Bellaterra Journal of Teaching & Learning Language & Literature

Vol. 16(3), Ago/Set 2023, e1142

Writing on steroids? Accuracy of automatic corrective feedback in L2 competence development

Escriuen dopats? Precisió de la retroalimentació correctiva automàtica en el desenvolupament de la L2

Robert Martínez-Carrasco
Alicia Chabert
Universitat Jaume I

Abstract

The article examines the use of written corrective feedback by online grammar checkers as a tool for L2 competence development. While feedback on students' writing is considered an instrumental part in their learning process, providing comprehensive feedback may not always be possible. Grammar checkers can therefore become an interesting tool to scaffold the students' learning process as they promote self-directed learning. For the purposes of this study, a set of authentic EFL compositions in English (n=91; C1 CEFR) was subjected to a popular Automated Writing Evaluation (AWE) system, Grammarly, in order to analyse the feedback provided, namely the types and frequency of errors and the level of accuracy in error detection. The findings determine that, while grammar checkers may be useful as a complementary tool, they are developed with native speakers in mind, so the feedback does not always meet the EFL learners' needs.

Keywords: Grammarly; AWE; EFL; Corrective feedback; Writing competence assessment

Resum

L'article aprofundeix en la retroalimentació correctiva automàtica com a eina per al desenvolupament de la L2. Si bé la retroalimentació de la producció escrita es considera instrumental per a l'aprenentatge de la llengua estrangera, proporcionar-la pot no ser sempre possible, de manera que els correctors, en promoure l'aprenentatge autodirigit, poden convertir-se en una eina útil que apuntale l'aprenentatge de l'alumnat. Així doncs, l'estudi que presentem fa servir l'eina Grammarly per tal d'avaluar la producció escrita en anglés d'un grup d'estudiants hispanoparlants (n=91; C1 MECR) i analitzar-ne la retroalimentació, a saber, els tipus i freqüència d'errors i el nivell de precisió de l'eina. Els resultats suggereixen que, encara que els correctors gramaticals poden ser útils com a eina complementària, solen estar desenvolupats per a parlants natius, per la qual cosa la retroalimentació no sempre satisfà les necessitats dels estudiants.

Paraules clau: Grammarly, Avaluació automàtica de la producció escrita; Anglès com a llengua estrangera; Retroalimentació correctiva; Avaluació de la competència escrita

ISSN: 2013-6196 | https://doi.org/10.5565/rev/jtl3.1142

INTRODUCTION

Developing written communication has always been one of the main objectives in English language teaching in higher education, empowering students and allowing them to express themselves while transferring their knowledge to effectively read and communicate (Graham et al., 2018). However, the writing competence is potentially one of the most difficult abilities for L2 learners to master (Nunan, 1999), and academic writing has been particularly acknowledged as a challenge in higher education (Dolores et al., 2003), yet an essential part of English proficiency assessment and a staple on the English curricula. For this reason, educators aim to continuously identify new strategies and tools to help students improve this skill. With the development of Computer-assisted Language Learning (CALL), along with the current digitalisation of education, more and more tools and software are being introduced to aid students in their language improvement, practising and evaluating their own work. Technology integration in the classroom is not a new challenge anymore, since multimedia platforms, web 2.0 technology, computer-mediated communication technologies (CMC) and the Internet have been well-established since the beginning of the 21st century (Schindler et al., 2017). This range of technologies, as well as teacher training on digital tools, has allowed the implementation of new approaches and programmes in language teaching.

One of the current emerging tools for the development of writing skills is automated writing evaluation software (AWE, also known as 'automated essay evaluation', AEE), the use of which has quickly increased by both teachers and students alike due to its many advantages (Koltovskaia, 2020). While this software is considered relatively new, the use of AWE tools dates back to 1973, with Ellis Page's Project Essay Grade (PEG), a programme developed 'to predict the scores that a number of competent human raters would assign to a group of similar essays' (Page, 2003, p. 47). Page is considered the first of the automated essay scorers, yet this system would not be used until the later development of computing and the availability of technology in the classroom (Khoshnevisan, 2019). The first systems were better known as automated essay scoring (AES), which developed into more sophisticated AWE software based on natural-language processing, artificial intelligence, and latent semantic analysis (Cotos, 2014). At the same time, the focus shifted from summative assessment and scoring to more formative assessment and feedback (Grimes & Warschauer, 2010), which steered the development of these tools into more complex feedback-based software.

AWE programmes use artificial intelligence developed by computational linguistics to evaluate and score writings (Ferster et al., 2012). These days, AWE software is included in well-known writing tools such as *WriteToLearn, Criterion*^{SM}, *LanguageTool*, or *Grammarly*, to name a few. These systems help not only native students improve their writing skills, but also foreign language students improve their English writing skills, reducing the number of errors related to grammar, lexis, style and structure by giving them automated corrective feedback (Dizon & Gayed, 2021). Some of these programmes are considered grammar checkers, rather than AWE software, as they can provide instantaneous feedback and some metalinguistic explanations of grammatical errors. Yet, they are not typically moderated by the educator and do not evaluate the quality of the text (Woodworth & Barkaoui, 2020) nor the student's communicative competence as a whole. While these tools provide feedback on the grammatical and lexical aspects of language, they generally fail to offer a full scope[1] of the social, cultural and pragmatic factors (Canale & Swain, 1980) that allow students to use language effectively in a given social context. Besides, at times evaluation depends on the version of the programme and whether they are Premium (paying version) or Basic, such as the case of *Grammarly*.

In view of the above, this paper focuses on the use of *Grammarly* (Premium) for both feedback and scoring purposes, analysing the validity and frequency of errors and the score provided by the tool. First of all, an overview of the existing literature focusing on *Grammarly* and other AWE systems is provided, as well as the role of feedback and scoring in foreign language education and the use of *Grammarly* by natives and L2 learners. Then, the specific objectives and methodology of the study are presented, followed by the results obtained. Finally, the strengths and weaknesses of this work will be discussed, as well as future lines of research.

LITERATURE REVIEW

Grammarly

Founded in 2009, *Grammarly* is considered one of the most accurate and thorough programmes to date (Cavaleri & Dianati, 2016). This writing tool is powered by an advanced AI that delivers sophisticated communication suggestions in grammar,

[1] To some extent, *Grammarly* does assess clarity, engagement and delivery errors that relate to communicative competence (e.g., sociolinguistic competence). However, they fall beyond the scope of this study, which focuses on correctness errors exclusively.

sentence structure, word choice and readability, and adapts to individual users through advanced machine and deep learning (*Grammarly* 2022[2]). According to their official website, as of 2019 *Grammarly* was recognised by Fast Company as one of the world's most innovative AI companies and ranks among industry leaders on the Forbes Cloud 100 list and The Software Report's Top 100 software companies. At the same time, its user-friendly interface, as well as different paying and free options, makes it a very popular tool for educators, which in turn has sparked a research interest in its efficacy and usefulness (Gain et al., 2019; Huang et al., 2020; Koltovskaia, 2020; Sahu, 2020; Zinkevich & Ledeneva, 2021). Due to the popular use of *Grammarly* in the classroom, numerous studies have focused on its use for evaluation purposes (Dong & Shi, 2021; Ghufron & Rosyida, 2018; Nova, 2018; Park, 2019) and the student's perception of its efficacy as a supporting tool (Huang et al., 2020; Lailika, 2019; Pratama, 2020; Woodworth & Barkaoui, 2020). However, its use is still debated, as there seems to be there still exists a dichotomy between the benefit or hindrance it represents for the students' learning process. Similarly, there is a gap in the research of the validity and usefulness of the type of error identified (and lack thereof) in correlation to the score this software provides.

The present pace of the modern classroom, along with the high teaching workload, has fostered the use of some form of computer-generated assessment to aid the educator, as they may not always be able to provide students with immediate and frequent corrective feedback (Ferster et al., 2012; Woodworth & Barkaoui, 2020). Some researchers, in fact, recommend the integration of AWE feedback to complement and increase the efficacy of educator feedback, allowing them to spend less time on lower-order concerns, such as grammar, and focus more on higher-order concerns, such as content and discourse (Ranalli, 2018). Students, on the other hand, may also benefit from using AWE to practise writing and receive quick feedback, as well as for additional support, improving their writing skills (Wilson, 2017). Nowadays, students are mostly digital natives, which also means that electronic means of learning foreign languages are most appealing to them (Alakrash & Razak, 2021).

Yet, while the use of AWE software may promote learner autonomy and there is positive evidence of the reliability of AWE feedback and even research showing that automated feedback could improve the quality of L2 students writing (Li et al., 2015; Stevenson & Phakiti, 2014), the feedback provided from this type

[2] See: https://www.grammarly.com/

of software tends to be less accurate and more error-prone than teacher feedback, sometimes even contradicting teacher feedback (Woodworth & Barkaoui, 2020). Furthermore, AWE systems supply the same feedback to all learners regardless of their level, whether the corrected essay is written in their mother tongue or a foreign language, or their goals, treating all errors equally (ibid.), hence providing a flawed score. An educator might request L2 students to overly use the passive voice in a certain essay for them to demonstrate they understand its correct use, or it may be that the learners are using a specific dialect which is highlighted as erroneous in the tool. This begs the question of whether these systems do more harm than good when it comes to feedback and scoring, as comprehensive evaluative feedback on the student's communicative competence is not really provided due to the tool's limitations on assessment and feedback. This seems to be in contrast with contemporary, student-centred education scenarios where language is seen as a social tool that speakers use to make meaning (Savignon, 2002) and, consequently, the primary units of language are not merely its grammatical or structural features, but categories of functional and communicative meaning as exemplified in discourse (Richards & Rodgers, 1999).

Feedback and scoring

The main objective of these writing tools is to provide automated written feedback and automated scoring (Bai & Hu, 2017) but, as discussed, we may not always know to what extent their feedback is valid or even helpful. For decades, the role of feedback in Second Language Acquisition has been regarded as an integral part of language acquisition. Already in the early 1980s, Krashen (1982) proposed in his input hypothesis that, when exposed to enough comprehensible input, learners did not require formal grammar instruction, suggesting that feedback was an essential part of the learning process. Since then, most second language acquisition research has focused on negative feedback, as well as the response to learner errors and its effects on linguistic development (Leeman, 2007).

Even from the early years of Second Language Acquisition, research on written corrective feedback has investigated which errors should be corrected and how they should be corrected (Bitchener & Ferris, 2012). In fact, over the past three decades, theoretical perspectives in SLA focusing on the role of error and feedback have become more prominent, as AWE technology identifies a wider range of error types.

Grammarly (Premium) provides feedback on four aspects: correctness (e.g., punctuation, misspellings, incorrect verb form), clarity (e.g., unclear sentence,

passive voice misuse), engagement (e.g., word choice) and delivery (e.g., colloquialisms). This programme also classifies the errors from very low to critical. However, these errors were devised from a L1 perspective, which may affect its applicability on L2 writing (Park, 2019). Table 1 shows the type of error classification used in *Grammarly* according to the software's patent, which specifies that the quality evaluation engine may provide an indication of the criticality of the errors in the text submitted to the tool, such as the errors on Table 1, and explains that the "search criticality of the error or a relative quality indication for the search result may be indicated with each search result, webpage, and the like, such as with a color indication of the search result, a colored symbol (such as a circle, square, icon, and the like), a numeric indication" (Hoover et al., 2015, p. 28).

Table 1. *Grammarly* error classification (Hoover et al., 2015)

High/critical errors	Medium errors	Low errors	Very low/fyi errors
Sentence fragments	Incomplete comparisons	Awkwardness or wordiness	Word overuse
Verb-number agreement	Misplaced/dangling modifiers	Improper adverb or adjective form	Passive voice use
Tense consistency	Pronoun-antecedent agreement	Article errors	Synonyms and word suggestion
There, their, they're	Diction/word-choice errors	Omission of "that" from a noun clause	
Use of non-existent contractions (e.g., could've)	Misuse of passive construction (not to be confused with general passive use)	Incorrect use of subjunctive mood (these may fall under conditional errors)	
Unclear subject	Faulty Parallelism	Misspelling	
Unclear sentence construction/ run-on sentence			
Comma splice			
Made-up words			
Appropriate word order			
Correct pronoun use			

Because the feedback in this programme is instant, it has the potential to increase students' awareness of their mistakes, as well as help them learn grammar, vocabulary, and punctuation (Khoshnevisan, 2019). The issue may come with the accuracy of the detected errors and the false positives obtained, while considering whether the essays were written by native speakers (and if so, depending on the dialect of English they use) or additional language speakers.

Another aspect of AWE tools is automated scoring. Based on the number of errors and their category, the programme users receive a score (1 to 100) along with the automated feedback. Yet, this score is far from straightforward given the

complexity of all nuances involved in the writing context. As Elliot and Klobucar state: "because writing is a complex socio-cognitive construct, the issues involved in its measurement are often as complex as the construct itself" (2013, p. 20). While a numeric score can be useful to provide students with a vague idea of their level, it is vital for scoring systems to identify the features of language that characterise learners at different proficiency levels rather than just quantifying errors. This would allow learners who are still developing in their language to demonstrate their writing competence (Weigle, 2013). At the same time, when comparing AWE scoring to that of human educators, it has been observed in previous research that automated programmes tend to generally provide higher scores and that longer essays are also scored higher (Khoshnevisan, 2019). Yet, the differences are not only found between humans and systems in regard to scoring, but also on the purpose of the tool when used by native speakers or L2 learners.

Grammarly use: Differences between natives and L2 learners

According to its patent, Grammarly was created for a number of purposes, including English grammar improvement for language learners and ESL/EFL students. However, it also claimed to focus on the professional context, such as professionals improving quality of their written work, writers using the grammar checking facility to correct bad writing habits and improve the quality of their writing in general, or anyone who switches from one writing genre or context to another, for instance, from personal emails to professional emails, and so on (Hoover et al., 2015). Yet, these are very different purposes that might not necessarily work together. Despite these claims, *Grammarly*, as well as other grammar checkers and AWE systems, were created from the native speaker perspective, i.e., based on how native speakers would articulate their sentences and how the tool would be used. Grammar and spelling errors differ between native and non-native usage (Lastres-López & Manalastas, 2017), and the context in which the tool is used has a great impact on the usability and applicability of the feedback. Therefore, AWE can improve some writing aspects and is also widely used to support L2 students in their writing development (Dizon & Gayed, 2021).

Previous research has focused on some of the limitations of this type of tool for English language learning. For instance, Heil et al. (2016) already identified significant limitations in the use of certain mobile-assisted language learning tools due to the limited corrective feedback and their inability to adapt to the individual learning process. Ranalli et al. (2017) also pointed out that the tool´s feedback

accuracy may be an issue, and John and Woll (2020) highlighted that there are a few inaccurate replacement forms and false positives actively misleading the user.

OBJECTIVES

Unlike prior research on the topic, this paper's goal is to analyse how *Grammarly*'s error categorisation fits within EFL assessment, focusing on the type of errors flagged, the accuracy of these errors and the purpose of the EFL writer when using this tool, all from the perspective of English as an L2. The authors' research aims to answer the following research questions:

- What type of errors are captured by *Grammarly*?
- Are the errors identified by *Grammarly* actual errors?
- To what extent is *Grammarly* a useful tool for L2 learning?

Even if *Grammarly* may be used as a way to support L2 students' written production, our hypothesis is that the fact that it was created with the aim of enhancing communicative effectiveness from a native speakers' perspective will negatively affect its use as a proficiency assessment tool in L2 settings. Besides, as a learning tool, we believe that *Grammarly*'s level of accuracy in terms of error identification and classification may still require the teacher's feedback to support the students' learning. The objectives of the study may be summarised as follows:

- To analyse the type of errors detected by *Grammarly*
- To analyse the total number and frequency of errors detected
- To determine the level of accuracy of the errors flagged, particularly the number of false positives and errors for which the suggested correction fails to solve the problem.

METHODOLOGY

Research design

As noted, the motivation behind this study arises from first-hand observation of the ESL classroom and the students' intuitive use of *Grammarly* to scaffold their learning process. There is a general belief among students that texts corrected by *Grammarly* should help them achieve at least the pass mark in their writing assignments. However, this is not always the case.

In light of the discrepancies between the evaluation of written work by educators and the automatic feedback provided by *Grammarly*, two groups of

students (n=91) enrolled in the module TI0916 Advanced English for Translators (C1, CEFR) at Universitat Jaume I (Spain) participated in the following study.

A written task was designed involving the composition of a formal letter (250-300 words). In the letter, the students were asked to address a well-known online fashion brand to express their concern about the environmental impact of the fashion industry. The task was conducted in the classroom and students were given 90 minutes to complete it, during which they had no access to support materials or guides.

Data collection & analysis

After the texts were collected, they were digitised, and the corresponding *Grammarly* performance reports were downloaded. Regarding the refinement criteria of the tool, performance reports for the texts were obtained by selecting the following options:

- **Domain:** E-mail, as, from the options available, it most closely resembled the genre required from the students.
- **Intent:** Convince, as the text had a clear persuasive purpose.
- **Audience & Formality:** Expert – Formal, respectively, as a higher register of formal communication was required from the students.

In order to analyse the data extracted from the reports, a data matrix was created, and data was standardised taking both quantitative and qualitative criteria into account. Thus, firstly, global data from the reports was recorded, including score, total number of issues, number of advanced and critical issues, number of words, % of unique words, % of rare words as well as the total number of errors in each category of analysis (incomplete sentences, misuse of quantifiers, spelling, etc.). Then, all the elements flagged by *Grammarly* were classified, one by one, and each error was individually evaluated.

In this first stage of analysis, the "mixed dialects" error category was omitted, because the students did not receive any instructions regarding the dialect of English they should employ for their writing (beyond coherence in its use). Nevertheless, *Grammarly*, to produce its report, requires the introduction of the English variant used and flags any element that does not conform to that variant. In our case, "US English" was used by default for the analysis of the texts, which meant *Grammarly* incorrectly flagged all texts written in other variants of English.

In order to conduct a detailed analysis of errors, a triple encoding of data by three human raters was carried out: the two project investigators, both EFL teachers; and a third outside rater, external to the investigation. All three raters had at least ten years' professional experience teaching English. Raters individually evaluated the errors flagged according to three criteria:

1) The error is correctly flagged and the typology of the error is correct

2) The error is correctly flagged, but the typology assigned to the error is incorrect

3) It is not an error (false positive)

After the individual evaluation, a joint evaluation of the errors took place. In cases where all three raters agreed with the coding of the error, the answer was directly transferred to the data matrix. In cases where there was a discrepancy in the coding of an error, the raters discussed the case until reaching a joint decision.

The following section delves into the results obtained, exploring the typology and incidence of the errors flagged, the human evaluation of those errors, and the false positives and partially correct errors.

RESULTS

General metrics and contextual analysis

Table 2 displays an overview of the data extracted from the texts. As shown, the texts had a regular extension for those requested in EFL exam situations[3] ($\bar{x}$=280.93 words; SD= 22.46).

With regards to the quality of vocabulary used, *Grammarly* distinguishes between "unique words", that is, the vocabulary diversity resulting from calculating the percentage of words used only once in the text; and "rare words", a measurement of the depth of vocabulary obtained "by identifying words that are not among the 5,000 most common English words". In the study at hand, students reached an average of 60% of unique words in their texts ($\bar{x}$=59.15; SD=4.09), and the percentage was no lower than 48% in any of the 91 cases under analysis. In the

[3] While in TOEFL, for example, there is no maximum length, the exam informs students that "an effective response will contain a minimum of 300 words" (in the independent writing task) or 150-225 words (in the integrated writing task). In CAE, on the other hand, students are asked to "write their answer in 220 – 260 words in an appropriate style" for both tasks. The instructions for the study we are conducting requested texts of between 250-300 words.

case of "rare words", the percentage was remarkably lower ($\bar{x}$=26.53; SD=4.17). However, if we consider the profile of the students, who are non-native speakers of English, the result need not be interpreted in a negative light.

Table 2. General metrics and writing issues of texts under analysis (n=91)

	Total	$\bar{X}$	SD	CI
Score	-	74.21	10.82	(46;96)
Issues	1876	20.62	6.19	(7;38)
Critical issues	546	6	3.67	(0;17)
Advances issues	1330	14.62	4.61	(6;28)
Words	-	280.93	22.46	(233;346)
Unique words (%)	-	59.15	4.09	(48;68)
Rare words (%)	-	26.53	4.17	(18;37)

With regards to the errors detected, *Grammarly* flagged a total of 1876 errors, with an average of 20.62 errors per text (SD=6.19). Of these errors, 70.9% were considered advanced errors, that is, reflective of the level of the students; while 29.1% were considered critical errors. However, *Grammarly* reports do not provide information about which of the flagged errors belong to one group or to the other, and thus it was not possible to conduct an alternative human revision to question *Grammarly*'s classifications.

Lastly, a controversial[4] aspect of *Grammarly*'s general metrics is the performance score. For this platform, the performance score is no more than the indicator of "how accurate [a text] is compared to documents written by other *Grammarly* users who set the same goals". Accuracy, in this case, is measured by considering the total word count and the number and types of writing issues detected. It is not a comprehensive evaluation of the overall quality of the text as such, but rather an evaluation of how the text is positioned among the set of texts that have previously passed through the correction engine. This approach to scoring neglects, as we have seen, other language proficiencies (discourse, sociolinguistic, and strategic competence) that are equally important in communicative learning scenarios (Bachman & Palmer, 2010; Canale, 1983; Celce-Murcia, 2007; Chabert & Agost, 2020).

In this case, we can see that the performance score of the analysed texts was particularly high ($\bar{x}$=74.21; SD=10.82). Remarkably, only three of the texts were

[4] Students often confuse *Grammarly's* performance score with the grade their text deserves, disregarding both discursive and contextual aspects that *Grammarly* can hardly evaluate, such as meeting the instructions successfully, the quality of the written production in relation to the level of language requested, the complexity of the content, etc.

below 50%, and all three were very close to 50% (Text 54 = 46%; Text 6 = 49%; Text 72 = 49%). In one case (Text 83) the score even reached 96%.

Typology and incidence of errors detected

Table 3 displays the total set of errors detected by *Grammarly*, divided according to the classifications proposed by the tool (correctness, clarity, engagement, delivery).

Table 3. Typologies and incidence of errors detected

Correctness (n=761)		Incomplete sentence	17	**Clarity (n=834)**	
Punctuation compound sentences	215	Conjunction	15	Wordy sentence	353
Determiner	84	Subject-verb agreement	10	Passive voice misuse	233
Coma misuse between clauses	79	Closing punctuation	8	Unclear sentence	164
Wrong or missing preposition	64	Misuse of quantifiers	8	Unclear paragraph	61
Confused words	42	Misuse of semicolons, quotations, etc.	7	Intricate text	29
Verb form	40	Text inconsistency	7	Hard-to-read text	2
Pronoun use	38	Misuse of modifiers	7	Outdated language	2
Misplaced words/phrases	27	Faulty tense sequence	3	**Engagement (n=150)**	
Spelling	25	Unknown words	2	Word choice	150
Incorrect noun number	23	Faulty parallelism	2	**Delivery (n=30)**	
Incorrect phrasing	20			Colloquialism	18
Improper format	18			Tone suggestion	12

The errors most frequently flagged were errors of clarity (n=834), closely followed by errors of correctness (n=761). With regard to errors of clarity, the categories "wordy sentence" and "passive voice misuse" have the highest incidence (n=353; n=223, respectively).

In contrast to the use that a native speaker could make of *Grammarly*, the main goal of the texts written by EFL students is to mobilise their discursive, lexical, semantic, syntactic, and grammatical skills, while demonstrating that they have reached a particular level in written production of the foreign language. In an exam situation, the student is not asked for a functional and useful text, but rather a text that shows that they can, at the very least, master the particularities of a genre (essays, reports, proposals, etc.) and that they are able to incorporate certain grammatical structures, complex lexical items, and specific discourse strategies.

A good example of this is the use of the passive voice, something which *Grammarly* tends to mark as an error of clarity. In the following two examples

(Tables 4), extracted from the corpus of *Grammarly* reports, it can be noted how the passive voice, far from posing a clarity problem, is used by students as part of a discourse strategy, without entailing an error.

Table 4. Passive voice misuse samples

Code	Extract
T51-5	(...) the worrying conditions in which labourers* **are forced** to work.
T88-17	If nothing **is done**, this will cause great damage to future generations.

The same could be said in the case of wordy sentences. In an effort to raise the register of the text, students at the most advanced levels tend to expand and amplify structures which, from the perspective of the economy of language, could be simplified (Table 5).

Table 5. Wordy sentence samples

Code	Extract	Correction
T46-13	Furthermore, producing poor-quality clothing makes that **a significant amount of** microplastics ends up in the ocean	*many*
T1-19	You claim to be committed to improve* the work conditions of your employees **in order for them to** (...)	*for them to*

Rather than simplifying and making their message sound more straightforward, the priority of students is to show their versatility and ability to integrate complex ideas, opting for formality and a more elaborate style which, without doubt, adds a somewhat stilted (or artificial) dimension to their text. The following two examples, both classified as "unclear sentences" by *Grammarly*, provide further evidence of this (Table 6).

Table 6. Unclear sentence samples

Code	Extract
T4-15	With regard to recycling, I would like to remark how, these days, most clothes are manufactured from new materials, which inevitably leads to more and more pollution.
T89-3	The overproduction of clothes not only contributes to the throwaway culture that surrounds us, but also negatively affects the environment.

In relation to errors of correctness, which are fewer when compared to clarity errors, it can be observed how *Grammarly* flags, above all, punctuation errors, followed by determiners, preposition errors and confused words. In fact, if we add up the four categories of punctuation-related errors (punctuation in compound sentences, comma misuse between clauses, closing punctuation, misuse of semicolons,

quotations marks, etc.), they account for 40.6% of correctness errors (n=309). Table 7 shows some sentences where *Grammarly* flags the misuse of punctuation.

Table 7. Punctuation errors flagged

Code	Extract	Correction
T63-16	It is inacceptable for a brand like yours to hide such relevant information about the problems delocalisation triggers around the **globe; not** to mention the work conditions you force your workers to have in your factories in underdeveloped countries.	*globe, not*
T38-8	Yet, clothes in landfills take hundreds of years to **degrade and** the chemicals in them gradually contaminate the soil and groundwater.	*degrade, and*
T15-9	The results showed that most of them had fire, **building and** electrical hazards.	*building, and*

While there is no doubt of the need for good punctuation in both mother tongue(s) and foreign language(s), fixing existing punctuation errors in the text may give learners the false feeling that the quality of their text improves substantially when it actually only does so superficially, in the best of cases.

Human evaluation of flagged errors

Finally, as mentioned above, a triple individual check of the errors flagged was carried out so as to validate the level of accuracy of the errors. Due to space limitations, we will only delve into the results relating to correctness, as these are the ones with the greatest variability of response.

As can be seen in Table 8, the raters estimated 71.98% accuracy of the errors flagged (SD=36.26). Besides, in 11.27% of cases (SD=2.06), although the fragment did present an error, the error was not well catalogued, the solution was not satisfactory, or it did not solve the problem.

The highest level of accuracy was detected in the errors of subject-verb agreement (n=10), faulty text sequence (n=3), and faulty parallelism (n=2), all three with an accuracy level of 100%. However, these are errors with a low incidence in the total number of correctness errors (15 of the 847 total).

We will now analyse in more detail the most representative cases where the errors detected by *Grammarly* were not considered errors per se or where *Grammarly*'s accuracy was considered to be only partially correct.

Table 8. Human evaluation of correctness errors

Error	Total	Correct	%	Not correct	%	Partly correct	%
Punctuation compound sentences	215	169	78,6	40	18.60	6	2.79
Determiner	84	66	78,57	12	14,29	6	7,14
Comma misuse bet. clauses	79	69	87,34	9	11,39	1	1,27
Wrong missing preposition	64	54	84,38	4	6,25	6	9,38
Confused words	42	33	78,57	5	11,9	4	9,52
Verb form	40	28	70	7	17.5	5	12.5
Pronoun use	38	28	73,68	5	13,16	5	13,16
Misplaced words or phrases	27	12	44,44	13	48,15	2	7,41
Spelling	25	18	72	3	12	4	16
Incorrect noun number	23	16	69,67	5	21,74	2	8,7
Incorrect phrasing	20	16	80	0	0	4	20
Improper Format	18	15	83,33	1	5,56	2	11,11
Incomplete sentence	17	13	76,47	3	11,65	1	5,88
Conjunction	15	11	73,33	1	6,67	3	20
Subject-verb agreement	10	10	100	0	0	0	0
Closing punctuation	8	6	75	0	0	2	25
Misuse of quantifiers	8	4	50	2	25	2	25
Misuse of semicolons quotations	7	5	71,43	2	28,57	0	0
Text inconsistency	7	0	0	7	100	0	0
Misuse of modifiers	7	6	85,71	0	0	1	14,29
Faulty tense sequence	3	3	100	0	0	0	0
Unknown words	2	0	0	1	50	1	50
Faulty Parallelism	2	0	100	2	0	0	0
TOTAL	**847**	**582**	-	**122**	-	**57**	-
x̄	-	-	**71.98**	-	**17.76**	-	**11.27**
SD	-	-	**36,26**	-	**8,26**	-	**2,06**

False positives

In most cases, the false positives provided acceptable linguistic forms in English, both in the flagged error and in the proposed solution. Table 9 shows some examples.

Example T1-33 flags the use of the past participle in the case of "shown/showed". While it is true that "shown" is the predominant past participle in English, the Oxford English Dictionary clarifies that it only became common in the 19th century. The original "showed", less used these days, is still used in cases of present or past perfect. In this particular sentence, it could be argued that "showed" is merely the elided version of "that you have showed/shown".

Table 9. Cases where there is no error and the proposed solution is valid

Code	Extract	Correction
T1-33	I would be willing to contribute with my knowledge and give advice to the brand so as to address the lack of care **showed** towards environmental issues until now.	*shown*
T38-8	And all these comments should be taken into account so as to collaborate together and find solutions **which** radically cut back on the ecological footprint.	*that radically*
T44-14	One of their former employees, Van Chou from Bangladesh, declared **in 2016** that they were treated like dogs.	*(Misplaced words/phrases)*
T79-5	Not only does it have millions of customers worldwide, but also a powerful media campaign thanks to a great deal of microinfluencers who advertise their products.	*(Incomplete sentence)*

The same happens with the relative pronoun "that/which" in defining relative clauses, considered a pronoun use error according to *Grammarly* [Example T38-8]. The use of "which" for defining relative clauses is more than widespread in contemporary English. In fact, opting for the pronoun "which" instead of "that" may help students to increase the register of their text, as the latter is usually more associated with spoken English.

In other cases, as shown in Table 10, *Grammarly* suggested an error as a solution in a fragment that did not present any error originally. This seems to apply to all types of errors.

In the first case (Example T21-17), *Grammarly* shows a determiner use error, perhaps induced by the structure "child labour/forced labour". As both adjectives share the same noun, *Grammarly* did not detect that "child" should be read together with "labour".

In Example T62-13, a pronoun use error, *Grammarly* suggests "them" probably because of the closest reference available in the sentence ("concerns"). However, the student is referring to something which transcends the sentence level. With this "it", the student is pointing to the idea that they were developing throughout the text.

Table 10. Cases where there is no error but where the solution introduces an error in the text

Code	Extract	Correction
T21-17	Moreover, your company claims that is has never used **child** or forced labour.	*a child, the child*
T62-13	I very much hope your will take all my concerns into account and do something about **it**.	*it → them*

Another element that *Grammarly* flags systematically is purpose clauses and the place they occupy in the sentence. Although it is true that starting the sentence with a purpose clause breaks the logical order of the sentence, its use emphasises the discourse. In the following cases (See Table 11) the learners modulate their discourse effectively without this resulting in error.

Table 11. False errors flagged on the position of purpose clauses

Code	Extract	Correction
T7-9	**To avoid this**, it would be convenient to take some measures, such as using recycled materials to avoid the use of raw materials.	*Misplaced words or phrases*
T74-3	**To solve this problem**, more sustainable measures should be taken.	

Finally, other common false positives are listed in Table 12: penalising the use of numbers instead of their written equivalent (Examples T4-10 or T74-3, categorised as improper formatting errors), or confusing names or references from the learner's own language and culture as spelling errors (Examples T24-14 or T25-22).

Table 12. Other false errors flagged

Code	Extract	Correction
T4-10	Therefore their lifespan will not be higher than **2** years.	*two*
T74-3	It's been **3** years since Shein became very popular.	*three*
T24-14	Yours, Pere **Daviu**.	*David*
T25-22	Yours faithfully, **Marina de** Nicasio.	*Marinade*

Partially correct errors

In some cases, *Grammarly* was able to detect an incorrect text sequence but failed to solve the problem. In the first case listed in Table 13 (Example T68-17/18), for instance, we see how *Grammarly* identifies the grammatical error "for avoid". However, it suggests two changes in the text so that the first solution to the error invalidates the second. Thus, if we accept error 17 (changing the preposition "for" to the preposition "to"), the error in the sentence is solved. However, if we also accept error 18 (avoid → avoiding) we find that the resulting structure ("to avoiding") is still an error.

In the second case (Example T58-3/4), the same problematic structure ("business' practices") has been flagged twice, both as a confused word and as an incorrect noun number. Both, however, lead to the same error solution.

This duplicity in error cataloguing has a direct impact on the score of the text, as two different errors have been flagged (wrong preposition-incorrect verb form) when there is only one error.

Table 13. Grammarly flags an error as two different errors

Code	Extract	Error 1	Error 2
T68-17/18	You are not thinking about doing anything **for avoid** this situation in the future.	*For → to (Wrong or missing prepositions)*	*Avoid → avoiding (Incorrect verb forms)*
T58-3/4	As a potential customer I like to keep track of the **business'** practices	*Business' → business (Confused words)*	*Business' → business (Incorrect noun number)*

In other cases, Grammarly finds the error but fails to catalogue it and, consequently, provide a valid solution. Several examples are given in Table 14. Example T6-12, for instance, shows an omission of a textual fragment. The students start their sentence with "In the same way as" and leave blank the part that would complete the first part of the sentence. *Grammarly* detects this as an unnecessary addition of the conjunction "as" and proposes to remove it. However, without the missing fragment in the sentence, the sentence makes no sense. Something similar happens in Example T28-11, where an incorrect revision by the learner causes "scarce" to become "scare", something which *Grammarly* interprets as an incorrect verb form when it should be a spelling error.

Examples T54-34 and T67-7 are particularly paradigmatic in that they reveal the influence of the learner's L1 on their written production in L2. In the first case, because, in both Catalan and Spanish[5], the term "advice" (*consell, consejo*) is countable, hence the "advices" in the text. This is not a problem of pronoun use, but of how the learner has internalised the use and form of "advice" in parallel to how they use it in their own language.

In the second example (T67-7) the influence of *consumisme/consumismo* (consumerism) affects the way the learner attempts to write "overconsumerism". *Grammarly's* suggestion (overconsume) does not fit the learner's sentence, which needs a noun to make any sense.

[5] The participants of the study live in a bilingual area of Spain where both Catalan and Spanish are spoken.

Table 14. Grammarly flags an error and its proposed solution is not correct

Code	Extract	Correction
T1-15	There are many compelling arguments **giving lie** to your declarations.	*the lie*
T6-12	In the same way as, the labour conditions of garment workers (…)	*as*
T6-31	I would like to offer my help and you bear in mind it.	*it in mind*
T28-11	Your company will no longer be able to continue manufacturing because national resources will become increasingly **scare**.	*scare → scared*
T54-34	**These are my advices**, I hope you find them helpful.	*These are → This are*
T67-7	In addition, because of its low prices, this fashion label is encouraging **over-consumism**.	*overconsumism → overconsumerism*
T64-9	In other words, it contributes to the culture of using and throwing away clothes quickly, which leads **to damage** the environment.	*to the*

Table 15 shows errors related to punctuation. Within this category we also find cases where, although human evaluation would not classify them as errors per se, they could improve the reading of the text. They are, in any case, stylistic additions which may improve the quality of a text which, in principle, had no errors.

Table 15. Grammarly flags an improvement as an error

Code	Extract	Correction
T81-2	In the newspaper article we are told that some international clothing brands (…)	*article,*
T50-6	I am not only referring to the fabrics, I am also talking about the way your company obtains the energy.	*, but,*

When it comes to clauses, there is a certain tendency for *Grammarly* to only suggest the comma that closes the clause and not the comma that opens it, which hinders the reading of the text to the point of turning a text with a somewhat complex reading into a badly punctuated text. In the case of Example T73-6, the comma after "disaster" necessarily requires a comma after "how" in order for the clause to be punctuated coherently. However, *Grammarly* only marks as an error the absence of the second comma (Table 16).

Table 16. Use of commas to close (and not to open) sentences

Code	Extract	Correction
T73-6	Yesterday, I read an article that explained how even after the Bangladesh **disaster** garment workers are still losing their lives making those clothes.	*disaster,*
T87-17	Furthermore, with greener policies people would feel more attracted to your brand because day by **day** more and more people are becoming aware of the type of clothes they buy and the impact they have.	*day,*

Finally, it can also be observed that, in some cases, the solution to certain problems at sentence level is not the introduction of more punctuation, but the segmentation of sentences. The examples given in Table 17, all flagged as punctuation errors in compound/complex sentences or comma misuse within clauses, would be significantly improved if the learner had chosen to segment the sentence.

Table 17. Punctuation errors that could be resolved with sentence segmentation

Code	Extract	Correction
T46-11	In addition, it is possible that they don't spend money on sustainable **products and** that is why their prices are so cheap.	*, and*
T42-1/2	Shein is an international e-commerce company that sells affordable products online **and** because of **that** it is considered a business related to the fast fashion culture, such as Primark.	*1) and,* *2) that,*

CONCLUDING REMARKS

This study has analysed the affordances and limitations of the use of *Grammarly* as an automated writing evaluation (AWE) tool in the English L2 classroom. The findings offer insight into the reality of the feedback provided by the tool and its implications for writing improvement in English as a Foreign Language. As discussed in Bailey and Lee's exploratory study of *Grammarly* (2020), there is a need for further analysis on how AWE programmes perform across different L2 error types, which this paper has aimed to shed some light on. The results obtained demonstrate a high level of error accuracy (an estimated 71.98%), which supports the use of this tool as a scaffolding tool in L2 learning thanks to its immediate and relatively accurate corrective feedback. However, both educators and learners should take into account that this software was created with native speakers in mind, meaning that the type of input expected would be different. Not only would English native users naturally tend to make shorter, purpose-oriented sentences when compared to natives of other languages, but the tool's primary function was that of focusing on communication efficiency, rather than language proficiency demonstration. EFL learners, on the other hand, aim to show their ability to express their own ideas and demonstrate their language proficiency by using complex sentences, specific vocabulary and incorporating certain grammatical structures (i.e., passive voice, which is often flagged as an error). At the same time, the rare words percentage is often contributed by the student's mother tongue (in this case Catalan or Spanish, which are Latin-based languages), which can potentially unknowingly change the register used.

When using *Grammarly,* learners should also consider the fact that some of its proposals are incorrect or not fully correct, so this tool cannot be solely taken as a perfect solution but rather a support tool to their own learning and teacher feedback, especially if we take into account that *Grammarly* cannot provide comprehensive feedback on certain elements (strategic, discursive or sociolinguistic) that truly allow learners to use language effectively in a social context. This especially affects sentence structure and punctuation. As it was observed, punctuation-related errors (punctuation in compound sentences, comma misuse between clauses, closing punctuation, misuse of semicolons, etc.) account for 40.6% of correctness errors. However, sentence segmentation would improve the whole coherence of the text. This would be a great improvement to this tool that would require this AWE system to view the text as a whole in order to determine the best segmentation for coherence and correctness. All in all, *Grammarly* seems to be an effective tool for English assessment as long as this evaluation is complemented by teacher feedback and the proposed tool correction is proofread.

Despite its perceived effectiveness, it should be noted that our research focused exclusively on a relatively small sample (n=91) located in a particular geographical and sociolinguistic setting (a bilingual region in Spain), with higher education students. Therefore, as future avenues of research, it would be advisable to apply the same research methodology to other EFL students with different linguistic and sociocultural backgrounds. Another interesting approach would be to use other tool refinement criteria (domain, intent, audience and formality), so as to test the effectiveness of other features of the tool. Finally, applying the same research conditions to other students with different levels of linguistic competence (beginners, intermediate, upper-intermediate) may also help us understand the overall applicability of *Grammarly*.

REFERENCES

Alakrash, H. M., & Razak, N. A. (2021). Technology-based language learning: Investigation of digital technology and digital literacy. *Sustainability (Switzerland)*, *13*(21). https://doi.org/10.3390/su132112304

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: how do students respond? *Educational Psychology*, *37*(1), 67-81. https://doi.org/10.1080/01443410.2016.1223275

Bailey, D., & Lee, A. R. (2020). An Exploratory Study of Grammarly in the Language Learning Context: An Analysis of Test-Based, Textbook-Based and Facebook Corpora. *TESOL International Journal*, *15*(2), 4-27.

Bitchener, J., & Ferris, D. R. (2012). *Written Corrective Feedback in Second Language Acquisition and Writing*. Routledge.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication* (pp. 2-27). Routledge.

Canale, M. & Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics*, *1*(1), 1-47.

Cavaleri, M., & Dianati, S. (2016). You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students. *Journal of Academic Language & Learning*, *10*(1), 223.

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E. Alcón Soler & M. P. Safont Jordà (Eds.), *Intercultural language use and language learning* (pp. 41-57). Springer.

Chabert, A. & Agost, R. (2020). Communicative language teaching: Is there a place for L1 in L2 learning? *European Journal of Language Policy*, *12*(1), 55-83. https://doi.org/10.3828/ejlp.2020.4

Cotos, E. (2014). *Genre-Based Automated Writing Evaluation for L2 Research Writing: From Design to Evaluation and Enhancement*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137333377_1

Dizon, G., & Gayed, J. M. (2021). Examining the Impact of Grammarly on The Quality of Mobile L2 Writing. *JALT CALL Journal*, *17*(2), 74-92. https://doi.org/10.29140/JALTCALL.V17N2.336

Dolores, P., Keselman, A., & Monopoli, M. (2003). The academic writing of community college remedial students: Text and learner variables. *Higher Education*, *45*, 19-42.

Dong, Y., & Shi, L. (2021). Using Grammarly to support students' source-based writing practices. *Assessing Writing*, *50*(September), 100564. https://doi.org/10.1016/j.asw.2021.100564

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 16-35). Routledge.

Ferster, B., Hammond, T. C., Curby Alexander, R., & Lyman, H. (2012). Automated formative assessment as a tool to scaffold student documentary writing. *Journal of Interactive Learning Research*, *23*(1), 81-99.

Gain, A., Rao, M., & Bhat, K. S. (2019). Usage of Grammarly - online grammar and spelling checker tool at the health sciences library, Manipal Academy of Higher Education, Manipal: A Study. *Library Philosophy and Practice*, *2019*.

Ghufron, M. A., & Rosyida, F. (2018). The Role of Grammarly in Assessing English as a Foreign Language (EFL) Writing. *Lingua Cultura*, *12*(4), 395. https://doi.org/10.21512/lc.v12i4.4582

Graham, S., MacArthur, C. A., & Fitzgerald, J. (2018). *Best practices in writing instruction* (Third). Guilford Press.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*(6), 1-43.

Heil, C. R., Wu, J. S., Lee, J. J., & Schmidt, T. (2016). A review of mobile language learning applications: Trends, challenges, and opportunities. *The EuroCALL Review, 24*(2), 32–50. https://doi.org/10.4995/eurocall.2016.6402

Hoover, B., Lytvyn, M., & Shevchenko, O. (2015). *Systems and Methods for Advanced Grammar Checking* (Patent No. US 9002700 B2). U.S. Patent and Trademark Office.

Huang, H. W., Li, Z., & Taylor, L. (2020). The Effectiveness of Using Grammarly to Improve Students' Writing Skills. In *ACM International Conference Proceeding Series, 2020* (pp. 122-127). https://doi.org/10.1145/3402569.3402594

John, P., & Woll, N. (2020). Using grammar checkers in an ESL context: An investigation of automatic corrective feedback. CALICO Journal, *37*(2), 169-192. https://doi.org/10.1558/cj.36523

Khoshnevisan, B. (2019). The affordances and constraints of automatic writing evaluation (AWE) tools: A case for Grammarly. *ARTESOL EFL Journal*, *2*(2), 12-25.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, *44*(September 2019), 100450. https://doi.org/10.1016/j.asw.2020.100450

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. Pergamon Press Inc. http://www.sdkrashen.com/content/books/principles_and_practice.pdf

Lastres-López, C., & Manalastas, G. (2017). Errors in L1 and L2 University Students' Writing in English: Grammar, Spelling and Punctuation. *RAEL: Revista Electrónica de Lingüística Aplicada*, *16*(2), 118-135

Lailika, H. I. (2019). *Students' Peceptions of The Use of Grammarly as an Online Grammar Checker in Thesis Writing*. http://digilib.uinsby.ac.id/34607/

Leeman, J. (2007). Feedback in L2 learning: Responding to errors during practice. In R. DeKeyser (Ed.), *Practice in a Second Language: Perspectives from Linguistics and Psychology* (pp. 111-138). Cambridge University Press. https://doi.org/10.1017/cbo9780511667275.007

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1-18. https://doi.org/10.1016/j.jslw.2014.10.004

Nova, M. (2018). Utilizing Grammarly in Evaluating Academic Writing: a Narrative Research on Efl Students' Experience. *Premise: Journal of English Education*, *7*(1), 80-97. https://doi.org/10.24127/pj.v7i1.1300

Nunan, D. (1999). *Second language teaching & learning*. Heinle & Heinle Publishers.

Page, E. B. (2003). Project Essay Grade: PEG. In M. Shermis & B. JC (Eds.), *Automated essay scoring: A cross-disciplinary perspective.* (pp. 43-54). Lawrence Erlbaum Associates Publishers.

Park, J. (2019). An AI-based English Grammar Checker vs. Human Raters in Evaluating EFL Learners' Writing. *Multimedia-Assisted Language Learning*, *22*(1), 112-131. http://journal.kamall.or.kr/wp-content/uploads/2019/3/Park_22_1_04.pdfhttp://www.kamall.or.kr

Pratama, Y. D. (2020). The Investigation of Using Grammarly As Online Grammar Checker in the Process of Writing. *English Ideas: Journal of English Language Education*, *1*(1), 46-54.

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

Richards, J. C., & Rodgers, T. (1999). *Approaches and methods in language teaching*. Cambridge University Press.

Sahu, S. (2020). Evaluating performance of different grammar checking tools. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 2227-2233. https://doi.org/10.30534/ijatcse/2020/201922020

Savignon, S. J. (2002). Communicative Language Teaching: Linguistic Theory and Classroom Practice. In S. J. Savignon (Ed.), *Interpreting Communicative Language Teaching: Contexts and Concerns in Teacher Education* (pp.1-28). Yale University Press.

Schindler, L. A., Burkholder, G. J., Morad, O. A., & Marsh, C. (2017). Computer-based technology and student engagement: a critical review of the literature. *International Journal of Educational Technology in Higher Education*, *14*(1). https://doi.org/10.1186/s41239-017-0063-0

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51-65. https://doi.org/10.1016/j.asw.2013.11.007

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

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, *30*(4), 691-718. https://doi.org/10.1007/s11145-016-9695-z

Woodworth, J., & Barkaoui, K. (2020). Perspectives on Using Automated Writing Evaluation Systems to Provide Written Corrective Feedback in the ESL Classroom. *TESL Canada Journal*, *37*(2), 234-247. https://doi.org/10.18806/tesl.v37i2.1340

Zinkevich, N. A., & Ledeneva, T. V. (2021). Using *Grammarly* to Enhance Students' Academic Writing Skills. *Professional Discourse & Communication*, *3*(4), 51-63. https://doi.org/10.24833/2687-0126-2021-3-4-51-63

———— & ————

ROBERT MARTÍNEZ-CARRASCO

Holds a PhD in Applied Languages, Literature and Translation from Jaume I University (Spain). His research interests include translation education, critical pedagogy and the construction of gender, sex and identity. At Universitat Jaume I, he teaches courses in Catalan-English translation and English linguistics.

rcarrasc@uji.es
https://orcid.org/0000-0002-2148-8637

ALICIA CHABERT

Holds a PhD in Applied Languages, Literature and Translation from Jaume I University (Spain). Her research interests include the role of the mother tongue in foreign language learning and English as a Lingua Francia. At Universitat Jaume I, she teaches Spanish-English translation.

chabert@uji.es
https://orcid.org/0000-0003-4599-8524

———— & ————

Martínez-Carrasco, R. & Chabert, A. (2023). Writing on steroids? Accuracy of automatic corrective feedback in L2 competence development. *Bellaterra Journal of Teaching & Learning Language & Literature*, *16*(3), e1142. https://doi.org/10.5565/rev/jtl3.1142

———— & ————

Rebut / Recibido / Received / Reçu: 17-05-2022
Acceptat / Aceptado / Accepted / Accepté: 14-04-2023

https://revistes.uab.cat/jtl3/