Comparing the accuracy and effectiveness of Wordvice AI Proofreader to two automated editing tools and human editors

Kevin Heintz¹, Younghoon Roh², Jonghwan Lee²

¹Department of Research & Development, Wordvice Editing Service, Des Moines, IA, USA; ²Department of Research & Development, Wordvice Editing Service, Seoul, Korea

Abstract

Purpose: Wordvice AI Proofreader is a recently developed web-based artificial intelligence-driven text processor that provides real-time automated proofreading and editing of user-input text. It aims to compare its accuracy and effectiveness to expert proofreading by human editors and two other popular proofreading applications—automated writing analysis tools of Google Docs, and Microsoft Word. Because this tool was primarily designed for use by academic authors to proofread their manuscript drafts, the comparison of this tool's efficacy to other tools was intended to establish the usefulness of this particular field for these authors.

Methods: We performed a comparative analysis of proofreading completed by the Wordvice AI Proofreader, by experienced human academic editors, and by two other popular proofreading applications. The number of errors accurately reported and the overall usefulness of the vocabulary suggestions was measured using a General Language Evaluation Understanding metric and open dataset comparisons.

Results: In the majority of texts analyzed, the Wordvice AI Proofreader achieved performance levels at or near that of the human editors, identifying similar errors and offering comparable suggestions in the majority of sample passages. The Wordvice AI Proofreader also had higher performance and greater consistency than that of the other two proofreading applications evaluated.

Conclusion: We found that the overall functionality of the Wordvice artificial intelligence proofreading tool is comparable to that of a human proofreader and equal or superior to that of two other programs with built-in automated writing evaluation proofreaders used by tens of millions of users: Google Docs and Microsoft Word.

Keywords

Artificial intelligence; Natural language processing; English proofreading; Writing assistant; Human editing

Received: July 23, 2021
Accepted: November 8, 2021

Correspondence to Kevin Heintz
content@wordvice.com

ORCID
Kevin Heintz
<https://orcid.org/0000-0002-8964-2573>
Younghoon Roh
<https://orcid.org/0000-0002-1008-5699>
Jonghwan Lee
<https://orcid.org/0000-0002-4660-6046>

Introduction

Background/rationale: The use of English in all areas of academic publishing and the need for nearly all non-native English-speaking researchers to compose research studies in English have created difficulties for non-native English speakers worldwide attempting to publish their work in international journals. Faced with the time-consuming process of self-editing before submission to journals, many researchers are now using Automated Writing Analysis tools to edit their work and enhance their academic writing development [1,2]. These include grammatical error correction (GEC) programs that automatically identify and correct objective errors in text entered by the user. At the time of this study, most popular GEC tools are branded automated English proofreading programs that include Grammarly [3], Ginger Grammar Checker [4], and Hemingway Editor [5], all of which were developed using natural language processing (NLP) techniques; NLP is a type of artificial intelligence (AI) technology that allows computers to interpret and understand text in the same way a human does.

Although these AI writing and proofreading programs continue to grow in popularity, reviews regarding the effectiveness of these programs at large are inconsistent. Similar studies to the present one have analyzed the effectiveness of NLP text editors and their potential to approach the level revision of expert human proofreading [6-8]. At least one 2016 article [9] evaluates popular GEC tools and comes to the terse conclusion that “grammar checkers do not work.” The jury appears to be out on the overall usefulness of modern GEC programs in correcting writing.

However, Napoles et al. [10] propose applying the Generalized Language Evaluation Understanding (GLEU) metric, a variant of the Bilingual Evaluation Understudy (BLEU) algorithm that “accounts for both the source and the reference” text, to establish a ground truth ranking that is rooted in judgements by human editors. Similarly, the present study applies a GLEU metric to more accurately compare the accuracy of these automated proofreading tools with that of revision by human editors. While the practical application of many of these programs is evidenced by their success in the marketplace of writing and proofreading aids, gaps remain in how accurate and consistent certain AI proofreading programs are in correcting grammatical and spelling errors.

Objectives: It aimed to analyze the effectiveness of the Wordvice AI Proofreader [11], a web-based AI-driven text processor that provides real-time automated proofreading and editing of user-input text. We also compared its effectiveness to expert proofreading by human editors and two other popular writing tools with proofreading and grammar checking applications, Google Docs [12] and Microsoft (MS) Word [13].

Methods

Ethics statement: This is not a human subject study. Therefore, neither approval by the institutional review board nor the obtainment of informed consent is required.

Study design: This was a comparative study using qualitative open dataset and quantitative GLEU metric of comparison.

Setting: The Wordvice AI Proofreader tool was measured in terms of its ability to identify and correct objective errors, and it was evaluated by comparing its performance to that of experienced human proofreaders and to two other commercial AI writing assistant tools with proofreading features: MS Word and Google Docs in June 2021. By combining the application of a quantitative GLEU metric with a qualitative open-dataset comparison, this study compared the effectiveness of the Wordvice AI Proofreader with that of other editing methods, both in the correction of “objective errors” (grammar, punctuation, and spelling) and in the identification and correction of more “subjective” stylistic issues (including weak academic language and terms).

Data sources

Open datasets

The performance of the Wordvice AI Proofreader was measured using the JHU FLuency-Extended GUG (JFLEG) open dataset 1 [14], a dataset developed by researchers at Johns Hopkins University and consisting of a total of 1,501 sentences, 800 of which were used to comprise Dataset 1 in the experiment (<https://github.com/keisks/jfleg>). The JFLEG data consists of sentence pairs, showing the input text and the results of proofreading by professional editors. These datasets assess improvements in sentence fluency (style revisions), rather than recording all objective error corrections. According to Sakaguchi et al. [15], unnatural sentences can result when the annotator collects only the minimum revision data within a range of error types, and letting the annotator rephrase or rewrite a given sentence can result in more comprehensible and natural sentences. Thus, the JFLEG data was applied with the aim of assessing improvements in textual fluency rather than simple grammatical correction.

Because many research authors using automated writing assistant tools are English as a second language writers, the proofread data was based on sentences written by non-native English speakers. This was designed to create a more accurate sample pool for likely users of the AI Proofreader. “Proofread data” refers to data that has been corrected by professional native speakers with master’s and doctoral degrees in the academic domain. The data were constructed in pairs: sentence before receiving proofreading and sentence after receiving proofreading.

The sample data used in the experiment consisted of 1,245 sentences (i.e., 1,245 pairs of sentences were assessed both before and after proofreading), and these sentences were derived from eight academic domains: arts and humanities, biosciences, business and economics, computer science and mathematics, engineering and technology, medicine, physical sciences, and social sciences. Table 1 summarizes the number of sentences applied from each academic domain (Dataset 2).

GLEU-derived datasets

The GLEU metric was used to create four datasets of comparison. The first dataset (Dataset 3), GLEU 1 (T1, P1), compares the correctness of the output sentence text of the Wordvice AI Proofreader (“predicted sentence,” P1) with that of human proofreaders (“ground truth sentence,” T1). The second dataset (Dataset 4), GLEU 2 (T1, P2), compares the correctness of the Wordvice AI Proofreader’s predicted sentence (P1). The third dataset (Dataset 5), GLEU 3 (T1, P2), compares the correctness of MS Word’s predicted sentence (P2). The fourth dataset (Dataset 6), GLEU 4 (T1, P3), compares the correctness of Google Doc’s predicted sentence (P4).

Table 1. Summary of experiment dataset

Subject area	No. of sentences
Arts and humanities	57
Biosciences	54
Business and economics	58
Computer science and mathematics	60
Engineering and technology	52
Medicine	53
Physical sciences	55
Social sciences	56
JFLEG	800
Total	1,245

JFLEG, JHU Fluency-Extended GUG.

Table 2. Comparison of the corrections and improvements of the sentences before correction, the sentences after correction of the comparative methods, and the sentences after the correction by Wordvice AI Proofreader

Correction method	Stylistic improvement		Objective errors	
	Fluency improvement	Vocabulary improvement	Determiner/article correction	Spelling correction
Human editing	Yes	Yes	Yes	Yes
Wordvice AI Proofreader	Yes	Intermediate	Yes	Yes
Google Docs	Yes	No	Intermediate	Yes
Microsoft Word	No	No	Intermediate	Yes

Measurement (evaluation metrics)

Error type comparison

A qualitative comparison was performed on T1, P1, P2, and P3 for categories including stylistic improvement (fluency, vocabulary) and objective errors (determiner/article correction, spell correction). Table 2 illustrates these details for each writing correction method (human proofreading, Wordvice AI, MS Word, and Google Docs).

A GLEU metric [16] was used to evaluate the performance of all proofreading types (T1, P1, P2, and P3). GLEU is an indicator based on the BLEU metric [17] and measures the number of overlapping words by comparing ground truth sentences and predicted sentences with n-gram to assign high scores to sequence words. To calculate GLEU score, we record all sub-sequences of 1, 2, 3, or 4 tokens in a given predicted and ground truth sentence. We then compute a recall (Equation 1), which is the ratio of the number of matching n-grams to the number of total n-grams in the ground truth sentence; we also compute a precision (Equation 2), which is the ratio of the number of matching n-grams to the number of total n-grams in the predicted sequence [18]. Python library (https://www.nltk.org/_modules/nltk/translate/gleu_score.html) was used for the calculation of GLEU.

$$\text{Recall} = \frac{\text{number of matching n-grams}}{\text{number of total n-grams in the ground truth sentence}} \quad (\text{Equation 1})$$

$$\text{Precision} = \frac{\text{number of matching n-grams}}{\text{number of total n-grams in the predicted sentence}} \quad (\text{Equation 2})$$

The GLEU score is then simply the minimum of recall and precision. This GLEU score’s range is always between 0 (no matches) and 1 (complete match). As with the BLEU metric, the higher the GLEU score, the higher the percentage of identified and corrected errors and issues captured by the proofreading tool. These are expressed as a percentage of the total revisions applied in the ground truth model (human-edited text), including objective errors and stylistic issues. The closer to the ground truth editing results, the higher the performance score and the better the editing quality.

Statistical methods: Descriptive statistics were applied for

comparison between the target program and other editing tools.

Results

Quantitative results based on GLEU

Comparison of all automated writing evaluation proofreaders

Table 3 shows the average GLEU score in terms of percentages of corrections made by the Wordvice AI Proofreader and other automated proofreading tools as compared to the ground truth sentences. As an average of total corrections made, the Wordvice AI Proofreader had the highest performance of the Automated Writing Analysis proofreading tools, performing 77% of the corrections applied by the human editor.

Based on the dataset of 1,245 sentences used in the experiment, the proofreading performance of Wordvice AI achieved a maximum of 11.2%P and a minimum of 3.0%P compared to those of Google Docs proofreader. Additionally, the GLEU score of the Wordvice AI-revised text was higher by 13.0%P at maximum on average compared to sentences before proofreading.

Analysis of variance was used to determine the statistical significance of the values. Comparisons made between Wordvice AI, Google Docs, and MS Word proofreading tools revealed a statistically significant difference in proofreading

performance (analysis of variance, $P < 0.05$) (Table 4).

Comparison of Wordvice AI Proofreader and Google Docs proofreading tool

Google Docs proofreader's results scored second in total corrections. Our comparative method confirmed that the deviation of Wordvice AI performance was smaller than that of the performance of Google Docs and MS Word proofreaders.

The proofreading performance of Wordvice AI (with a variation of 5.4%) was more consistent in terms of percentage of errors corrected compared to MS Word (with a variation of 5.6%), but was slightly less consistent than the Google Docs proofreader (with a variation of 5%).

Comparison of Wordvice AI Proofreader and MS Word proofreading tool

We compared the AI Proofreader's performance in the specific academic subject area compared to Google Docs and MS Word, as listed in the Methods section (Tables 3, 5, 6). In each of the eight subject areas, the Wordvice AI Proofreader showed the highest proofreading performance, by total percentage of ground truth sentence corrections applied, at 79.4%. When compared using the GLEU method, MS Word applied the lowest amount of revision and was closest to the original source text in terms of revised to unrevised text. Of the three proofreading tools, MS Word applied the least amount of editing. Table 3 shows the comparison between performance of

Table 3. Percentage of appropriate corrections of all automated proofreaders compared to ground truth sentence (100% correct)

Subject area	Original sentence (%)	Wordvice AI Proofreader (%)	Google Docs (%)	Microsoft Word (%)
Arts and humanities	61.5	78.5	73.2	65.1
Biosciences	62.6	75.7	68.5	64.9
Business and economics	66.5	79.4	68.2	67.1
Computer science and mathematics	65.1	74.5	71.5	66.5
Engineering and technology	64.3	74.1	67.5	65.8
Medicine	61.5	80.5	73.5	62.9
Physical sciences	67.8	78.3	73.4	66.5
Social sciences	65.8	78.1	71.5	68.5
Average	64.4	77.4	70.9	65.9

Table 4. One-way analysis of variance results of proofreading performance analysis between Automated Writing Analysis tools

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F	P-value	F crit
Between groups	0.052960333	2	0.026480167	54.78317838	4.65E-09	3.466800112
Withing groups	0.010150625	21	0.000483363			
Total	0.063110958	23				

Table 5. Comparison of Wordvice AI Proofreader performance to Google Docs proofreader by academic subject area

Subject area	Wordvice AI Proofreader (%)	Google Docs (%)	Difference (%)
Arts and humanities	78.5	72.2	6.3
Biosciences	75.7	70.5	5.2
Business and economics	79.4	69.8	9.6
Computer science and mathematics	74.5	71.5	3.0
Engineering and technology	74.1	68.5	5.6
Medicine	80.5	73.5	7.0
Physical sciences	77.2	73.4	3.8
Social sciences	78.1	72.5	5.6

Table 6. Comparison of Wordvice AI Proofreader performance to Microsoft Word's proofreader by academic subject area

Subject area	Wordvice AI (%)	Microsoft Word (%)	Difference (%)
Arts and humanities	78.5	65.1	13.4
Biosciences	75.7	64.9	10.8
Business and economics	79.4	67.1	12.3
Computer science and mathematics	74.5	66.5	8.0
Engineering and technology	74.1	65.8	8.3
Medicine	80.5	62.9	17.6
Physical sciences	77.2	66.5	10.7
Social sciences	78.1	68.5	9.6

the Wordvice AI Proofreader and Google Docs.

The Wordvice AI Proofreader exhibited a higher performance metric over MS Word in every subject area. As illustrated in Table 4, the Wordvice AI Proofreader outperformed the MS Word proofreader by 17.6%P in the subject area of medicine and by 8.1%P in computer science and mathematics. It also exhibited an 11.4% total average performance advantage over MS Word in each subject area.

Qualitative results

Qualitative results were derived from an open dataset by applying a set of error category criteria (Table 2). These criteria are applied to the input sentences before proofreading, input sentences proofread by MS Word and Google Docs, and sentences proofread by Wordvice AI.

Criteria 1. Fluency improvement (stylistic improvement)

The Wordvice AI Proofreader improved sentence fluency by editing awkward expressions, similar to revision applied in documents edited by editing experts ("human editing"). In Table 7, "point" was used to indicate how different editing applications can interpret the intended or "correct" meaning of words that have multiple potential meanings. In the original

sentence instance, "point" means pointing a finger or positioning something in a particular direction. However, "point out" means indicating the problem, and thus the original term "point" was changed to "point out" by human editing. Because our study considers the sentence revised by human editing as 100% correct, this result accurately conveys the intended meaning of the sentence—here, "point out" is more appropriate than "point."

Google Docs applied the same correction, changing "points" to "points out." However, it did not correct the misspelling "scond," the intended meaning of which human editing recognized as "second." However, Wordvice AI corrected both of these errors perfectly, following the human editor's revisions. MS Word did not detect or correct either error in this sentence.

Criteria 2. Vocabulary improvement (stylistic improvement)

Wordvice AI Proofreader applied appropriate terminology to convey sentence meaning in the same manner as the human editor. Human editing removed the unnecessary definite article "the" from the phrase "the most countries" to capture the intended meaning of "an unspecified majority"; it also

Table 7. Comparative sentence example evaluating fluency improvement

Fluency improvement	Sentence
Original (source text)	Scond, Menzied points that chinese ships in the 1400s used very distinctive anchors that were round stones with a hole in the middle.
Human editing	Second, Menzied points out that chinese ships in the 1400s used very distinctive anchors that were round stones with a hole in the middle.
Wordvice AI Proofreader	Second, Menzied points out that chinese ships in the 1400s used very distinctive anchors that were round stones with a hole in the middle.
Google Doc	Scond, Menzied points out that chinese ships in the 1400s used very distinctive anchors that were round stones with a hole in the middle.
Microsoft Word	Second, Menzies points those Chinese ships in the 1400s used very distinctive anchors that were round stones with a hole in the middle.

Text marked in red denotes incorrect alterations to the input text; text marked in blue denotes correct alterations to the input text.

Table 8. Comparative sentence example evaluating vocabulary improvement

Vocabulary improvement	Sentence
Original (source text)	Unfortunately in the most of the countries the functioning of the public transport is not perfectly organised.
Human editing	Unfortunately in most countries , public transport is not perfectly organised.
Wordvice AI Proofreader	Unfortunately, in most countries , the functioning of public transport is not completely organized.
Google Doc	Unfortunately in most of the countries the functioning of the public transport is not perfectly organised.
Microsoft Word	Unfortunately, in most of the countries the functioning of the public transport is not perfectly organized.

Text marked in red denotes incorrect alterations to the input text; text marked in blue denotes correct alterations to the input text; text marked in pink denotes a style edit to improve clarity or meaning.

Table 9. Comparative sentence example evaluating determiner and article correction

Determiner/article correction	Sentence
Original (source text)	He said in other words that the more flouride may create damage in human body, specifically the bone .
Human editing	He said, in other words , that the more fluoride may create damage to the human body, specifically the bones .
Wordvice AI Proofreader	He said, in other words , that the more fluoride may create damage to the human body, specifically the bones .
Google Doc	He said in other words that the more fluoride may create damage in the human body, specifically the bone.
Microsoft Word	He said in other words that the more fluoride may create damage in human body, specifically the bone.

Text marked in red denotes incorrect alterations to the input text; text marked in blue denotes correct alterations to the input text.

changed the phrase “functioning of the public transport” to “public transport” to reduce wordiness (Table 8).

Similarly, Wordvice AI improved the clarity of the sentence by removing the unnecessary article “the” from the above-mentioned sentence. In addition, Wordvice AI was able to improve the clarity of sentences by inserting a comma and the word “completely,” neither of which revisions were made by human editing. Furthermore, neither Google Docs nor MS Word performed these or any additional revisions to the text.

Criteria 3. Determiner/article correction (objective errors)

In the grammar assessment, Wordvice AI exhibited the same

level of performance as human editing. Table 9 shows that the objective errors identified and corrected by Wordvice AI were the same as those corrected by human editing. A comma is required before the phrase “in other words” to convey the correct meaning, but the comma is omitted in the original. Both the human edit and Wordvice AI edit detected the error and added a comma appropriately.

Additionally, the definite article “the” should be deleted from the original sentence because it is unnecessary in this usage, and both the human edit and Wordvice AI edit performed this revision correctly. Finally, because the human body is not composed of one bone, but multiple bones, the

Table 10. Comparative sentence example evaluating spelling correction

Spelling correction	Sentence
Original (source text)	Lastly, for the economic reason, it is not beneficial becasue the cost of the equipment abd staff required to control fires is very expensive.
Human editing	Lastly, for economic reasons, it is not beneficial because the cost of the equipment and staff required to control fires is very expensive.
Wordvice AI Proofreader	Lastly, for economic reasons, it is not beneficial because the cost of the equipment and staff required to control fires is very expensive.
Google Doc	Lastly, for economic reasons, it is not beneficial because the cost of the equipment and staff required to control fires is very expensive.
Microsoft Word	Lastly, for the economic reason, it is not beneficial because the cost of the equipment and staff required to control fires is very expensive.

Text marked in red denotes incorrect alterations to the input text; text marked in blue denotes correct alterations to the input text.

term “bone” should be revised to “bones.” Both Wordvice AI and the human editor recognized this error and corrected it appropriately. However, Google Docs and MS Word did not detect or correct these errors.

Criteria 4. Spelling correction (objective errors)

The ability to recognize and correct misspellings was exhibited not only by Wordvice AI, but also by all the other proofreading methods we compared (Table 10). In this original sentence, the misspelled word “becasue” should be revised to “because,” and the misspelled word “abd” should be revised to “and.” Each of the proofreading tools accurately recognized the corresponding spelling mistakes and corrected them.

Discussion

Key results: In terms of the accurately revised text, as evaluated by the GLEU metric, Wordvice AI exhibited the highest proofreading score compared to the other proofreading applications, identifying and correcting 77% of the human editor-corrected text. The Wordvice AI Proofreader scored an average of 12.8%P higher than both Google Doc and MS Word in terms of total errors corrected. The proofreading performance of Wordvice AI (variation of 5.4%) was more consistent in terms of percentage of errors corrected compared to MS Word (variation of 5.6%) but was slightly less consistent than Google Docs (variation of 5%). These results indicate that Wordvice AI Proofreader is more thorough than these other two proofreading tools in terms of the percentage of errors identified, though it does not edit stylistic or subjective issues as extensively as the human editor.

Additionally, Wordvice AI Proofreader exhibited consistent levels of proofreading among all academic subject areas evaluated in the GLEU comparison. Variability in editing perfor-

mance among these subject areas was also relatively small, with only a 6.4%P difference between the lowest and highest average editing applied compared to the human proofreader. Both Google Docs and MS Word exhibited similar degrees of variability in performance throughout all subject areas. The highest percentage of appropriate corrections recorded for these automated writing evaluation proofreaders (Google Docs: medicine 73.5%) was still lower than Wordvice Proofreader’s lowest average (medicine 80.5%).

Interpretation: The Wordvice AI Proofreader identifies and corrects writing and language errors in any main academic domain. This tool could be especially useful for researchers writing manuscripts to check the accuracy of their writing in English before submitting their draft to a professional proofreader, who can provide additional stylistic editing and. NLP applications like the Wordvice AI Proofreader may exhibit greater accuracy in correcting objective errors than more widely-used applications like MS Word and Google Docs before the input text is derived primarily from academic writing samples. Similar AI proofreaders trained on academic texts (such as Trinka and Ginger) may also prove more useful for research authors than general proofreading tools such as Grammarly, Hemingway Editor, and Ginger, among others.

Suggestion of further studies: By training the software with more sample texts, the Wordvice AI Proofreader could potentially exhibit performance and accuracy levels even closer to those of human editors. However, due to the current output limits of NLP and AI, human editing by professional editors remains the most comprehensive and effective form of text revision, especially for authors of academic documents, which require the understanding of jargon and natural expressions in English.

Conclusion: In most of the texts analyzed, the Wordvice AI Proofreader performed at or near the level of the human edi-

tor, identifying similar errors and offering comparable suggestions in the majority of sample passages. The AI Proofreader also had higher performance and greater consistency than the other two proofreading applications evaluated. When used alongside professional editing and proofreading to ensure natural expressions and flow, Wordvice AI Proofreader has the potential to improve manuscript writing efficiency and help users to communicate more effectively with the global scientific community.

Conflict of Interest

The authors are employees of Wordvice. Except for that, no potential conflict of interest relevant to this article was reported.

Funding

The authors received no financial support for this study.

Data Availability

Dataset file is available from the Harvard Dataverse at: <https://doi.org/10.7910/DVN/KZ1MYX>

Dataset 1. Eight hundred sentence pairs out of 1,501 from JHU Fluency-Extended GUG (JFLEG) open dataset, which were used for assessing improvements in textual fluency (<https://github.com/keisks/jfleg>).

Dataset 2. Four hundred forty-five sentences from eight academic domains, derived from Wordvice's academic document data: arts and humanities, biosciences, business and economics, computer science and mathematics, engineering and technology, medicine, physical sciences, and social sciences.

Dataset 3. One thousand two hundred forty-five sentences composed of 800 JFLEG data and 445 academic sentence data edited by human editing experts.

Dataset 4. One thousand two hundred forty-five sentences composed of 800 JFLEG data and 445 academic sentence data edited by Wordvice AI.

Dataset 5. One thousand two hundred forty-five sentences composed of 800 JFLEG data and 445 academic sentence data edited by MS-Word.

Dataset 6. One thousand two hundred forty-five sentences composed of 800 JFLEG data and 445 academic sentence data edited by Google Docs.

References

1. Warschauer M, Ware P. Automated writing evaluation: defining the classroom research agenda. *Lang Teach Res* 2006;10:157-80. <https://doi.org/10.1191/1362168806lr190oa>
2. Daudaravicius V, Banchs RE, Volodina E, Napoles C. A report on the automatic evaluation of scientific writing shared task. Paper presented at: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications; 2016 Jun; San Diego, CA, USA. p. 53-

- 62.
3. Grammarly [Internet]. San Francisco, CA: Grammarly; 2021 [cited 2021 Aug 20]. Available from: <https://www.grammarly.com/>
4. Ginger Grammar Checker [Internet]. Lexington, KY: Ginger Software; 2021 [cited 2021 Aug 22]. Available from: <https://www.gingersoftware.com/grammarcheck>
5. Hemingway Editor [Internet]. Durham, NC: 38 Long LLC; 2021 [cited 2021 Aug 22]. Available from: <https://hemingwayapp.com/>
6. Leacock C, Chodorow M, Gamon M, Tetreault J. Automated grammatical error detection for language learners [Internet]. Williston, VT: Morgan & Claypool Publishers; 2010 [cited 2021 Aug 22]. Available from: <https://doi.org/10.2200/S00275ED1V01Y201006HLT009>
7. Montgomery DJ, Karlan GR, Coutinho M. The effectiveness of word processor spell checker programs to produce target words for misspellings generated by students with learning disabilities. *J Spec Ed Tech* 2001;16:27-42. <https://doi.org/10.1177/016264340101600202>
8. Dale R, Viethen J. The automated writing assistance landscape in 2021. *Nat Lang Eng* 2021;27:511-8. <https://doi.org/10.1017/S1351324921000164>
9. Perelman L. Grammar checkers do not work. *WLN J Writ Cent Scholarsh* 2016;40:11-9.
10. Napoles C, Sakaguchi K, Post M, Tetreault J. Ground truth for grammatical error correction metrics. Paper presented at: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing; 2015 Jul 26-31; Beijing, China. p. 588-93.
11. Wordvice AI Proofreader [Internet]. Seoul: Wordvice; 2021 [cited 2021 Aug 23]. Available from: <https://wordvice.ai>
12. Google Docs [Internet]. Mountain View, CA: Alphabet; 2021 [cited 2021 Aug 22]. Available from: <https://www.google.com/docs/about/>
13. Microsoft Word [Internet]. Redmond, WA: Microsoft; 2021 [cited 2021 Aug 22]. Available from: <https://www.microsoft.com/microsoft-365/word>
14. Napoles C, Sakaguchi K, Tetreault J. JFLEG: a fluency corpus and benchmark for grammatical error correction. *arXiv:1702.04066 [cs.CL] [Preprint]*. 2017 [cited 2021 Aug 22]. Available from: <https://arxiv.org/pdf/1702.04066.pdf>
15. Sakaguchi K, Napoles C, Post M, Tetreault J. Reassessing the goals of grammatical error correction: fluency instead of grammaticality. *Trans Assoc Comput Linguist* 2016;4:169-82. https://doi.org/10.1162/tacl_a_00091
16. Mutton A, Dras M, Wan S, Dale R. GLEU: automatic evaluation of sentence-level fluency. Paper presented at: Proceedings of the 45th Annual Meeting of the Association of

- Computational Linguistics; 2007 Jun; Prague, Czech Republic. p. 344-51.
17. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. Paper presented at: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2002 Jul; Philadelphia, PA, USA. p. 311-8. <https://doi.org/10.3115/1073083.1073135>
18. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144v2 [cs.CL] [Preprint]. 2016 [cited 2021 Aug 19]. Available from: <https://arxiv.org/pdf/1609.08144v2.pdf>