and prepositions. As shown in Table 1, results are typically presented in terms of precision (percentage of errors flagged that are in fact errors) and recall (percentage of actual errors flagged).

The pervasively low rates of precision across the studies in Table 1, with even the best rate (80%) meaning 20% of errors flagged are false alarms, are troubling news for ESL learners. Since L2 writers lack native speaker intuition, presumably they are ill-equipped to recognize and reject false alarms. Combined with the low (albeit varied: 18–72%) rates for recall, which indicate degree of coverage, the impression is that GCs suffer from limited accuracy in error detection.[2]

Nonetheless, the findings in Table 1 are mainly restricted to preposition and determiner errors, which may not be representative of performance across a

Table 1
Precision and Recall Rates in Previous Studies on GCs

Study & GC	Target	Precision	Recall
Chen (2009) Microsoft ESL Assistant vs. NTNU	articles, verbs, SVA, run-ons/fragments, spelling, compounds	50% vs. 61%	30% vs. 72%
Chodorow, Tetreault, & Han (2007) A maximum entropy classifier	prepositions	80%	30%
De Felice & Pulman (2008) A maximum entropy classifier	prepositions, determiners	66.7%	70%
Gamon et al. (2009) Microsoft ESL Assistant	prepositions, articles, plural nouns	NA	18% (preps) 37% (articles) 27% (plurals)
Yi, Gao, & Dolan (2008) A web-frequency algorithm	prepositions, determiners	62%	41%

broader set of error types. As De Felice and Pulman (2008) note, preposition and determiner choice is contextually determined, so these types of errors are particularly challenging. Chen (2009) covers a wider set of errors, but this study mixes grammatical errors (articles, verbs, subject–verb agreement) and strictly writing issues (spelling, run-ons/fragments), which makes it harder to assess the GCs' ability to provide grammatical feedback. In our view, to establish the appropriacy of using GCs in an ESL context, especially from a writing-to-learn perspective, an investigation of a wide range of grammatical errors is required.

As shown in Table 2, some research on the grammar function of AWE systems encompasses a broader set of error types, although, as with Chen (2009) above, there is a tendency to combine grammatical and writing errors. The two sets of results for Han et al. (2006) in Table 2 illustrate well the tendency for performance to involve a trade-off between precision and recall: higher rates of precision are often achieved at the expense of recall or vice versa. The first set of results (52% precision and 80% recall) refer to Criterion's overall performance, whereas the second set (90% and 40%) refer to the results when the system was set to ignore low confidence cases. By eliminating these uncertain errors, precision rates increase, but recall rates plummet. In addition, the findings for AWE systems reveal ranges in rates of precision (49–90%) and recall

Table 2
Precision and Recall Rates in Previous Studies on AWE Systems

Study & AWE system	Target	Precision	Recall
Dikli & Bleyle (2014) *Criterion*	many grammatical/ writing errors	NA	16.5% (94 of 570 errors)
Feng, Saricaoglu, & Chukharev-Hudilainen (2016) *CyWrite*	articles, SVA, run-ons, quantifiers	56–75%*	30–76%*
Han, Chodorow, & Leacock (2006) *Criterion*	articles	52% 90%**	80% 40%**
Hoang & Kunnan (2016) *MY Access*	many grammatical/ writing errors	73%	39.6%
Lavolette, Polio, & Kahng (2014) *Criterion*	many grammatical/ writing errors	89%***	54%
Liu & Kunnan (2016) *WriteToLearn*	many grammatical/ writing errors	49%	18.7%
Tetreault & Chodorow (2008) *Criterion*	prepositions	84%	19%

* across the four error types
** with the system set to ignore low confidence cases
*** 14% of these were correctly identified but miscoded in terms of the error category

(16.5–76%) similar to those for GCs, so the same concerns apply concerning the appropriacy of using these tools to provide feedback in an ESL setting. One further drawback to AWE systems is that they can be prohibitively expensive and inaccessible beyond classroom settings, meaning they lack the appeal of free (Virtual Writing Tutor) or partially free (Grammarly) online GCs.

Based on the above review, we consider that the feasibility of using feedback from online GCs to promote accuracy in L2 English merits further investigation. Previous studies have tended to adopt either a narrow focus or else to incorporate strictly writing issues such as spelling and punctuation, which are mechanical rather than grammatical in nature. Our approach is to employ a wide scope on grammatical errors, eschewing writing errors and focusing on the potential for written corrective feedback to promote wider L2 learning (i.e., as an opportunity for writing-to-learn). By cutting a wide swathe, we can determine which error types GCs handle more successfully. Such information on error-dependent reliability is invaluable for L2 teachers and learners alike.

In addition, we compare the performance of Grammarly and Virtual Writing Tutor with that of the ubiquitous word processing software, Microsoft Word. As Figures 1 and 2 show, corrective feedback from the GCs, as with Microsoft Word, is direct: the error in the text is underlined, a replacement form is proposed, and a metalinguistic explanation is provided.

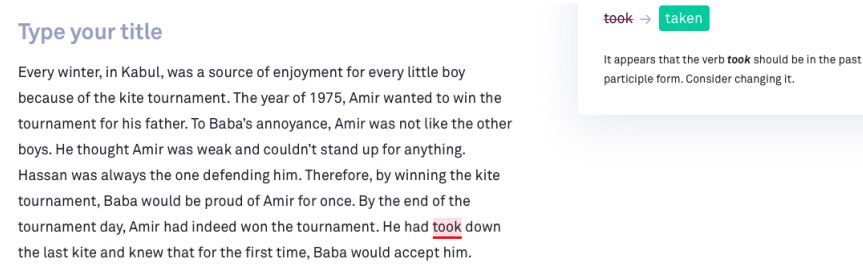

Figure 1. Screenshot of feedback in Grammarly.

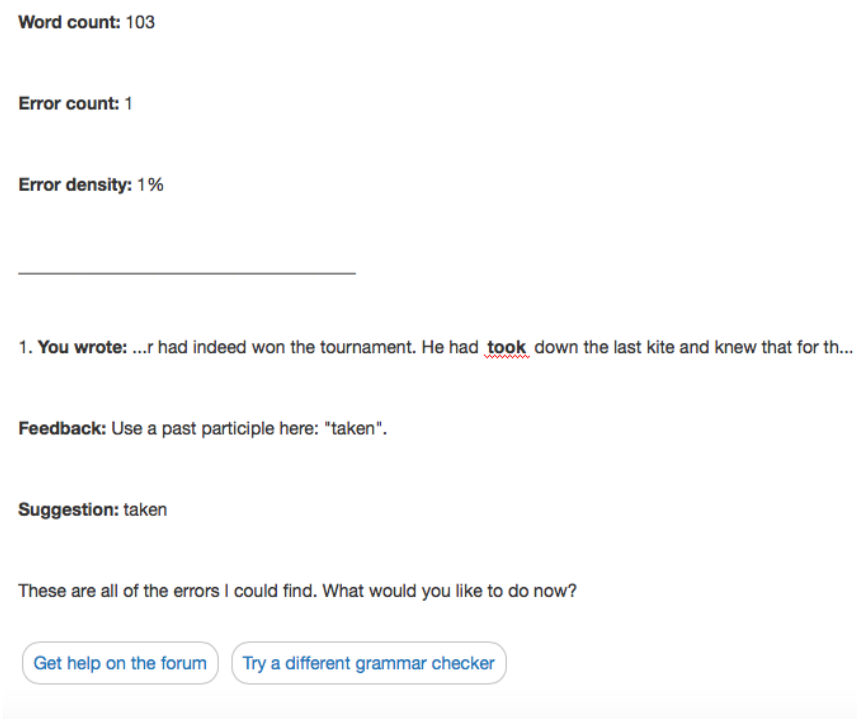

Figure 2. Screenshot of feedback in Virtual Writing Tutor.

In brief, our study addresses the following research questions:

1. In terms of error coverage, replacement forms, and false alarms, how accurate is the corrective feedback from Grammarly, Virtual Writing Tutor and Microsoft Word?
2. Does the technology perform better on certain grammatical errors than others?
3. Is there a difference in how the technology handles authentic L2 writing errors versus errors appearing in specially designed simple sentences?

3. Methodology

To assess the accuracy of automatic corrective feedback, our study examined how GCs handle grammatical errors in: (1) authentic compositions produced by francophone ESL learners; and (2) a set of simple sentences we composed ourselves containing errors typical of francophone ESL learners.

The compositions were 50 handwritten essays produced by 28 adult francophone learners of English (n_{male} = 10, n_{female} = 18; age 21–36). The participants were in the second year of a four-year TESL program at a university in Quebec (Canada) situated in a French-speaking region. When they arrive at the university, these undergraduates have typically been learning English in a classroom setting for 10 years (i.e., from the age of nine), so despite some variation in proficiency, they are best characterized as advanced learners. The compositions were produced under exam conditions (two essays in three hours) as responses to questions on novels that the students had read for a course in their teacher-training program. In composing the essays, students had access to their notes, dictionaries, and grammar references. In using errors from the essays to assess the GCs' performance, our aim was to establish how well the technology handles actual, sometimes opaque and convoluted, L2 written output.

Unlike the compositions, the 129 sentences developed for the study were not authentic samples of L2 writing. Instead, they were generated by the researchers to instantiate errors that francophone ESL learners typically make. To construct the sentences, we relied partly on personal experience, drawing on prior knowledge of learner errors, and we also consulted records we have compiled listing errors our students made in various contexts (exams, essays, and other assignments) over a number of years. These records were originally compiled to give whole class feedback to our students on representative errors from their own writing. To ensure we had not overlooked any common errors, we further consulted with two other applied linguistics professors (Mariane Gazaille

at Université du Québec à Trois-Rivières and Walcir Cardoso at Concordia University), both of whom are highly familiar with francophone ESL learner output. The resulting sentences have the advantage of covering a broader range of errors than might have occurred in the compositions: we were able to include errors that advanced learners might no longer make or that were missing from the compositions due to incidental gaps. By embedding the errors in quite simple sentences, the primary aim was to assess the GCs' performance under optimal conditions. That is, if a GC fails to identify an error in a composition, this failure could be due to factors such as syntactic complexity or the presence of multiple errors in a sentence rather than an absolute inability to handle the grammatical issue in question. Indeed, while only 52 of the 358 sentences from the compositions were syntactically simple (i.e., containing a single clause), fully 96 of the 129 sentences we generated were simple. Typically, these were short sentences, and in all cases, they contained only a single error. In short, the specially designed sentences constitute a useful counterpoint to the authentic errors. Results from the two sets of data provide a multifaceted portrait of the GCs' abilities.

3.1 Compositions

As a first step, the 50 handwritten compositions were transcribed into a Word document, with the grammar checking function turned off. They were then coded by the two researchers for a set of grammatical categories established beforehand and modified subsequently as needed (a system of annotation similar to that in Granger, 2003). Both researchers analyzed the first five essays. We then compared our analyses in order to calibrate our classification of errors. The rest of the essays were divided up between us, and we consulted each other only when unsure of how to categorize a particular error token. If the nature of an error remained impossible to pinpoint (e.g., because the intended meaning was opaque), we excluded the token from the analysis. To maintain a grammatical focus, we overlooked not only spelling and punctuation errors but also errors of usage such as incorrect word choice. As a final step, the first author went over all of the identified errors to ensure consistency in classification.

A total of 358 errors were found in the target categories among the 23,108 words comprising the 50 authentic compositions (1 error for every 65 words). The error distribution across the various categories is presented in Table 3. The numbers in the left-hand column show the total errors in a given category, and the numbers in the columns on the right represent the total errors within each subcategory.

Table 3
Error Count in Compositions (Researcher Coding)

Category	**Subcategory**			
VERBS	tense-aspect	verb form	subj-V agreement	tense shift
164	42	58	18	46
NOUNS	plural	possessive	pronoun	
57	29	18	10	
PREPOSITIONS	wrong	missing	unnecessary	
71	45	16	10	
WORDS	word order	word form		
34	14	20		
MISC.	determiner	relative clause		
32	19	13		
GRAND TOTAL				
358				

Many of the categories in the coding system should be quite transparent (e.g., an overgeneralization such as "teached" is readily classified as an error in verb form), but others may be more elusive. By "tense shift", we mean inconsistent use of verb tense at the discourse level: primarily, this involves shifts between past and present in contexts where either tense is acceptable, but where, for coherence, one tense should be maintained throughout. In the category of plural nouns appear cases where the learner fails to pluralize a noun or employs a non-count noun as count. The latter error is common among francophones since a number of English non-count nouns are count in French (e.g., "research" vs. "les recherches" or "homework" vs. "les devoirs"). Errors with the possessive involve inappropriate use of either apostrophe + *s* (generally reserved for animate possessors) or the periphrastic possessive with "of" (e.g., "the dog of Peter"—unusual with animate possessors). Pronoun errors concern problems of anaphora (i.e., incorrect reference, including use of the wrong relative pronoun in, say, "the book *who* is on the table"). The category "Words" covers questions of syntax (word order) and morphology (word form). Under "word order", we would include the misapplication of subject-auxiliary inversion or *do*-insertion

in embedded questions. Using "cowardness" instead of "cowardice" is an example of a "word form" error. The category "relative clauses" refers to the misuse of commas to distinguish between restrictive and non-restrictive relative clauses. The decision was made to include this exceptional punctuation issue since the ability to distinguish between restrictive and non-restrictive relative clauses has wider implications, notably for relative pronoun choice ("that" can only be used with restrictive relative clauses). Once the essays were thus coded and the consistency in classification verified by the first author, we ran all 358 errors through the GCs.

3.2 Sentences

To complement the authentic errors from student compositions, we composed a series of 129 simple sentences containing errors that francophone ESL learners typically make in the grammatical categories identified previously. The principal purpose was to provide an optimal context for error identification and correction. In addition, we were able to include errors that were missing from the compositions (due to a task effect or to our learners being beyond that error stage). For example, francophones have difficulty distinguishing between "since" and "for" used with time expressions that indicate either a starting point ("since Tuesday") or duration ("for three days"). French employs "depuis" in both contexts ("depuis mardi", "depuis trois jours"), so francophones tend to use "since" even for expressions of duration. This characteristic error was not found in the compositions, but we incorporated it into one of the sentences. In brief, the simple sentences constitute an optimal and broad context for testing the GCs.

4. Results

The results are presented first in terms of the GCs' coverage for errors from the compositions and sentences respectively, and then in terms of the number of inaccurate replacement forms and false alarms.

4.1 Coverage (Compositions)

Table 4 shows how the three GCs performed with respect to the set of composition errors. The number of errors in each subcategory is given in parentheses in the left-hand column. The three right-hand columns indicate the number of errors flagged by each GC and the percentage of the total this represents.

Table 4
Error Identification by Grammar Checkers (Compositions)

		Microsoft Word	Grammarly	Virtual Writing Tutor
Verbs	Tense-aspect (42)	3 (7.1%)	4 (9.5%)*	6 (14.3%)
	Verb form (58)	3 (5.2 %)	25 (43.1%)	25 (43.1%)
	Subj-V agreement (18)	0 (0%)	15 (83.3%)	6 (33.3%)
	Tense shift (46)	0 (0%)	0 (0%)	0 (0%)
	Total: 164	**6 (3.7%)**	**44 (26.8%)**	**37 (22.6%)**
Nouns	Plural (29)	3 (10.3%)	15 (51.7%)	10 (34.5%)
	Possessive (18)	0 (0%)	5 (27.8%)	4 (22.2%)
	Pronoun (10)	0 (0%)	1 (1%)	0 (0%)
	Total: 57	**3 (5.3%)**	**21 (36.8%)**	**14 (24.6%)**
Preps	Wrong prep (45)	1 (2.2%)	9 (20%)	1 (2.2%)
	Missing prep (16)	0 (0%)	3 (18.8%)	0 (0%)
	Unnecessary prep (10)	0 (0%)	3 (30%)	2 (20%)
	Total: 71	**1 (1.4%)**	**15 (21.1%)**	**3 (4.2%)**
Words	Word order (14)	0 (0%)	1 (7.1%)	3 (21.4%)
	Word form (20)	6 (30%)	12 (60%)	11 (55%)
	Total: 34	**6 (17.6%)**	**13 (38.2%)**	**14 (41.2%)**
Misc.	Determiner (19)	0 (0%)	5 (26.3%)	3 (15.8%)
	Relative clause (13)	2 (15.4%)	8 (61.5%)**	0 (0%)
	Total: 32	**2 (6.3%)**	**13 (40.6%)**	**3 (9.4%)**
	Grand total: 358	**18 (5.0%)**	**106 (29.6%)**	**71 (19.8%)**

* 2 of the 4 verb tense-aspect errors flagged by Grammarly were indicated to be "advanced issues" (i.e., accessible only to fee-paying subscribers)
** these 8 relative clause errors were flagged as "advanced issues"

The grand totals in Table 4 create an immediate impression that the GCs provide poor overall coverage and that, compared with Grammarly (29.6%) and Virtual Writing Tutor (19.8%), Microsoft Word is particularly ill-equipped to identify errors (5.0%). Despite the poor overall coverage, however, there are some grammatical subcategories in which Grammarly, and at times Virtual Writing Tutor, perform better. Grammarly is relatively strong on "verb form"

errors (43.1%) and certainly on "subject-verb agreement" errors (83.3%). It likewise performs well on errors involving plural nouns (51.7%), word forms (60%), and relative clauses (61.5%). Conversely, Grammarly, like the other GCs, shows especially poor coverage on "tense-aspect" (9.5%), "tense shift" (0%), "pronoun" (1%), and "word order" (7.1%) errors.[3] Exceptionally, Virtual Writing Tutor outperforms Grammarly on "tense-aspect" (14.3%) and "word order" (21.4%) errors.

4.2 Coverage (Simple Sentences)

Table 5 provides the distribution of sentence errors identified by the different GCs across the various categories and subcategories. The figures in parentheses in the left-hand column again indicate the number of errors run through the GCs, while the figures in the right-hand columns show the number of errors flagged and the percentage of the total this represents.

Table 5
Error Identification by Grammar Checkers (Simple Sentences)

		Microsoft Word	Grammarly	Virtual Writing Tutor
Verbs	Tense-aspect (9)	1 (11.1%)	4 (44.4%)	0 (0%)
	Verb form (13)	2 (15.4%)	8 (61.5%)	8 (61.5%)
	Subj-V agreement (6)	0 (0%)	6 (100%)	6 (100%)
	Tense shift (2)	0 (0%)	0 (0%)	0 (0%)
	Total: 30	**3 (10%)**	**18 (60%)**	**14 (46.7%)**
Nouns	**Plural (20)**	4 (20%)	11 (55%)	11 (55%)
	Possessive (4)	0 (0%)	0 (0%)	0 (0%)
	Pronoun (5)	0 (0%)	2 (40%)	0 (0%)
	Total: 29	**4 (13.8%)**	**13 (44.8%)**	**11 (37.9%)**
Preps	Wrong prep (10)	0 (0%)	8 (80%)	8 (80%)
	Missing prep (4)	0 (0%)	2 (50%)	2 (50%)
	Unnecessary prep (7)	0 (0%)	3 (42.9%)	2 (28.6%)
	Total: 21	**0 (0%)**	**13 (61.9%)**	**12 (57.1%)**
Words	Word order (18)	3 (16.7%)	7 (38.9%)	3 (38.9%)
	Word form (10)	6 (60%)	7 (70%)	7 (70%)
	Total: 28	**9 (32.1%)**	**14 (50%)**	**10 (35.7%)**

Misc.	Determiner (13)	1 (7.7%)	4 (30.8%)	4 (30.8%)
	Relative clause (8)	2 (25%)	1 (12.5%)	0 (0%)
	Total: 21	**3 (14.3%)**	**5 (23.8%)**	**4 (19.0%)**
	Grand total: 129	**19 (14.7%)**	**63* (48.8%)**	**51 (39.5%)**

* 10 errors were flagged as "advanced issues" accessible only to fee-paying subscribers

Total error detection rates are higher for the errors in simple sentences than for those in student compositions. The differences in error detection rates are considerable: 29.6% vs. 48.8% (Grammarly), 19.8% vs. 39.5% (Virtual Writing Tutor) and 5.0% vs. 14.7% (Microsoft Word). Still, with even the best performance catching just under half of the errors, the GCs continue to exhibit limited overall coverage.

As with the composition errors, the GCs performed better in some categories than others. Both Grammarly and Virtual Writing Tutor were particularly strong at identifying verb form (61.5%), subject-verb agreement (100%), plural noun (55%), and word form (70%) errors in the simple sentences. Unusually, while Grammarly performed well on relative clause errors in the compositions (61.5%), it did a poor job of identifying these errors in the simple sentences (12.5%). In the preposition category, however, both Grammarly and Virtual Writing Tutor performed well with the simple sentences (especially wrong prepositions: 80%), which was not the case for errors from the compositions. In sum, more overall errors were detected in the simple sentences than the compositions, with strong performances generally being observed in the same error categories. It seems that these are the categories that GCs are truly well-equipped to handle.

4.3 Replacement Forms

Aside from error coverage, we also analyzed the GCs for accuracy of replacement forms and number of false alarms. Although the issue of appropriate replacement forms is clearly pertinent to the use of GCs by ESL learners, this aspect of GC performance has not been addressed in previous studies, which have limited their analysis to questions of precision and recall. In Table 6, the results for inaccurate replacement forms are presented as fractions (number of inaccurate forms per errors flagged) and percentages.

As seen from the percentages in Table 6, all three GCs show greater likelihood of inaccurate replacements for errors in the compositions than in the simple sentences. While these inaccuracies are relatively frequent from Microsoft Word, they are comparatively rare from either of the online GCs (especially Grammarly).

Table 6
Number and Percentage of Inaccurate Replacement Forms

	Microsoft Word	Grammarly	Virtual Writing Tutor
Compositions	5/18 (27.8%)	5/106 (4.7%)	9/71 (12.7%)
Simple sentences	2/19 (10.5%)	1/63 (1.6%)	1/51 (2.0%)

To give an idea of the nature of inaccurate replacements, Microsoft Word misinterpreted an *-ed* overgeneralization error from the composition data ("Dunstan **seeked** for help and Mary arrived") as a spelling rather than a verb form error. Consequently, the system proposed "seeded" and "sleeked" as possible corrections rather than "sought". Grammarly inaccurately suggested "personalities" rather than "people" as a replacement for "person" in "Frank Bascombe in *The Sportswriter* by Richard Ford had to mourn two different **person** in the book". While the suggested replacement forms are not ungrammatical, they are nonetheless inappropriate in the context.

The miscues on the sentence data occurred with both verb form and word form errors. Among the verb form errors it flagged, Microsoft Word proposed one inaccurate replacement form, and Virtual Writing Tutor proposed two. Specifically, Microsoft Word had difficulty with "She **teached** in Japan last year", failing to recognize that an irregular past tense form is required; instead it suggested *teaches, teacher,* and *reached* as possible corrections. Among the word form errors, all three GCs had difficulty with "You drove **quicklier** than I did": Microsoft Word and Grammarly both suggested replacing "quicklier" with just "quickly", rather than the intended comparative "more quickly"; Virtual Writing Tutor tagged the form as a "possible spelling mistake", without proposing a replacement. In response to the overgeneralized nominalizing suffix *-ness* in "There is a lot of **obeseness** in North America", Microsoft Word suggested "baseness, bossiness, obscenest" rather than the intended "obesity", while Virtual Writing Tutor identified it as either a "possible spelling mistake" (with no replacement form) or a count noun which should be pluralized ("a lot of obesenesses"), which completely misses the mark; Grammarly failed to flag the error.

The GCs also occasionally provided partially inaccurate forms, where the correct form was given only after one or more inaccurate forms, whether on the compositions (3 partially inaccurate replacements for Microsoft Word, 5 for Grammarly, and 1 for Virtual Writing Tutor) or on the simple sentences (1 for Microsoft Word and 2 for Virtual Writing Tutor). For example, Virtual Writing Tutor had two miscues involving passive voice. For the sentences "We were **expose** to dangerous levels of radiation" and "John is **motivate** to

learn Russian", where the past participle is required, the GC suggested "We expose/were exposing/were exposed" and "John motivate [without -*s*]/is motivating/is motivated" as possible corrections (i.e., in both cases, only the third replacement form is accurate). Though such complete or partial inaccuracies are relatively rare, it is important to note that the GCs' feedback on detected errors is not always reliable.

4.4 False Alarms

The data were also analyzed for false alarms, where a GC mistakenly flags a correct form as an error. The rates in Table 6 were arrived at by running the 129 sentences and the 50 compositions in their entirety through the GCs.

Table 7
Number of False Alarms

	Microsoft Word	Grammarly	Virtual Writing Tutor
Compositions	13	4	30
Simple sentences	0	0	0

On the compositions, Grammarly does the best job of avoiding false alarms. Virtual Writing Tutor, with 30 false alarms, is the poorest performer, although these are spread out over 50 compositions or 23,108 words (1 false alarm per 770 words), so the performance is not as critical as it might initially seem. The relatively low rate of 13 false alarms for Microsoft Word is probably a function of its low error identification rate leading to a lower likelihood of flagging a correct form as an error. The absence of false alarms in the simple sentences is probably partly due to lack of opportunity (the sentences contained only 1055 words). Another factor could be the lower complexity of the sentences, making it less likely the GC will perform an inaccurate parse.

4.5 Summary of Results

For our first research question, our study sought to establish how accurate the corrective feedback from Grammarly, Virtual Writing Tutor, and Microsoft Word is in terms of coverage, replacement forms, and false alarms. In terms of coverage, the two online GCs considerably outperformed Microsoft Word, and Grammarly generally outperformed Virtual Writing Tutor. All the same, none of the GCs were strong at providing comprehensive coverage: overall scores for all three GCs, whether on the compositions or simple sentences, never exceeded 50%. In terms of inaccurate replacement forms and false alarms,

Grammarly was best at avoiding these pitfalls, but all of the three tools had relatively low rates of actively inaccurate feedback.

With respect to our second research question, our results show considerable variation in coverage depending on error type. The grammatical errors that GCs are best at handling involve verb form, subject-verb agreement, plural noun, and word form errors. Relative clause errors were flagged more often in the composition than the simple sentence data; wrong preposition errors were flagged at a high rate in the simple sentence data only.

Regarding our third research question, the GCs showed considerable variation in how they handle authentic L2 writing errors versus errors appearing in specially designed simple sentences. Rates of coverage were much higher for errors in the simple sentences. In other words, actual L2 output appears to hamper the GCs' ability to detect errors. The difficulty posed by authentic writing is reflected also in different rates of inaccurate replacement forms and false alarms (higher rates in compositions than simple sentences), although the pattern observed for false alarms can also be attributed to differences in opportunity.

5. Discussion

Unlike previous research on GCs and AWE systems (see the studies in Tables 1 and 2 in the Background section), the current study examined the accuracy of automatic corrective feedback on a wide range of grammatical error types, eschewing purely writing issues such as spelling, punctuation and run-ons/fragments, and usage issues such as word choice. The aim was to compare the reliability of different GCs (Grammarly, Virtual Writing Tutor, and Microsoft Word) in terms of coverage, appropriacy of replacement forms, and avoidance of false alarms. The degree of reliability, particularly on actual L2 compositions, determines the extent to which GCs can relieve teachers of the corrective feedback burden and afford learners greater autonomy.

The GCs we examined offer poor overall coverage (the best being Grammarly at 48.8% on the simple sentences), which is in line with what previous research has observed. Other studies showing higher coverage (e.g., De Felice & Pulman, 2008, with 70% coverage on preposition and determiner errors, and Han, Chodorow, & Leacock, 2006, with 80% coverage on article errors) also reveal high rates of false alarms (66.7% and 52% precision respectively, meaning 33.3% and 48% of errors flagged were false alarms), a pattern indicative of the trade-off between coverage and false alarms. The three GCs in our study, particularly Grammarly, showed low rates of false alarms on the compositions and none on the simple sentences; even the worst offender, Virtual Writing Tutor, had only 30 false alarms across the 50 compositions. In terms

of replacement forms, which previous studies have tended not to report on, we found very low rates of inaccurate replacements for the online GCs (especially Grammarly) and higher rates for Microsoft Word. In sum, in terms of accuracy of corrective feedback, the two online GCs outperform Microsoft Word, and Grammarly outperforms Virtual Writing Tutor (although only paid subscribers have access to full coverage).

While overall coverage rates were unimpressive, the performance of the GCs varies widely according to error type. Grammarly (and to a degree Virtual Writing Tutor) is better at detecting errors in verb forms, subject-verb agreement, plural nouns, and word forms, as well as relative clauses (composition data only) and wrong prepositions (sentence data only). In addition, the GCs are better at detecting errors in specially designed simple sentences than in sentences extracted from authentic compositions. This finding points to the difficulty of identifying errors in the often opaque context of actual L2 writing. Our findings have implications for the use of GCs in an ESL context, which we discuss next.

5.1 Pedagogical Implications

For the purposes of ESL learning, it is preferable that GCs be conservative, sacrificing coverage to avoid actively inaccurate corrective feedback. While it is unfortunate when a GC overlooks an error, from an ESL learner's perspective it is far more confusing for a GC to indicate erroneously that something is wrong or to propose an inaccurate replacement form. In this sense, the feedback from Grammarly and Virtual Writing Tutor is well suited for use in an ESL context. Microsoft Word's coverage, however, is so low that it is of little use. Indeed, one of the implications of our findings is that we would recommend that learners not rely on the feedback from Microsoft Word; to benefit from wider coverage, it is worth turning to an online GC such as Grammarly or Virtual Writing Tutor. Whatever system they use, learners still need to realize that it will not catch everything, so using a GC does not obviate scrutiny of their own writing for errors. From a teacher's perspective, the low overall coverage, particularly on authentic errors, means GCs cannot be relied upon to provide comprehensive corrective feedback on L2 compositions. Nonetheless, GCs can be used in circumscribed fashion to target the specific error types they perform well on. In this way, GCs can afford learners with at least partial autonomy and relieve teachers of part of the corrective feedback burden, acting as a precursor to teacher feedback.

Because the feedback from GCs is not entirely reliable, it would be best for teachers to familiarize learners with the limitations of the technology. To do so, teachers can design special focus-on-form activities using GCs to target

particular grammar points (e.g., error types the GCs are good at detecting). For example, ESL textbooks typically contain readings accompanied by some form of meaning-focused activity. We suggest that teachers subsequently modify the passage, inserting a particular type of error. The resulting activity involves students first analyzing the text for errors (i.e., attempting to identify, correct, and explain the errors themselves) and then running the text through a GC to see whether it has the same answers as they do. The teacher should check beforehand whether the GC detects all of the inserted errors and whether it provides any inaccurate feedback (incorrect replacement forms or false alarms). Students can then be forewarned of the GC's missteps, and part of the activity is to identify where the GC trips up (see Activity 1 in the Appendix for an example targeting subject-verb agreement). The idea is that such an activity performs a dual function: (1) it encourages learners to look critically at writing to try to detect errors, which will serve them when they themselves are composing; and (2) it gets them used to using GCs, sensitizing them to the sometimes fallible feedback provided.

Another type of focus-on-form activity teachers can develop involves sets of sentences with particular error types, again emphasizing categories of error on which the GCs perform well. Activity 2 in the Appendix provides an example targeting wrong, missing, or unnecessary prepositions. The sentences were developed with our francophone students in mind, containing typical errors based on L1 transfer. For example, the sentences with a wrong preposition, "I like participating **to** this activity" and "Are you satisfied **of** the service at the hotel?", employ prepositions used in their French equivalents, "J'aime participer **à** cette activité" and "Êtes-vous satisfait **du** service à l'hôtel?". The sentences with a missing preposition, "I am waiting _ the bus" and "I enjoy listening _ the radio", lack prepositions in their French counterparts, "J'attends l'autobus" and "J'aime écouter la radio." Finally, the one sentence with an unnecessary preposition, "The teacher asked **to** the students to be quiet", corresponds to the French sentence, "L'enseignant a demandé **aux** étudiants de se calmer." ESL teachers can develop similar activities to target the characteristic errors of their own students.

Once learners are familiarized with GCs, including the fallibility of their feedback, they can use the technology on their own writing, initially in more circumscribed fashion. For example, in a two-step process, teachers can first get learners to scrutinize their own compositions for a particular type of error (e.g., subject-verb agreement). Next, learners run their writing through a GC to see whether the tool flags any errors in the grammatical category of interest. The idea is that learners can overlook unrelated feedback, focusing purely on improving accuracy in the chosen category. Eventually, learners can pay attention to all the feedback provided. Relatedly, they can keep logs of the

type and number of errors a GC flags in their compositions over the course of a semester. This way they can keep track of their own progression, using the GC as a means of self-evaluation.

5.2 Limitations

Among the limitations of our study is the fact that our assessment is restricted to three resources, so it does not offer a comprehensive account of what is available. We focused our analysis on two leading, readily available GCs, one designed for L1 writers (Grammarly) and the other for L2 writers (Virtual Writing Tutor), alongside the grammar checking function in Microsoft Word. While we assumed that ESL users would prefer services that require little or no payment, the possibility remains that other, potentially costly, GCs might perform better. This scenario seems unlikely, however, given that previous studies show no indication of the existence of high-end supercheckers.

Another shortcoming, due to the nature of the study, is that the analysis employs only descriptive statistics. This means that, particularly given the sometimes low number of error tokens tested—our study does not have the sort of data set found in corpus-based approaches such as Granger (2003)—we cannot be sure of the statistical significance of our results. Nonetheless, the portrait that emerges is clear and consistent across the findings; an analysis employing inferential statistics or a larger corpus would be unlikely to change the global message.

6. Conclusion

We have proposed two avenues for incorporating GCs into the ESL classroom: for provision of automatic corrective feedback on student compositions or for use with special focus-on-form activities. The former potential use is what first attracted our attention: if GCs provide accurate, comprehensive corrective feedback on student writing, this could relieve teachers of a time-consuming task and provide learners with greater autonomy. Our results show the feedback from GCs to be limited in coverage such that GCs cannot entirely replace human feedback. Nonetheless, they can be used effectively to target particular types of errors in student writing, focusing on grammatical categories in which the systems are strongest. This application requires some kind of training in the use of GCs, for example via specially designed activities that target specific error types. Such activities will familiarize learners with the strengths and weaknesses of GCs and ultimately train them to bring a critical eye to their own writing.

There is room for more research on GCs in the future. Notably, we have yet to assess the accuracy of their metalinguistic feedback. Do certain GCs offer better metalinguistic explanations than others? What kinds of terminology should learners be familiar with to benefit from these explanations? Our impression is that Grammarly and Virtual Writing Tutor both provide quite detailed metalinguistic information, whereas any explanations provided by Microsoft Word tend to be cursory. Clearly, however, a systematic study should examine the relative strengths of GCs in this area. In addition, the evaluation of GCs is an ongoing process, which will need to be updated as new generations or new iterations of the resources emerge. As it stands, while GCs provide only a partial solution to the problem of written corrective feedback, they nonetheless constitute a useful tool for integrating a focus on form into the ESL classroom.

Notes

1. As an anonymous reviewer pointed out, this observation takes for granted that teachers actually devote the time to supplying corrective feedback, which may not be the case. The point would benefit from empirical verification.
2. Rather than "recall" (percentage of actual errors flagged) and "precision" (percentage of errors flagged that are actual errors), we use the terms "coverage" and "false alarms", partly because their meaning is more transparent. In addition, measures of precision can be misleading, since, on the same piece of writing, 50% precision would be reported for a GC that flags 48 errors of which 24 are actual errors and for another GC that flags 4 errors of which 2 are actual errors. To assess a GC's performance, we thus find it more pertinent to report on the number of false alarms generated.
3. The finding that none of the tools caught errors in the "tense shift" category is not surprising, given that the technology is not designed to detect errors beyond the sentence boundary (see the explanation offered by Nicholas Walker, the developer of Virtual Writing Tutor: https://virtual-writing-tutor.blogspot.com/2013/01/grammarcheckerfail.html). We are grateful to an anonymous reviewer for pointing this out to us.

Acknowledgments

We would like to acknowledge the invaluable input at various stages of the research project from our colleagues, Mariane Gazaille and Walcir Cardoso, and of our research assistant, Michel Monier. We are responsible for any enduring errors and infelicities.

About the Authors

Paul John is an Associate Professor in the Department of Modern Languages at the University of Quebec in Trois-Rivières (Canada). While his main research focuses on L2 phonological acquisition, including the use of neuroimaging to investigate L2 phonological perception, he is also interested in computer-assisted language

learning. His recent projects in this field have explored the use of text-to-speech and grammar-checking technology for L2 learning.

Nina Woll is an Associate Professor in the Department of Modern Languages at the University of Quebec in Trois-Rivières (Canada). Her research interests include psycholinguistic processes in the acquisition of additional languages, specifically with regard to the development of metalinguistic and crosslinguistic awareness in instructed settings.

References

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*, 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing, 19*, 207–217. https://doi.org/10.1016/j.jslw.2010.10.002

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chen, H.-J. H. (2009). Evaluating two web-based grammar checkers – Microsoft ESL Assistant and NTNU statistical grammar checker. *Computational Linguistics and Chinese Language Processing, 14*(2), 161–180.

Chodorow, M., Tetreault, J. R., & Han, N.-R. (2007). Detection of Grammatical Errors Involving Prepositions. In F. Costello, J. Kelleher and M. Volk (Eds.), *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions* (pp. 25–30). https://doi.org/10.3115/1654629.1654635

De Felice, R., & Pulman, S. G. (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In D. Scott and H. Uszkoreit (Eds.), *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)* (pp. 169–176). https://doi.org/10.3115/1599081.1599103

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

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal, 1*, 3–18. https://doi.org/10.5070/L2.V1I1.9054

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*, 353–371. https://doi.org/10.1016/j.system.2008.02.001

Feng, H.-H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016). Automated error detection for developing grammar proficiency of ESL learners. *CALICO Journal, 33*(1), 49–70. https://doi.org/10.1558/cj.v33i1.26507

Ferris, D. (2006). Does error feedback help student writers? New evidence on the short- and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104). Cambridge, England: Cambridge University Press. https://doi.org/10.1017/CBO9781139524742.007

Ferris, D. (2011). *Treatment of error in second language writing* (2nd ed.). Ann Arbor, MI: University of Michigan Press. https://doi.org/10.3998/mpub.2173290

Ferris, D., Liu, H., Sinha, A., & Senna, M. (2013). Written corrective feedback for individual L2 writers. *Journal of Second Language Writing, 22*, 307–329. https://doi.org/10.1016/j.jslw.2012.09.009

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. https://doi.org/10.1558/cj.v26i3.491-511

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–480.

Guénette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing, 16*, 40–53. https://doi.org/10.1016/j.jslw.2007.01.001

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

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly, 13*(4), 359–376. https://doi.org/10.1080/15434303.2016.1230121

Jurma, W. E., & Deidre, L. F. (1984). Effects of immediate instructor feedback on group discussion participants. *Central States Speech Journal, 35*(3), 178–86. https://doi.org/10.1080/10510978409368186

Lavolette, E., Polio, C., & Kahng, J. (2014). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology, 19*(2), 50–68.

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of *WriteToLearn*. *CALICO Journal, 33*(1), 71–91.

Samuels, S. J., & Wu, Y. (2003). The effects of immediate feedback on reading achievement. Technical report. Minneapolis: University of Minnesota. Retrieved from http://www.epsteineducation.com/home/articles/file/research/immediate_feedback.pdf

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning, 64*, 103–131. https://doi.org/10.1111/lang.12029

Tetreault, J. R., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. In D. Scott and H. Uszkoreit (Eds.), *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)* (pp. 865–872). https://doi.org/10.3115/1599081.1599190

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing, 16*(4), 255–272. https://doi.org/10.1016/j.jslw.2007.06.003

Yi, X., Gao, J., & Dolan, W. B. (2008). A web-based English proofing system for English as a second language users. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)* (pp. 619–624).

Appendix

Activity 1 (Subject-Verb Agreement)

Adapted from: Sarchuk, N., and Payne, D. (2001). *Bookmark. Grammar.* Anjou, QC: Les Éditions CEC.

Instructions: The following paragraph contains a number of verbs in the Simple Present, but there are some errors in subject-verb agreement (twelve in all). Can you identify and correct the twelve errors? Check your answers by copying the paragraph into Grammarly. The grammar checker flags and corrects ten of the twelve errors. Which two errors does it miss? (NB—To facilitate identification, the errors are presented in bold here.)

Let me introduce my good friend Brian to you. Brian is 20 years old, and he **study** at the university. At this point, he doesn't **knows** exactly what career he **want**, but he is interested in work that **involve** environmental protection, especially in developing countries. His concerns **is** that, in the name of progress, companies **doesn't** think about the future. They **forgets** that natural resources won't last forever. Brian **want** to focus on a slower, more sustainable development over the long term. Brian **like** to travel, meet new people, learn new languages, and **plays** sports such as soccer and hockey. He **don't** enjoy staying in the same place for a long time, at least, not for now. He is adventurous and ready for new experiences. There **are** a lot more information I can tell you about Brian, but that will come later.

Activity 2 (Prepositions)

Instructions: Each of the following sentences contains a preposition error. In some cases, the wrong preposition is used. In other cases, a preposition is missing or an unnecessary preposition is used. Can you identify and correct the preposition errors? Check your answers by copying the sentences into Grammarly. The grammar checker flags and corrects eighteen of the twenty errors. Which two errors does it miss? What are the correct forms? (NB—To facilitate identification, wrong or unnecessary prepositions are in bold here, and a space indicates missing prepositions.)

1. We moved **at** Montreal in 2008.
2. I met her **to** the restaurant.
3. I drove **at** the restaurant.
4. We stayed **to** a hotel.
5. I am waiting __ the bus.
6. I enjoy listening __ the radio.

7. The teacher asked **to** the students to be quiet.
8. They have been abusing **of** drugs.
9. I was born **at** Montreal.
10. My best friend is angry **against** me.
11. They are responsible **of** the mess.
12. She is very interested **to** jazz.
13. These are problems associated **to** adolescence.
14. I like participating **to** this activity.
15. I saw an old woman who suffered **of** dementia.
16. It depends **of** how interested they are.
17. For her birthday, she asked __ a new cell phone.
18. I am worried **of** my son. He does nothing all day long.
19. Are you satisfied **of** the service at the hotel?
20. They are fed up **of** their noisy neighbours.