Towards Universal Accessibility on the Web: Do Grammar Checking Tools Improve Text Readability?

Hitesh Mohan Kaushik[1], Evelyn Eika[1] and Frode Eika Sandnes[1,2][0000-0001-7781-748X]

[1] Oslo Metropolitan University, 0130 Oslo, Norway
[2] Kristiania University College, 0153 Oslo, Norway
hmkaushik@yahoo.com,evelyn.eika@oslomet.no,frodes@oslomet.no

Abstract. Readable text is a key ingredient in a universally accessible web. WCAG2.1 recommends that text should be readable by someone with basic schooling, a criterion that is hard to quantify and implement. Writers rely on qualitative clear-language recommendations, their own experience, and tools. This study set out to investigate if one class of such tools, automatic grammar checkers, has a measurable effect on the readability of text. A controlled experiment was conducted employing 15 participants who brought a piece of their own writing to the experiment tasked with improving the text using a grammar checker. Changes in readability of the text before and after applying the grammar tool were measured. Results show that there were significant reductions in error rates by applying the grammar tool, while there were no significant effects on readability. The results suggest that other automatic tools beside grammar checkers are needed to improve readability. These results have implications for web content providers.

Keywords: universal accessibility, readability, web texts, grammar checkers, clear language, writing assessment.

1 Introduction

With increased access to digital devices, information is more available than ever, governments have been pushing towards digital societies. While it is the fastest medium to share information on the internet, it should also be of concern to make that information accessible to everyone [45]. Acknowledging this, US government in 1998 announced a plan for implementing a system of plain language for the writing of government regulations. The objective of plain language is to make regulations clearer and easier for the average person to understand [37]. According to WCAG2.1 guideline 3.1, content providers should "make text content readable and understandable" [52]. Guideline 3.1.5 addresses the reading level and implies that the limit is centered around "lower secondary education level". When a text requires the reader to have an education level beyond this, content providers are recommended to offer an alternative simplified version of

This is a post-peer-review, pre-copyedit version of a conference proceeding published in HCII 2020: Universal Access in Human-Computer Interaction. Design Approaches and Supporting Technologies, 14th International Conference, UAHCI 2020 Proceedings, Part I, which is part of the Lecture Notes in Computer Science book series (volume 12188). The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-49282-3_19

the texts. This is challenging [15-166] as it is both difficult to quantify and also language dependent. WCAG2.1 also addresses unusual words (3.1.3) and abbreviations (3.1.4).

According to Statistics Norway 2019, about 34.1% of the Norwegian population above 16 years of age had higher education. In the United States 32.2% of all adults of age 18 and above had higher education as per 2018. As per 2011 census, only 6.7% of the population in India attained higher education. In order to prevent information discrimination in society, it is important that the information is accessible and succinct. Additionally, governments and organizations such as hospitals and businesses are also providing online services. Meade and Smith [38] described the importance of readable and understandable texts in healthcare. Text provides vital information, including advice on how to prevent unhealthy habits and actively participate in diagnosis processes. Health-care summaries that are difficult to read prevent patients from becoming active and responsible partners that make informed conscious decisions.

Quality writing is a challenging task, be it for academics or professionals, for business or private communications [13]. Graduates often lack necessary writing skills for business across disciplines including public relations, journalism, and communication [25]. Ideally, a good quality text is comprehensible, readable, and communicative. Words on paper help establish a bond between sender and receiver, and this bond breaks if the writer attempts to persuade a reader with poorly structured sentences that fail to forge trust and create coherence [25]. It is essential to address the different factors that improve text quality and readability. These include vocabulary, sentence structure, subject verb agreement, use of correct tense, content, and other grammar conventions. Improved writing skills help the writer to express the ideas more clearly and accurately. Writing is intended for sharing information; it might be in the form of a personalized letter, examination paper, published news, or a research article. It is the responsibility of an author to ensure that the text is legible and easily understood by the target group, such as a newspaper article which has the general public as an audience. This large audience includes readers with different backgrounds, education, literacy levels, disabilities, and people with English as a foreign language.

Although the notion of readability is relatively easy to comprehend as "easy to read", what exactly constitutes a readable text in practice is less obvious. Klare [32] defines readability as the level of difficulty of written text. Various factors affect text readability such as vocabulary, sentence length, semantics, readers' area of interest [24], education level and experience. Usage of widely known words and shorter sentences help make texts easier to read and understand [34]. Readability refers to how much sense the words and sentences make to readers, how clear the vocabulary and grammar are [7]. Readable texts benefit not only persons with limited education but also readers with learning disorders, cognitive disabilities, dyslexia, and ADHD [3, 4, 22]. Highly readable and concise texts are also beneficial for visually impaired readers who rely on screen reading technologies as it provides these users with a more rapid access to the texts [31].

A grammar checker is a type of writing aid. Web content providers may deploy such tools to improve content readability. Word-processors such as Microsoft Word come equipped with simple integrated spelling and grammar checkers. There are also third

party specialized commercial grammar checkers such as Grammarly and Ginger. Open source tools such as Language Pack provided the Open Office family of word-processors. The rationale for this study was to investigate if such tools have a measurable effect on the actual readability of texts. One of the widely used tools, Grammarly Premium claims to offer over 400 types of checks. It checks grammatical errors, provides vocabulary enhancement suggestions, detects plagiarism, and provides citation suggestions [19]. Moreover, Grammarly claims to provide writer support for improving the readability of the text by reducing sentence length and employing simple and exact words in all contexts [19]. Other products include GingerSoftware and WhiteSmoke, with similar claims of helping writers write better English. Based on this we formulated the following research questions:

1. Do grammar tools help improve readability?
2. Do writers' self-assessed writing abilities correspond with their actual writing abilities?
3. Do grammar tools help writers learn about writing?

2 Related work

2.1 Readability Formulas

Dale [111] discussed three aspects affecting readability, namely typography, readers' interest, and writing style. Typography refers to the choice of font, text size, text color, background color, spacing, line length, and line spacing. Typography is concerned with both the legibility of the text and its aesthetics. Legibility and readability concern the speed at which users can read the printed matter. Color contrast has been an important issue for readable texts on the web [6, 23, 42, 46-50]. Interest regards what grabs the readers' attention. Gilliland [18] inferred that readability when studied as interest leads to the analysis of subject matter and themes preferred by specific groups of readers. The style of writing concerns what types of vocabulary, sentence structure, and other expressional elements best suit the abilities of readers.

In order to determine difficulty level, three widely cited readability formulas were developed: the Flesch–Kincaid readability tests, the Dale-Chall readability formula, and the Gunning FOG Index. Flesch [17] introduced the two-part readability formula. The first part, the reading ease formula, uses only two variables, the number of syllables and the number of sentences for each 100-word sample. It predicts reading ease on a scale from 1 to 100, with 30 being very difficult and 70 being easy. The second part of Flesch's formula predicts human interest by counting the number of personal words (e.g., pronouns and names) and personal remarks (e.g., quotes, exclamations, and incomplete sentences). Dale-Chall formula was designed to correct certain shortcomings in the Flesch Reading Ease formula. It uses a sentence-length variable plus a percentage of hard words not found on the Dale-Chall list of 3,000 easy words, of which 80 percent are known to fourth-graders. The Gunning FOG Index [21] uses two variables: average sentence length and the number of words with more than two syllables per 100 words. Mc Laughlin [36] deduced that readability could be expressed as a relationship between

two variables which are measures of the difficulty experienced by people reading a given text and a measure of the linguistic characteristics of that text. They proposed the SMOG readability formula which was derived using regression analysis. SMOG was intended to eliminate the problem in existing formulas where one long word or sentence affects the readability of an easier text more than it will of a harder text.

To study the independent impact of different text attributes on readability, Pitler and Nenkova [43] identified six factors affecting readability: word length and sentence length, vocabulary, syntactic features, lexical coherence, entity coherence, and discourse relations. Of these, vocabulary and discourse relations had the strongest impact, followed by the average number of verb phrases and text length. The authors claimed that using word length and sentence length were less effective than the other features, while using a combination of all features produces the best results.

2.2 Readability beyond Formulas

Traditional readability formulas are regarded too simplistic and possibly do more harm than good as they do not consider other factors such as vocabulary, grammar, and background knowledge [**Error! Reference source not found.**, 51]. Wright [54] pointed out that readability formulas do not consider key factors such as document type, layout, acronyms, and abbreviations. Further, some longer words (e.g., *understanding*) are given low readability scores with the formulae but are easy to read, while short but less frequent words (e.g., *grasp*) are given a higher readability score but may be harder to read. In addition, proper nouns such as people's and place's names should not count negatively towards readability. A standard test of readability measure is how well its prediction matches with readers' actual comprehension using existing texts [2].

2.3 Readability on the Web

Jatowt and Tanaka [29] compared readability of three websites, namely Wikipedia, simple Wikipedia, and Britannica. They used both syntactical (Flesch Reading Ease) and familiarity-based approaches (New Dale-Chall formula) to determine the readability index. They found that the average word and sentence lengths were much higher on Wikipedia texts compared to those of simple Wikipedia. Britannica was also easier to read compared to Wikipedia. The study suggested that Wikipedia's emphasis on accuracy and coverage may have reduced readability compared to the other resources.

To study the impact of typographic features on readability, Yu and Miller [55] introduced a Firefox Extension named Froggy. The extension removed distractions from the web pages in the form of advertisements and transformed the text into a more readable format. The participants were positive towards the Jenga format, and they considered it easier to read and understand compared to the standard format. There was also a slight improvement in comprehension without affecting reading speed.

Chung et al. [10] focused on simplifying the text on news websites for improved readability for deaf people. They developed an online news display system simplifying syntactic structures and providing graphical representations. The system simplified complex sentences by identifying embedded clauses, and relocating them for simpler

structure, and then visually presenting the relationships among clauses. The evaluation showed that tested sentences were mostly correctly restored. More than half of the erroneous sentences were false relocations of adverbial clauses. The participants responded positively concerning system adequacy.

2.4 Evaluating writing

Writing assessment can be implemented based on holistic or analytic scales. The holistic evaluation involves reading to gain an overall impression of a writer's skill [9]; in contrast, the analytic scoring involves an itemized analysis to help identify weaknesses in a student's writing [33]. In holistic scoring, the rater makes an overall judgment concerning the quality of performance. In analytic scoring, the rater assigns a score to each of the dimensions being assessed [30]. These evaluations are often conducted using scoring rubrics to help analyze writing in a reliable and consistent manner [39]. A well-established scoring scheme for writing assessment included five categories: content, organization, vocabulary, language use, and mechanics [26]. Weigle [53] described this as one of the best known and most widely used analytic scales in ESL.

2.5 Grammar checkers and efficacy

Schraudner [40] examined the role of automated correction tools as a teacher's assistant to supplement efficient learning for English language learners. The participants were asked to weekly read a portion of a book and electronically respond with an explanation of the content. The results showed that commonly occurring errors were related to punctuation, conjunctions, and pronouns. Grammarly's category for sentence structure found direct translation errors. Sentence structure was created as object-subject-verb or subject-object-verb. The tools were deemed useful for the students' learning and planning; in particular, word choice, word frequency, and spelling were easy to monitor and target, except handling irregular past tense verbs. The tools helped improve learners' lexical abilities and the use of punctuations and prepositions.

Dale [12] conducted a comparative study of ten proofreading systems including Grammarly, Ginger, ProWritingAid, ClearEdits, Editor, Correct English, GrammarBase, GrammarCheck.net, SpellCheckPlus, and Style Writer 4. The study observed that the performance of these programs was unsatisfactory. Grammarly and Ginger performed better than the others.

Cavaleri and Dianati [8] surveyed the students' perceptions of Grammarly use in writing assignments. Students had mostly positive feedback with a few exceptions. The survey indicated long-term benefits as explanations and hints were helpful in understanding grammar rules. Some responded that they would only use Grammarly for proofreading. The grammar mistakes detected appeared to be minor and could have been resolved if they had read them carefully themselves. Some Grammarly recommendations were deemed incorrect or unclear.

Oneill and Russell [41] explored the role of feedback on grammar for those using Grammarly. The results were compared for those who received automatic advice from Grammarly and those who manually received advice from the advisors. The students

who received non-Grammarly advice were satisfied, but the students with Grammarly advice were strongly satisfied. Students' experience was largely positive regarding the use of Grammarly; they claim it improved their confidence. The major concern was the accuracy of the feedback. They also identified issues with passive voice, complex sentences, and vocabulary choices. Some students were not satisfied with the performance of Grammarly and even preferred feedback from MS Word over Grammarly. Students' prior knowledge of the English language also had effect on their responses. Students with a low IELTS score were highly satisfied with Grammarly. Students with the lowest scores who studied English at the university were most critical of Grammarly.

3 Method

3.1 Experimental design

A controlled experiment was configured with a pre-test/post-test design. Automatic grammar tool was the independent within-groups variable with two levels, namely without tool and with tool. The dependent variables included ratio of errors per word, Grammarly readability score, Gunning Fog Index, SMOG readability score, and raters' language scores.

3.2 Participants

Fifteen participants were recruited; six were female and nine were male. All were students at Oslo Metropolitan University, randomly selected from graduate programs where English was the language of tuition. Most participants were in the 27 to 34 age range, with the youngest being 21 and the oldest close to 50. The participants had minimum qualification of English proficiency with experience in academic writing. Participants either had at least a TOEFL score of 90 or had been studying English for more than 13 years. None of the participants reported having any reading disabilities such as dyslexia.

Eleven participants reported having English as their second language (L2), three participants reporting having English as their third language (L3), while one participant reported using English as the fourth language (L4). Nine participants heard about readability, while six participants were unfamiliar with this term. Moreover, ten participants had heard about grammar checking tools such as Grammarly, while five reported that they had not heard about such tools.

3.3 Equipment

A Thinkpad laptop computer with screen size of 13 inches and touchpad was used for the experiments with Microsoft Word word-processor and a full version of Grammarly installed. An initial test revealed that there were minimal functional differences between Grammarly and Ginger, and Grammarly was chosen as it appears to have a larger

market share. Grammarly also has its own built-in readability metric. The screen activity was recorded using ShareX video recording software.

3.4 Task

The participants were asked to bring a recently written text document on a USB stick. This text was to be academic coursework as part of their studies. These texts were used as a pre-experiment sample. Participants' own writing was chosen to make the experiment more engaging and help motivate the participants as they had a chance to improve their own writing.

The grammar-checking task involved editing their own text using the feedback from the grammar tool. The grammar tool would suggest simple grammar corrections and other changes such as restructuring sentences to improve clarity. Participants had the option of accepting suggestions, reject suggestions, or edit the text based on the feedback.

3.5 Procedure

Potential participants were contacted personally and briefed about the purpose and tasks for this project. A suitable time to perform experiment was agreed upon, which allowed potential participants extra time to decide about their participation.

The participants conducted the tasks individually at the university campus in a meeting room, which reduced the influences of external noise and ensured constant conditions for each participant. This setup was controlled in a natural environment since all the participants were students at the university. Participants were given a copy of consent form explaining the purpose of the study. Consent was given orally. The experiment was completed in a single session for each participant. Each session started with a questionnaire to gather information about the participant's understanding of readability and their self-assessment of proficiency in written English (How much help is needed with vocabulary, grammar, sentence formation, punctuations, content, writing style, and voice). It also included questions about common writing issues and any prior experiences with grammar checker software.

Next, the text sample was loaded, and the participant was asked to run Grammarly on the text sample they had brought to the session. The participants were asked to use the feedback and improve the text as per their understanding. Participants were encouraged to work independently, but on-spot guidance was provided. The computer screen was recorded without audio enabled to facilitate in-depth analyses of the users' interaction with the grammar tool.

After completing the grammar-check, the participants were asked to answer a post-test questionnaire about their overall experience and their opinions about the problems identified and corrections suggested by the software.

Participation was voluntary and anonymous. The participants' identities were not recorded and the meta-information in the participants text documents were deleted. Therefore, no personal identifying information was stored, and the General Data Protection Regulations (GDPR) did not apply.

3.6 Observations

The number of errors in the documents was reported by Grammarly and verified by a review of the screen recordings. The ratio of errors per word per writing was computed by taking the number of errors reported divided by the number of words in the text. This allowed the error scores to be compared across writings with varying lengths. The Gunning FOG Index was calculated using an online tool (http://gunning-fog-index.com/fog.cgi) and the Grammarly readability scores were provided directly by the software tools. The SMOG scores were computed using online readability checker tool (https://readabilityformulas.com/).

The texts were also manually assessed based on a scoring rubric of five criteria. A scoring rubric guides assessors what features to scrutinize as they read; these descriptors are useful because they give evaluators a sense of what aspects of a student's writing should be critiqued [30, 33]. Style, vocabulary, grammar, mechanics, and clarity are the criteria identified for evaluation. These categories are tweaked version of scoring rubric from [26], which is widely used in its original or adapted form [5, **Error! Reference source not found.**, 27, 33, 35, 44]. This study excludes the content criterion and uses analytic scoring as content evaluation involves testing subject-specific knowledge, which is not affected by grammar checker tools.

3.7 Analysis

The observations were analyzed using the statistical analysis software JASP version 0.11.0.0 [28].

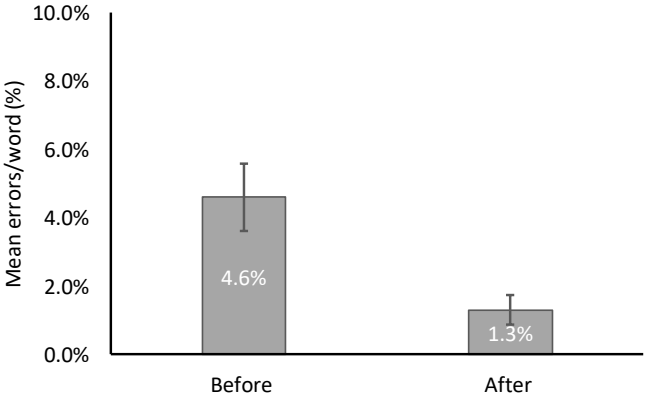

Fig. 1. The mean number of errors before and after applying the grammar checker. Error bars show 95% confidence intervals.

4 Results

Shapiro Wilks tests revealed that the Grammarly readability scores ($W = 0.433$, $p < .001$) and Gunning FOG scores ($W = 0.769$, $p = .002$) were not normally distributed,

while the ratio of errors per word and SMOG measurements did not deviate from the normal distribution. The pre-check and post-check scores for Grammarly readability and Gunning FOG were therefore analyzed using non-parametric procedures, while the errors per word and SMOG measurements were analyzed using paired t-tests.

A paired t-test test revealed that grammar checking had a significant effect on the percentage of errors per word ($t(14) = 6.437$, $p < .001$, Cohen's $d = 1.662$), as there were nearly three times the percentage of errors per word in the pre-checked texts ($M = 4.6$, $SD = 1.8$) compared to the post-checked texts ($M = 1.3$, $SD = 0.8$). In terms of percentage, the results show that the participants followed about one third of the advice provided by the grammar tool ($M = 31.8$, $SD = 18.3$) but rejected the remaining advice. It is also worth noting that the rate of error per word prior to grammar checking correlated strongly with language levels (L2, L3, and L4) of the participants ($r(15) = .656$, $p = .008$), confirming that less experienced learners made more mistakes than more experienced learners.

No significant effect of the grammar checker could be observed for the Grammarly readability score ($W = 15.0$, $p = .93$); the scores before checking ($M = 38.7$, $SD = 10.7$) were marginally larger than the scores after checking ($M = 37.2$, $SD = 12.4$). Similarly, no significant effect of the grammar checker could be observed on the Gunning FOG index ($W = 39$, $p = 1.0$); the scores before checking ($M = 15.157$, $SD = 2.319$) were nearly the same as the scores after checking ($M = 15.153$, $SD = 2.33$).

There was also no significant effect of the grammar tool on the SMOG measures ($t(14) = 1.317$, $p = .209$) as the mean pre-check score ($M = 11.693$, $SD = 1.65$) was nearly identical to the mean post-check score ($M = 11.633$, $SD = 1.616$). The SMOG score did not correlate with the participants' self-reported English writing skills; however, the SMOG readability scores correlated positively ($r(15) = .524$, $p = .045$) with the participants' language level (L2, L3, or L4).

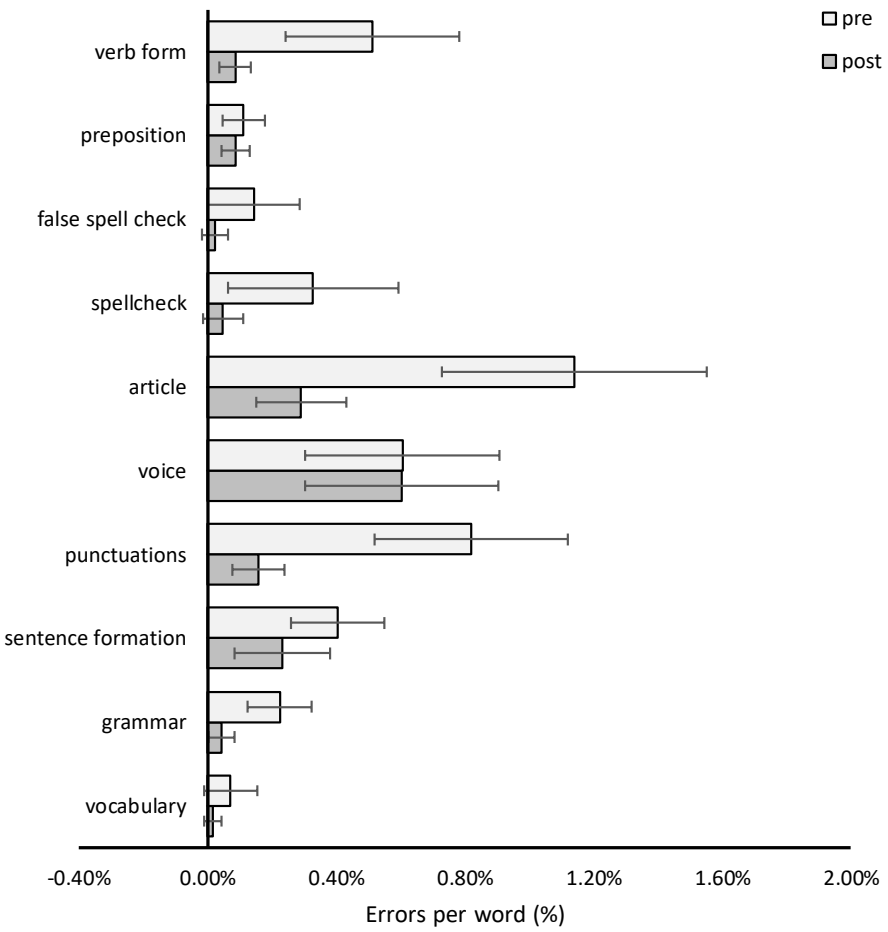

Fig. 2. Distribution of mean errors types before and after grammar check. Error bars show 95% confidence intervals.

Fig. 2 shows the distribution of error types before and after applying the grammar tool. As can be seen, verb form, article, and punctuations are error categories that were effectively eliminated by the grammar tool. Voice, preposition, and vocabulary were associated with only minimal improvements. Fig. 2 also shows that the spread in ratio of errors was generally smaller for the texts after the grammar check.

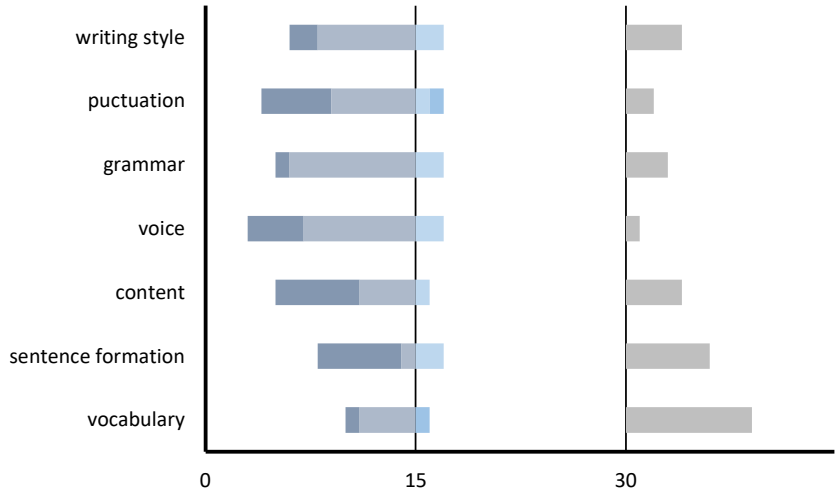

Fig. 3. Diverging stacked bar chart showing participants' self-reported writing abilities. The Grey-blue left bars show the number of negative responses and the blue right bars show positive responses. The grey bars on the right show neutral responses.

Fig. 3 shows the participants' self-reported English writing abilities. Overall, all the responses strongly leaned towards the negative side with the fewest positive ratings of participants' own abilities in terms of content. The largest number of negative responses was associated with voice, while vocabulary was associated with the fewest negative responses. Clearly, voice was the feature of writing with the fewest neutral responses while vocabulary was associated with the largest number of neutral responses.

We also correlated the participants' self-reported writing abilities with the objective readability metrics. The only significant positive correlation was observed between writing style and the Grammarly readability score ($r_s(15) = .553$, $p = .033$) and a significant negative correlation between writing style and the Gunning FOG scores ($r(15) = -.514$, $p = .05$). Participants' self-reporting writing style correlated positively with the error rate after grammar checking ($r_s(15) = .556$, $p = .031$).

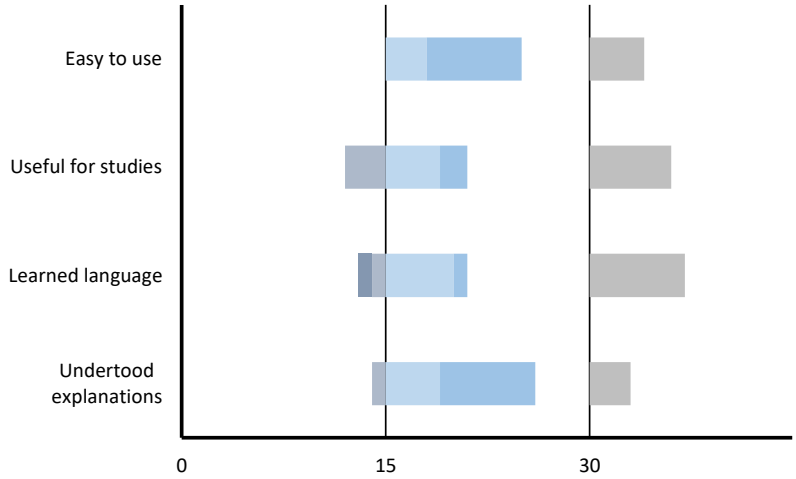

Fig. 4. Diverging stacked bar chart showing the participants' subjective opinions about the grammar-checking tool. The Grey-blue left bars show the number of negative responses and the blue right bars show positive responses. The grey bars on the right show neutral responses.

Fig. 4 shows the participants' perceptions about the grammar-checking tool. Unlike the participants' rather pessimistic rating of their own writing abilities, the perceptions of the grammar tool were positive with all results strongly leaning towards positive responses. None of the participants reported that they found the tool difficult to use, and only one participant reported that they did not understand the explanations given by the tool. Although also tending towards positive responses, the two questions related to learning (i.e., if the tool would be useful for their studies and if they learned language using the tool) were associated with mostly neutral responses.

Correlation analyses show that the participants' responses to the question about whether the grammar tool helped the text correlated positively with the percentage of advice followed ($r_s(15) = .533$, $p = .041$). The error rate before correlated negatively with the participants' perception of how helpful the tool was for improving text ($r_s(15) = -.688$, $p = .005$). The error rate before also correlated negatively with the participants' perception of how relevant the advice provided by the tool was ($r_s(15) = -.560$, $p = .030$).

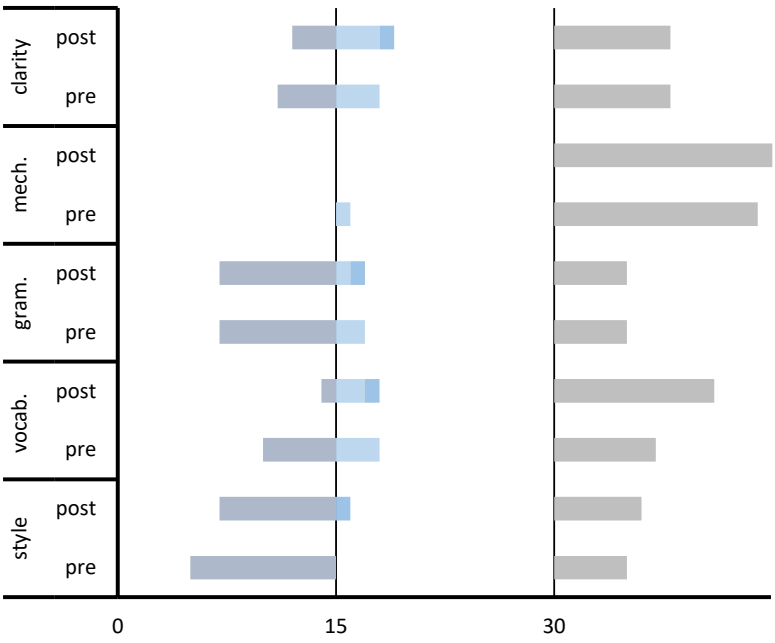

Fig. 5. Diverging stacked bar graph showing the median scores based on the three raters for the five-item scoring rubric (style, vocabulary, grammar, mechanics, and clarity). The Grey-blue left bars show the number of negative responses and the blue right bars show positive responses. The grey bars on the right show neutral responses.

Fig. 5 shows the median ratings of the texts based on the three raters. The median ratings of the three raters for each text were used in the analysis as it was a more robust measure than the mean. Analyses show that there was only a significant improvement effect of the grammar checker in terms of vocabulary and language ($W = 0.0$, $p = .018$) which started with a lower mean score ($M = 2.867$, $SD = .743$) and it ended with a higher score ($M = 3.2$, $SD = .676$) after the grammar checking. The diverging stacked bar graph in Fig. 4 shows that most of the scores on the negative side of the scale became neutral. A visible improvement can also be spotted for clarity and style, although these improvements are not statistically significant. Moreover, Fig. 4 shows that there were very little change in terms of mechanics and grammar. In fact, the post-grammar check scores for mechanics were slightly lower than the pre-check scores.

The median ratings of the texts before checking were correlated with the participants' self-assessed writing abilities. Style was found to correlate negatively with sentence formation ($r_s(15) = -.543$, $p = .036$) and positively with voice ($r_s(15) = -.557$, $p = .031$), and clarity was correlated negatively with sentence formation ($r_s(15) = -.686$, $p = .005$). There were no significant correlations between the automatic readability indices (FOG, SMOG, and Grammary readability score) and the manual ratings of the texts.

Table 1. Inter-rater agreement based mean Spearman correlations (ρ).

	Pre	Post
Style	0.36	0.20
Vocabulary	0.14	0.12
Grammar	0.36	0.22
Mechanics	0.30	-0.33
Clarity	0.14	0.12

The overall inter-rater agreement was 0.16 which is very low. Table 1 lists the detailed inter-rater agreements for the individual rubrics in the pre and post conditions. The inter-rater agreement was computed using the mean Spearman correlations of all three rater-pair combinations as the ratings were ordinal Likert values. Clearly, the raters agree more on the pre-checked text compared to the post-checked texts. Next, the agreements were higher for style and grammar ($\rho = 0.36$) and lower for vocabulary and clarity ($\rho = 0.14$).

5 Discussion

The significant effect of a grammar checker on the reduction of errors is as one would expect, as the purpose of a grammar checker is to identify and help correct errors. Similarly, the positive effect of the grammar tool on the writers' vocabulary improvement is as expected and as also reflected in the raters' assessment since the grammar tool suggests alternative words and phrases. Readability, or the lack of readability, on the other hand is not an error; the results clearly show that there is no significant effect of the grammar tool in terms of readability. One could argue that errors and readability represent two perpendicular dimensions: It is possible to envisage a grammatically correct and error free text that is very hard to read, and a text that is very easy to read but with many trivial grammar and spelling mistakes. In other words, our results do not give support to the claims made by the tool developers that the tool helps improve readability.

It is interesting to observe that most of the participants had a negative perception of their own writing abilities, while they had an overall positive perception of the grammar tool. It would be interesting to also contrast the results with a cohort of native English speakers. Perhaps we would observe the opposite pattern, namely a positive perception of their own writing ability and a negative perception of the tools. Put differently, those who are aware of their shortcomings may be more perceptible to assistance compared to those who may not have such shortcomings and hence will not find the tools valuable.

The danger of this positive perception of the grammar tool is that they may give a false sense of security for learners of English, especially because of the advertising claim that the tools help improve readability. Users may perceive that the texts are improved as many changes are suggested and feedback is provided. However, the elimination of grammar errors should not be mistaken for readability enhancement.

The fact that the percentage of advice followed was far from 100% suggests that participants do not blindly follow the advice provided by the tool. They make individual assessments of the suggestions and reject some proposals. This is an encouraging result. The positive connection between the perceived helpfulness of the tool to improve the text and the percentage of advice followed shows that participants who followed more advice were also more satisfied with the tool. Most of the participants gave positive feedback about the tool, while one of them was highly critical and did not find it effective. Six of the participants raised the issue of false-positives and said they were overwhelmed with the amount of errors reported. Three participants also noted that the tool lacks technical vocabulary.

It is not clear as to why the number of errors before the grammar check correlates negatively with the participants' perception of the helpfulness of the tool and the relevance of the advice. This result seems to suggest that participants who made fewer errors initially had a more positive perception of the tool than participants who made many mistakes initially.

Clearly, grammar tools did not help with improving readability. It seems that content providers still need to rely on language expertise and manual editorial review work on their content in order to ensure universally accessible text with high level of readability.

It is also interesting to observe that the SMOG scores correlated with the participants' reported language level (L2, L3, L4) while both FOG and Grammarly scores correlated with the participants' self-reported writing style. It would be worthwhile to explore this in more detail, but it may be advisable to base any speculation on a larger sample of participants.

6 Limitations

This study was conducted with a relatively small sample of participants. The observations of the dependent variables did not adhere to a normal distribution. It could be that a larger sample would yield normally distributed observations. If so, parametric testing procedures could be applied. Although the cohort is narrow in the sense that it only included students, the cohort was still quite wide in that it included both undergraduate and postgraduate students from a range of disciplines. Our goal was to narrow the cohort to a single class, but not enough participants volunteered to participate.

Another potential limitation lies in the texts provided by the participants. These were on a diverse set of topics and contexts. Although this was a necessary practical adaptation, it would have been beneficial if these texts were on the same topics and contexts.

The inter-rater agreement appears to be low, suggesting that the devised manual rating procedures need to be further refined. One problem may have been that some of the rubrics overlapped and no instructions were given on what the various levels of the scales meant, leaving the scoring up to personal interpretation. Moreover, the different cultural and academic backgrounds of the raters could have been an influencing factor on the inter-rater agreement. Also, as the manual ratings did not correlate with the automatic readability measures, less emphasis has been placed on the manual rating re-

sults. Finally, this study was based on English, while the WCAG readability requirement is language neutral. Hence, it may differ concerning how to approach readability in various languages.

7 Conclusions

This paper investigated the effects of the grammar checking tool on the readability of texts. As expected, the results reveal that the grammar tool has a positive effect on reducing the ratio of errors. However, no significant effect on readability is detected. Our results therefore do not agree with the claims that these tools help readability. It is recommended that web content providers do not rely on grammar checking tools to ensure readability. Content providers are advised to consult manual editorial work involving language competences in order to contribute towards universal accessibility of web contents.

References

1. Bailin, A., Grafstein, A.: Grammar and Readability. Readability: Text and Context, pp. 65-96. Springer (2016).
2. Benjamin, R. G.: Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. Educational Psychology Review **24**, 63-88 (2012).
3. Berget, G., Sandnes, F. E.: Searching databases without query-building aids: implications for dyslexic users. Information Research **20**(4) (2015).
4. Berget, G., Mulvey, F., Sandnes, F. E.: Is visual content in textual search interfaces beneficial to dyslexic users?. International Journal of Human-Computer Studies **92**, 17-29 (2016).
5. Boye, A.: Teaching, Learning, & Professional Development Center. Retrieved from https://www.depts.ttu.edu/tlpdc/Resources/Teaching_resources/TLPDC_teaching_resources/StudentWriting.php (2017).
6. Brathovde, K., Farner, M. B., Brun, F. K., Sandnes, F. E.: Effectiveness of Color-Picking Interfaces among Non-Designers. In: International Conference on Cooperative Design, Visualization and Engineering, pp. 181-189. Springer, Cham. (2019). doi: 10.1007/978-3-030-30949-7_21
7. Brinck, T., Gergle, D., Wood, S. D.: Writing for the web. In: Usability for the Web: Designing web sites that work, pp. 244-301) San Francisco, CA, etc.: Morgan Kaufmann (2003).
8. Cavaleri, M.R., Dianati, S.: You want me to check your grammar again? The usefulness of an online grammar checker as perceived by students. Journal of Academic Language and Learning **10**, A223-A236 (2016).
9. Charney, D.: The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English **18**, 65-81 (1984).
10. Chung, J.-W., Min, H.-J., Kim, J., Park, J.C.: Enhancing readability of web documents by text augmentation for deaf people. Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, pp. Article 30. Association for Computing Machinery, Madrid, Spain (2013). https://doi.org/10.1145/2479787.2479808.
11. Dale, E., Chall, J.: The Concept of Readability. Elementary English **26**(1), 19-26 (1949).

12. Dale, R.: Checking in on grammar checking. Natural Language Engineering **22**, 491-495 (2016).
13. Dubay, W.: The Principles of Readability. CA 92627949, 631-3309 (2004).
14. Eika, E., Sandnes, F. E.: Assessing the Reading Level of Web Texts for WCAG2. 0 Compliance—Can It Be Done Automatically? In: Advances in Design for Inclusion, pp. 361-371, Springer (2016).
15. Eika, E., Sandnes, F. E.: Authoring WCAG2. 0-compliant texts for the web through text readability visualization. In: International Conference on Universal Access in Human-Computer Interaction, pp. 49-58. Springer, Cham (2016).
16. Eika, E.: Universally designed text on the web: towards readability criteria based on anti-patterns. Stud. Health Technol. Inform **229**, 461-470 (2016).
17. Flesch, R.: A new readability yardstick. Journal of Applied Psychology **32**(3), 221–233 (1948).
18. Gilliland, J.: The concept of readability. Reading **2**, 24-29 (1968). doi: 10.1111/j.1467-9345.1968.tb00749.x
19. Grammarly: Write your best with Grammarly. (n.d.). Retrieved from https://www.grammarly.com/
20. Gray, W. S., Leary, B. E.: What makes a book readable? Oxford, England: Univ. Chicago Press (1935).
21. Gunning, R.: The technique of clear writing. New York: McGraw-Hill (1971).
22. Habib, L., Berget, G., Sandnes, F. E., Sanderson, N., Kahn, P., Fagernes, S., Olcay, A.: Dyslexic students in higher education and virtual learning environments: an exploratory study. Journal of Computer Assisted Learning **28**(6), 574-584 (2012).
23. Hansen, F., Krivan, J. J., Sandnes, F. E.: Still Not Readable? An Interactive Tool for Recommending Color Pairs with Sufficient Contrast based on Existing Visual Designs. In: The 21st International ACM SIGACCESS Conference on Computers and Accessibility, pp. 636-638. ACM (2019). doi: 10.1145/3308561.3354585
24. Hargis, G.: Readability and computer documentation. ACM Journal of Computer Documentation **24**, 122-131 (2000).
25. Hines, R., Basso, J.: Do Communication Students Have the "Write Stuff"?: Practitioners Evaluate Writing Skills of Entry-Level Workers. Journal of Promotion Management **14**, 293-307 (2008). doi: 10.1080/10496490802625817
26. Jacobs, H., Zinkgraf, S., Wormuth, D., Hearfiel, V., Hughey, J.: Testing ESL Composition: a Practical Approach. (1981).
27. Janssen, G., Meier, V., Trace, J.: Building a better rubric: Mixed methods rubric revision. Assessing Writing **26**, (2015). doi: 10.1016/j.asw.2015.07.002
28. JASP Team: JASP (Version 0.11.1)[Computer software] (2019).
29. Jatowt, A., Tanaka, K.: Is wikipedia too difficult? comparative analysis of readability of wikipedia, simple wikipedia and britannica. Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 2607–2610. Association for Computing Machinery, Maui, Hawaii, USA (2012). doi: 10.1145/2396761.2398703.
30. Jönsson, A., Svingby, G.: The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review 2, 130-144 (2007). doi: 10.1016/j.edurev.2007.05.002
31. Kadayat, B. B., Eika, E.: Impact of Sentence length on the Readability of Web for Screen Reader Users. In International Conference on Universal Access in Human-Computer Interaction. Cham: Springer (2020).
32. Klare, G.: The measurement of readability: Useful information for communicators. ACM Journal of Computer Documentation 24, 107-121 (2000). doi:10.1145/344599.344630.

33. Klimova, B.: Evaluating Writing in English as a Second Language. Procedia - Social and Behavioral Sciences **28**, 390–394 (2011). doi: 10.1016/j.sbspro.2011.11.074
34. Lidwell, W., Holden, K., Butler, J.: Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design. Rockport Pub 198 (2010).
35. Mahon, R.: A Grading System for Composition Papers. The Clearing House 69, 280-282 (1996). doi: 10.1080/00098655.1996.10114317.
36. Mc Laughlin, G.H.: SMOG grading-a new readability formula. Journal of reading 12, 639-646 (1969).
37. McKinley, V.: Keeping it Simple: Making Regulations Write in Plan Language. Regulation 21, 30 (1998).
38. Meade, C., Smith, C.: Readability formulas: Cautions and criteria. Patient Education and Counseling **17**, 153-158 (1991). doi:10.1016/0738-3991(91)90017-Y
39. Moskal, B., Leydens, J.: Scoring Rubric Development: Validity and Reliability. Practical Assessment Research and Evaluation **7**, (2000).
40. Schraudner, M.: The online teacher's assistant: Using automated correction programs to supplement learning and lesson planning. CELE Journal 22, 128-140 (2014).
41. ONeill, R., Russell, A.: Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. Australasian Journal of Educational Technology **35**, (2019).
42. Pedersen, L. A., Einarsson, S. S., Rikheim, F. A., Sandnes, F. E.: User Interfaces in Dark Mode During Daytime – Improved Productivity or Just Cool-Looking? In: International Conference on Universal Access in Human-Computer Interaction. Cham: Springer (2020).
43. Pitler, E., Nenkova, A.: Revisiting readability: a unified framework for predicting text quality. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 186–195. Association for Computational Linguistics, Honolulu, Hawaii (2008).
44. Rakedzon, T.: "To make a long story short: A rubric for assessing graduate students' academic and popular science writing skills". Assessing Writing **32**, (2017). doi: 10.1016/j.asw.2016.12.004
45. Sandnes, F. E.: Universell utforming av IKT-systemer, Oslo: Universitetsforlaget, 2nd edition (2018).
46. Sandnes, F. E.: On-screen colour contrast for visually impaired readers: Selecting and exploring the limits of WCAG2. 0 colours. In: Information design: research and practice (Eds: Black, A., Lund, O., & Walker, S.), pp. 405-416 (2016).
47. Sandnes, F. E.: Understanding WCAG2. 0 color contrast requirements through 3D color space visualization. Studies in health technology and informatics **229**, 366-375 (2016). doi: 10.3233/978-1-61499-684-2-366
48. Sandnes, F. E., Zhao, A.: An interactive color picker that ensures WCAG2. 0 compliant color contrast levels. Procedia Computer Science **67**, 87-94 (2015). doi: 10.1016/j.procs.2015.09.252
49. Sandnes, F. E., Zhao, A.: A contrast colour selection scheme for WCAG2. 0-compliant web designs based on HSV-half-planes. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1233-1237. IEEE (2015). doi: 10.1109/SMC.2015.220
50. Sandnes, F. E.: An image-based visual strategy for working with color contrasts during design. In: International Conference on Computers Helping People with Special Needs, pp. 35-42. Springer, Cham (2018). doi: 10.1007/978-3-319-94277-3_7
51. Schriver, K. A.: Readability formulas in the new millennium: what's the use? ACM Journal of Computer Documentation **24**, 138-140 (2000).

52. W3C: Web Content Accessibility Guidelines (WCAG) 2.1. (2018, June 5). Retrieved from https://www.w3.org/TR/WCAG21/
53. Weigle, S.C.: Assessing Writing. Cambridge University Press (2002)
54. Wright, N.: Free eBook : StyleWriter's New BOG INDEX Readability Formula : Readability Software. Retrieved from http://www.stylewriter-usa.com/bog-index-readability-formula.php.
55. Yu, C.-H., Miller, R.C.: Enhancing web page readability for non-native readers. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2523–2532. Association for Computing Machinery, Atlanta, Georgia, USA (2010). https://doi.org/10.1145/1753326.1753709.