Challenges of using auto-correction tools for language learning

Conference Paper · March 2022

DOI: 10.1145/3506860.3506867

3 authors, including:

Leo Sylvio Rüdian
Humboldt-Universität zu Berlin
27 PUBLICATIONS **69** CITATIONS
SEE PROFILE

Moritz Dittmeyer
Ludwig-Maximilians-University of Munich
6 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Challenges of using auto-correction tools for language learning

Sylvio Rüdian
Humboldt-Universität zu Berlin,
Weizenbaum Institute e.V.
ruediasy@informatik.hu-berlin.de

Moritz Dittmeyer
Goethe-Institut e.V.
moritz.dittmeyer@goethe.de

Niels Pinkwart
Humboldt-Universität zu Berlin
pinkwart@hu-berlin.de

ABSTRACT

In language learning, getting corrective feedback for writing tasks is an essential didactical concept to improve learners' language skills. Although various tools for automatic correction do exist, open writing texts still need to be corrected manually by teachers to provide helpful feedback to learners. In this paper, we explore the usefulness of an auto-correction tool in the context of language learning. In the first step, we compare the corrections of 100 learner texts suggested by a correction tool with those done by human teachers and examine the differences. In a second step, we do a qualitative analysis, where we investigate the requirements that need to be tackled to make existing proofreading tools useful for language learning. The results reveal that the aim of enhancing texts by proofreading, in general, is quite different from the purpose of providing corrective feedback in language learning. Only one of four relevant errors (recall=.26) marked by human teachers is recorded correctly by the tool, whereas many expressions thought to be faulty by the tool are sometimes no errors at all (precision=.33). We provide and discuss the challenges that need to be addressed to adjust those tools for language learning.

CCS CONCEPTS

• **Applied computing** → Education; E-learning; Education; Interactive learning environments; • **Human-centered computing** → Human computer interaction (HCI); Interaction paradigms; Natural language interfaces.

KEYWORDS

Language learning, automated feedback, online course, written corrective feedback

ACM Reference Format:
Sylvio Rüdian, Moritz Dittmeyer, and Niels Pinkwart. 2022. Challenges of using auto-correction tools for language learning. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3506860.3506867

LAK22, March 21–25, 2022, Online, USA

ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506867

1 INTRODUCTION

Language is at the center of everyday communication. There are many reasons why people learn foreign languages, and there are many different ways they do so. Students learn languages at school to be prepared for future work in a globalized world. Many school graduates keep on learning languages to enhance their career opportunities. Others learn for pleasure or different private reasons. All language learners have in common that they need to learn and practice using a new language. To support this, teachers help students by providing corrective feedback, e.g. for students' written texts. Getting corrections and recommendations can be beneficial for students to optimize their language skills by correcting possible comprehension errors [1]. While this has been state of the art for hundreds of years, technology-enhanced solutions can be used nowadays that may support teaching and learning languages.

A large group of language learners is represented by refugees who aim to learn languages in order to get along in their new country of residence [2]. Missing language skills increase the risk of social exclusion. Thus there is a great need to learn languages. A major problem is that refugees learn their language basics according to the "Common European Framework of Reference for Languages" (CEFR) [3], which does not necessarily meet refugees' actual needs. Although descriptions for levels in the CEFR are generic, taught words or phrases for each level are based on contexts that are not useful for the immigration process [4]. However, refugees are in very stressful situations, with lots of fear and uncertainty. Thus, they need to learn how to communicate with authorities in a short period of time, instead of knowing how to describe elements of pictures (just as an example). Existing learning material for concrete everyday situations of refugees (e.g. with public authorities) is missing; the basics of generic language courses are insufficient for the required skills. This is why learners use tools like *Google translate,* which are freely available to learn languages [5]. They type in what they want to say and use translation tools as a basis to learn. However, generated translations often use advanced-level words and more complex grammar than learners would do [6]. The overwhelming use of more complex texts reduces the learner's understanding and control over their texts. Alternatively, learners use auto-correction tools to optimize their language skills. Auto-correction or automated writing evaluation tools help to optimize written matters by proofreading. Thus, learners have more control over the text and can optimize mistakes found by the tool. There are freely available auto-correction tools that provide concrete feedback on mistakes. This could be a good base to assist language learners in improving their writing skills. The benefits of giving feedback for open writing tasks automatically would be huge. If automated writing evaluation is working correctly and if the errors identified are similar to those of

human teachers, this would be a beneficial technology for supporting teachers in correcting texts and providing feedback. However, there is a need for these tools to work properly. Found mistakes must be classified as mistakes, and error-free texts or text passages should not be classified as faulty. This is especially important as language learners in contrast to native speakers are often not aware to distinguish between faulty corrections and correct ones in their writings.

Since numerous automated writing evaluation tools exist [7], we focus on one well-known example in this paper: The LanguageTool (version 5.5) for German texts is an open-source tool that was refined in the last years and is now used in many text-based applications [8]. The authors of the tool claim that it detects "errors, [. . .] grammar issues, commonly confused words, and punctuation oversights, [it] offers style suggestions: synonym replacements for overused words, concise rephrasing of wordy sentences, and formal alternatives to commonly used expressions" [9]. Its scope of application includes extensions for web browsers like Google Chrome, Mozilla Firefox, or Edge, Libre-Office or even mailing tools like Thunderbird. Thus nowadays, the LanguageTool has become one of the state-of-the-art tools widely used for everyday communication. Based on the description and the open-source architecture, we think this is a promising base for correcting texts of language learners instead of developing a new tool from scratch. Its description fulfills the demands for correcting texts in language learning. A deeper analysis of the current state is necessary to understand its usefulness for language learning. In this paper, we focus on two research questions. RQ1: What is the overlap of mistakes detected by teachers and the LanguageTool in open writing tasks of a language learning online course? RQ2: Which challenges need to be addressed to change the tool's target from proofreading to giving feedback for learners?

2 RELATED WORK

In language learning, feedback is a crucial component to enhance learning outcomes. The idea of giving feedback to learners is to minimize the gap between teacher expectations and learner skills [10]. In the literature, several different kinds of feedback are distinguished. One essential category is written corrective feedback, which aims at telling the learner that there is something linguistically wrong with what they have written [11]. As such, corrective feedback is descriptive in nature and often perceived as negative feedback [12]. There are many further distinctions regarding how written corrective feedback can and should be formulated to help students enhance their writing skills. One of its most ordinary forms is direct, explicit error correction, which identifies each error and suggests its correction. More sophisticated types of corrective feedback involve e.g. indirect or focused approaches [13]. A completely different category of feedback is to assign grades or numerical scores. Grading generally has another purpose than corrective feedback. It aims to provide the student with a summative assessment about his/her level of knowledge at the end of some learning unit instead of giving formative feedback about errors and susceptibilities, as this is primarily the case for corrective feedback [14]. Two areas of research in the field of computer assisted learning address this topic of feedback. They differ in purpose and details in given feedback. The purpose of automated essay scoring (AES) is to automatically classify the quality of written texts by predicting grades or scores for text submissions. The field of automated writing evaluation (AWE) focuses on finding and highlighting concrete errors. Doing so, AWE is primarily concerned with corrective feedback or, more precisely, direct, explicit error correction. In contrast to AES, it is used mainly for proofreading in native language contexts and is not necessarily connected with language learning. AES can be done using different approaches, e.g. Neural Models [15], abstract representations of texts using hierarchical classification methods [16], or combinations with online course contents [17] to be useful in education.

AES tools often address two areas: spelling and grammar. While the first part can be fulfilled with word lists, evaluating the grammar requires natural language processing. Therefore, applications focus on text's part-of-speech (POS) tag representation which can be derived automatically [18]. A valid grammar can be defined by a fixed set of rules that consist of POS tag sequences [19]. The fixed set of POS tags is represented by n-grams, where n represents the number of sequential POS tags [20]. Both, AES and AWE use this generic representation of texts, but they differ in the richness of derived feedback. The AWE of the LanguageTool [8] uses the spelling checker "Snakespell" [21] and the POS tagger of [18]. Rules for typical mistakes were added to provide useful feedback, including a possible correction that could also be beneficial for a language learner. Thus, instead of examining whether a text is valid according to rules that define grammar, typical mistakes were extracted and patterns were created (including a message, what the mistake is about). This is principally a good base to provide feedback for open writing tasks. According to the comparison by Näther [22], the LanguageTool in its nowadays version generates good results for correcting sentences. Naber stated that no corpus of uncorrected texts existed [8] and thus, precision and recall [23] could not be measured to determine its usefulness, especially for the field of language learning. Focusing on open writing tasks in language learning, we use a dataset of the Goethe-Institut in this paper for an evaluation to bridge the gap mentioned by Naber. Further, we ask teachers at the Goethe-Institut in qualitative interviews about the limitations and challenges that need to be overcome to become a valuable tool in language learning.

3 METHODOLOGY

First, 100 users of different language levels learning German as a foreign language participated in different German online courses[1] related to their levels. All courses contained at least one open writing task, where the learners faced an open question to write about. The courses were tutored by qualified German teachers to provide high-quality feedback. We collected all user submissions, including the corrected versions.

There was no "artificial" study setting that could be inferred by the teachers involved, as open writing tasks are always manually corrected by human teachers who also give feedback in other courses of the platform. Thus, teachers were not aware of being part of the study. This is important to avoid bias based on the feeling of being observed or evaluated within a study (Hawthorne effect [24]).

[1] https://www.goethe.de/

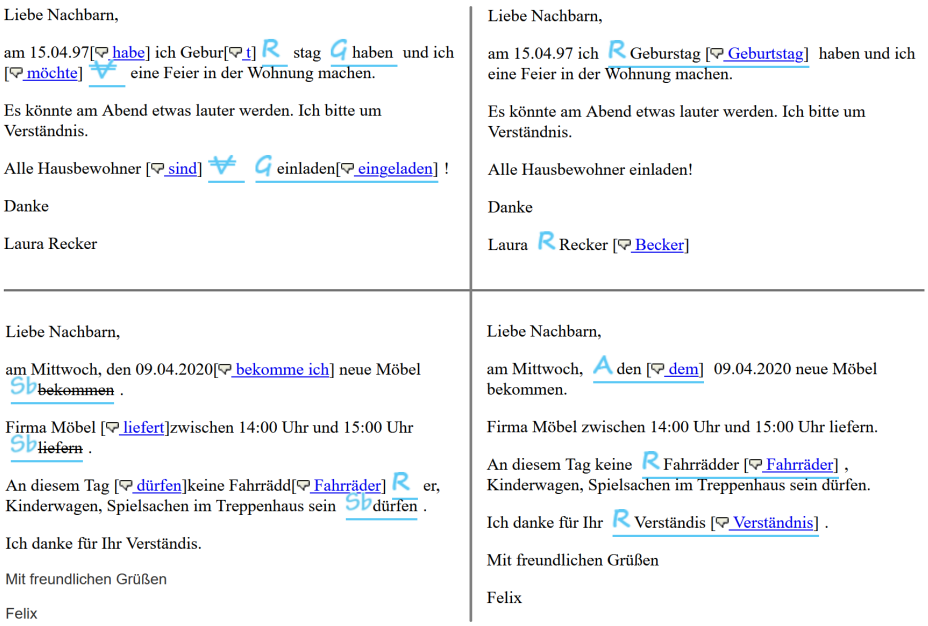

Figure 1: Examples of two corrections, texts on the left are corrected by teachers, texts on the right side are corrected by the LanguageTool.

To address our research question, we aimed to compare detected errors by teachers (T) and the LanguageTool (L) for the same texts. Therefore, the 100 texts have been corrected by the AWE tool, which underlines detected errors. This procedure is comparable with the approach of language teachers that correct learner texts. Then, detected mistakes by L were transferred into the visualization mode used by the teachers for a direct comparison. An example can be found in Figure 1

In the next step, the comparisons of all corrections took place. We counted all detected errors, which were the same in both corrections of T and L (true positives, TP); we counted all errors that could be found in the teachers' annotations but not in the corrections suggested by the AWE tool (false negative, FN) and vice versa (false positives, FP). We did not consider the type of error while comparing the corrections.

For each text, we used precision (P) and recall (R) to measure the similarity of annotations. This is a common metric to compare the similarity of annotations [23]. Finally, we used the mean of P and R over all texts to get insights into the overall dataset. This can be expressed by:

$P_\emptyset = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FP_i}$; $R_\emptyset = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i+FN_i}$ with n = 100. The results give a clear and data-driven view on the practicability of the AWE tool in the current version, without modifications.

In the next step, we aimed to get more insights into the differences of T and L to understand what needs to be optimized to change the AWE tool to be useful for language learning. We randomly chose 25 texts of different open writing tasks that have not been used in the first step, and conducted interviews with 19 experienced language teachers, who have a full-time job in educating German as a foreign language. In this qualitative study, the auto-correction of each text was presented to the teachers one by one, and they were asked to provide feedback, positive and negative aspects, and improvements so that the AWE tool becomes useful for them. All answers were collected and categorized to get a concrete gap between the teachers' expectations and the AWE tool at the current state. Based on this, we derived suggestions on how to improve the AWE tool.

4 RESULTS

Following the study design, we got $P_\emptyset = 0.33$ and $R_\emptyset = 0.26$ for our first investigation ($F_1 = 0.29$). From the result of $P_\emptyset$ it follows that one third of found mistakes by T are correct. $R_\emptyset$ describes how many relevant errors were found by L related to all relevant ones. The resulting value indicates that L classifies much more parts of the texts wrong then it is the case for the teachers. This answers RQ1. As theses precision and recall values question the practicality of the AWE tool, we moved on addressing the reasons to optimize the current state of the tool to make it useful for language learning. Therefore, in the second study, 25 texts were analyzed separately by 19 teachers. In the qualitative analysis, teachers classified 34.4% of the auto-corrected texts to be useful as a base for further corrections. Thus, the precision of correct identified errors is equal to the perceived usefulness in correcting open writing tasks in language learning. The identified challenges are grouped by their relations to "mistakes", "didactics" and "semantics", which are summarized in Table 1 to address RQ2.

Nevertheless, the AWE tool was not rejected completely. There were also positive comments, especially for the corrections that were classified to be useful. This was often the case for nouns

Table 1: Identified challenges

Class	Hints by tutors
Mistakes	- not all mistakes were found, especially grammar, phrasing, or syntax
	- incorrect marks
	- names of learners were often classified to be misspelled
	- proposals for corrections are often wrong
Didactics	- if the sentence requires a different structure, detailed feedback is not useful
	- selecting only mistakes that the learner already should know
	- avoid overwhelmed feedback
	- select only didactically useful mistakes
	- double empty spaces or formatting mistakes are not important
Semantics	- feedback for corrected words should fit the context
	- proper names should be excluded
	- missing words were not found
	- formal and informal speech should be detected

with spelling errors that the AWE tool detected conscientiously. Mistakes concerning the case sensitivity were also emphasized, which is an important topic in the German language as much more words have to start with a capital letter than in English. However, those automatic corrections were only classified to be helpful if they had a high precision. Thus, conversely, precision is one of the most important criteria.

Having a deeper look into the corrections of L, over 66% of mistakes that teachers highlighted were not detected. The AWE tool often marks errors that are no errors at all, which can also be seen in the low precision value. This was criticized a lot by teachers. From their perspective, it is an additional effort to remove incorrectly marked corrections manually. As the AWE tool should be helpful to reduce the time for correcting texts, removing errors that are no errors at all is too time-consuming. Thus, there is the need to optimize the recall as this is a knockout criterion, according to the teachers. Besides, if the correction will be given to the learner without a preview of the tutor, giving incorrect corrections is not debatable.

One problem of the AWE tool is very prominent: learners often provide their name at the end of the texts, e.g. when being instructed to write a letter. This name is very often marked by L as misspelled, although this is not the case. On top, suggestions for a correct version of this name will be provided, mainly no proper names, rather words of a dictionary that appear to be equal. An automatic correction providing students with this kind of feedback appears to be very unprofessional.

In our study, teachers often did not select all errors and focused on useful ones from a didactical perspective. It is often said that there is the need to show only errors if they are helpful for the learning process. L does not distinguish between "important" or "not important" errors from the perspective of language learning. Thus, it is not surprising that often more errors are found by L, which results in a low recall.

An extension in L helps to make texts semantically more correct if a logic expression was defined. If a date including the day of the week was detected that principally is not existing, the AWE tool gives a warning and marks it as an error. This is not helpful for learning languages. The punctuation extension detects unusual or missing punctuations, e.g. in dates. This is a detail that could be interesting for advanced learners only.

One deficiency that Naber [8] observed was the precision concerning the sentence boundary detection, which was 97%. In addition, the POS tagger reached an accuracy of 93.05%. Thus finding a match to a defined n-gram depends on the correct classification of the POS tags in advance.

We did not consider the correct classification of error types. Nevertheless, as we found out that the gap between errors detected by teachers and those detected by the AWE tool is high, and as some types cannot be detected by the AWE tool due to technical limitations (e.g. limiting the POS n-grams to 4 tags), it is not recommended to use this AWE tool for automated feedback in language learning without modifications. However, it can still be beneficial guidance for teachers to detect mistakes that they might have overlooked.

5 HOW TO BRIDGE GAPS & DISCUSSION

Based on the study, we identified several gaps so that we cannot recommend using the AWE tool's current version for language learning. This is not a surprising result as the aim of proofreading is different from providing corrective feedback to language learners. For the pre study, we used 100 texts. Based on this limitation and the short learner texts those lengths are typically for language learning beginners; this study gives a rough estimation whether the AWE tool can be used for language learning in its current version. Based on our results we do not claim that precision and recall will be generalizable if we focus on another purpose like proofreading. However, using the AWE tool as a basis for a new tool could be promising as it has not been criticized completely and detects at least some relevant errors correctly. This section provides some solutions to bridge identified gaps in order to make the AWE tool more suitable for language learning.

First, we discuss the class "Mistakes" of Table 1 and how to optimize it. We have different layers that need to be addressed. We begin with spelling. Nouns were often correctly identified to be faulty. Here, nouns were compared with lists of words. This works well for long words where many letters are required. For each word, a comparison to the word list can take place. If there is a match or

the noun was adjusted (e.g. for countable nouns, mainly adding "en" to German nouns, it does not need to be further considered. If there is no match, a syntactic alignment could be used to find the most similar existing word [25]. As this does not work on the semantic layer, the AWE tool also says that proper names, often classified as nouns, have a related "correct" word, which partly overlap in some characters. The example in Figure 1 illustrates this. Thus, proper names need to be identified, and they must be excluded as words "out of vocabulary". This approach was examined in the last decades [26], and further experiments show good results [27]. A combination with the AWE tool would fix this shortcoming.

Also, the correction of grammar is an issue. The architecture of L was not designed to find out whether grammar is correct. Instead, typical mistakes are recognized based on patterns. Thus, there is no way to find more grammar mistakes as the hand-crafted rules only represent a subset of possible errors. If the structure of a sentence is wrong, e.g. using a word that should be at the beginning but was used at the end, it cannot be detected. The identified problem can be traced back to the sequence length of POS tags, set to 4 [8]. A naïve solution is to increase the POS sequence analyzed. This allows to detect errors of more than 4 consecutive words, but it requires creating new rules for possible errors. However, this is of high interest for language learning as wrong word order in sentences is a common mistake for learners. Besides, the more POS tags were compared, the fewer mistakes will be found as each case of wrong orderings needs to be defined, following the AWE tool's design. In general, language can be defined by rules and patterns [28]. It is wise to use these rules and detect whether there is a pattern in the learner's text, that cannot be explained by the rules and expressions. Doing this on the representations of part-of-speech-tags, this analysis runs on a very abstract representation of language, where a finite number of correct combinations exists. If a deviation was found, then using a pre-defined pattern is useful to determine the error type and give explanations for possible transformations or provide examples. Nevertheless, if no pre-defined rule exists and it cannot be explained by existing rules, it is still a mistake, that cannot be detected. In an improved version of the AWE tool, detecting grammar issues should be the first step, followed by a classification. Otherwise, the precision and recall will be low.

Analyzing the syntax of sentences is also related to grammar. Identifying missing words consists of two challenges. First, the position of a possibly missing word needs to be located. Second, the missing word itself has to be identified. However, if the abstract POS representation is used, missing words can be identified by rules that represent grammar structures. The idea is to use existing rules (that represent typical sentence structures), and the AWE tool needs to find the most similar existing POS pattern. This can be realized with the alignment approach of Altschul & Erickson [25], which needs to be adjusted to align POS patterns (like in [29]). Thus, possible mistakes, based on gaps could be identified. Then, the position of missing words, including their part of speech, can be identified.

A prediction model using the bidirectional encoder representation from transformers (BERT [30]) could be beneficial to overcome the second challenge of identifying concrete missing words. Therefore, the identified gap needs to be masked with a [MASK] tag in the first step, which will be replaced by a predicted word that has a semantic relation to the sentence. Such an approach would be helpful to solve the problem of non-detecting correct missing words in learner texts. Besides, such a model can also be used to predict recommendations, in general, to reduce faulty suggestions when spelling mistakes were identified. It is important to note that pre-trained language models should be used for this approach as training bidirectional transformers is very expensive.

Besides the technical limitations, the discussed didactical appropriateness of the correction suggested by the AWE tool was heavily criticized. First, provided feedback should come along in different levels. If the structure of a sentence is wrong, then it is sufficient to give hints about the structure; words themselves do not need to be corrected. Corrections need to be more deeply filtered if a structural error for a phrase or sentence is detected. Then, the learner can focus on this error and will not be confused due to many markers. Second, if words were recommended as corrections, they should fit the content. Semantic analysis is usually used to determine whether the context (e.g., a sentence) is true based on previous information [31]. In the area of language learning for submissions of open writing tasks, the domain is mainly specific and pre-defined. Thus, the information is required whether suggested words have a relation to the context. Concept-based structures like the WordNet [32] can be used to overcome this limitation. The idea is that words, phrases, or sentences belong to a concept and concepts can be represented by a similarity score. If a word is recommended by the AWE tool, it should have a low similarity score according to the phrase and surrounding sentences. To do this a concrete threshold could be determined to specify if the suggested word fits the context. If this is not the case, it should not be used for recommendation.

From a didactical perspective, automatic corrections should be on a par with the standards of providing corrective feedback in focus on language learning. The idea is not to mark all errors as it is the case in proofreading. Instead, feedback should be helpful for the learner by highlighting misunderstandings. Thus, there is the need to select only mistakes that are important for the learning process. If rules are used to detect concrete grammar errors as it is the case for the AWE tool [8], all rules should be labeled with the particular error type. Then, if the tool's detection is the same as the teacher ones, labels of the error type could automatically be derived. Corrected texts by teachers could be used as a basis to distinguish between relevant and non-relevant mistakes. If errors have not been marked to be wrong by teachers, they do not need to be marked by the AWE tool either. This distinction works under the assumption that the overlap of found mistakes by teachers and the AWE tool are similar, which is not the case in the current version. Alternatively, rules could be applied, as stated above, that represent the correct language usage. If a difference to this gold standard was detected in previous texts and also marked as faulty by teachers, the underlying POS sequence should be extracted for further use. If a POS sequence of this set will be found in new submissions, these mistakes will be highlighted. Otherwise, they will not be considered faulty as the learner may not have the knowledge at the current language level, and thus these errors are not important from a didactical perspective.

One last identified deficiency of the AWE tool was to distinguish between formal and informal speech. This is not a knockout criterion but a possible extension to make the AWE tool more valuable

in language learning. The idea is that depending on the task, the style needs to be different (e.g. informal speech in a private letter). Sheika and Inkpen [33] have shown that they can distinguish between formal and informal language with high accuracy. Using this could be an interesting extension to provide feedback if the style was not met.

6 CONCLUSION

In this paper, we investigated the performance of a specific AWE tool in providing automated corrective feedback for open writing tasks in a German online course at the Goethe-Institut. To assess its viability, we compared errors detected by the AWE tool with those marked by human teachers. The results have shown that there are major differences in the resulting corrections. The significant challenges are based on the design of the AWE tool. It leads to possibly sparse rules as errors have to be formalized using POS n-grams and categorized according to the error type. Many possible rules used to detect errors have not been created manually yet. Thus, corrections are incomplete, which is in line with the results. The results have shown that the AWE tool at its current stage of development should not be used for language learning. However, we identified several challenges that need to be addressed for adjusting the AWE tool to be used for correcting texts of open writing tasks in language learning.

Overall, overcoming the challenges identified is possible. Integrating some further state of the art technology into the AWE tool makes it more valuable for the language learning. To the best of our knowledge, the three classes of challenges can be solved by changing the tool's focus from proofreading to language learning. If the AWE tool does not have the current limitations, it can assist teachers in correcting texts and the tool could also be used to provide feedback directly if the precision is much better than it is now. For many teachers, correcting texts is the most time-consuming task, and they aim to use the time to assist learners during the learning process rather than doing the "batch work" of correcting texts. The time that will be saved could be used to help learners in a more personalized way. Thus, we can conclude that, transferred with the appropriate effort, an adjusted version of the AWE tool investigated is a valuable resource to support the language learning process in online language courses.

ACKNOWLEDGMENTS

This work was supported by the Goethe-Institut e.V. and the German Federal Ministry of Education and Research (BMBF), grant number 16DII127 (Weizenbaum-Institute e.V.). The responsibility for the content of this publication remains with the authors.

REFERENCES

[1] R. E. Clark and R. E. Mayer, "E-learning and the science of instruction (2)," San Francisco, Jossey-Bass, 2008.
[2] L. Morrice, L. K. Tip, M. Collyer and R. Brown, "You can't have a good integration when you don't have a good communication': English-language learning among resettled refugees in England," in *Journal of Refugee Studies 34(1)*, 2021, pp. 681-699.
[3] Council of Europe, "A Common European Framework of Reference for Languages: Learning, Teaching, Assessment," in *Council for Cultural Co-operation*, Strasbourg, Cambridge University Press, 2001.
[4] G. Chamorro, M. d. C. Garrido-Hornos and M. Vázquez-Amador, "Exploring ESOL teachers' perspectives on the language learning experiences, challenges, and motivations of refugees and asylum seekers in the UK," in *International Review of Applied Linguistics in Language Teaching*, 2021.
[5] V. Abou-Khalil, S. Helou, B. Flanagan, N. Pinkwart and H. Ogata, "Language learning tool for refugees: Identifying the language learning needs of Syrian refugees through participatory design.," in *Languages 4(3)*, 2019, p. 71.
[6] S.-C. Tsai, "Using google translate in EFL drafts: a preliminary investigation," in *Computer Assisted Language Learning 32.5-6*, 2019, pp. 510-526.
[7] P. Vitartas, J. Heath, S. Midford, K.-L. Ong, D. Alahakoon and G. Sullivan-Mort, "Applications of Automatic Writing Evaluation to Guide the Understanding of Learning and Teaching," in *Show Me The Learning*, Ascilite, 2016.
[8] D. Naber, A Rule-Based Style and Grammar Checker (Diplomarbeit), Bielefeld, 2003.
[9] Languagetool, "Chrome Extension: Grammar and Spell Checker — LanguageTool," 16 09 2021. [Online]. Available: https://chrome.google.com/webstore/detail/grammar-and-spell-checker/oldceeleldhonbafppcapldpdifcinji?hl=en. [Accessed 24 09 2021].
[10] R. C. Ivanic and R. R. and Rimmershaw, "What am I supposed to make of this? The messages conveyed to students by tutors' written comments'," in *Student Writing in Higher Education: New Contexts*, Open University Press, 2000.
[11] Y. Sheen and R. Ellis, "Corrective feedback in language teaching," in *Handbook of research in second language teaching and learning (2)*, 2011, pp. 593-610.
[12] R. Ellis, "Corrective Feedback and Teacher Development," in *L2 Journal (1), eScholar-ship Repository*, 2009, pp. 3-18.
[13] J. D. R. Bitchener, "Written corrective feedback in second language acquisition and writing," Routledge, 2012.
[14] C. Garrison and M. Ehringhau, "Formative and Summative Assessments in the Classroom," amle, 2007.
[15] F. Nadeem, H. Nguyen, Y. Liu and M. Ostendorf, "Automated Essay Scoring with Discourse-Aware Neural Models," in *14th Workshop on Innovative Use of NLP for Building Educational Applications*, 2019, p. 484–493.
[16] D. McNamara, S. A. Crossley, R. Roscoe, L. Allen and J. Dai, "A hierarchical classification approach to automated essay scoring," in *Assessing Writing Volume 23*, Elsevier Ltd., 2015, pp. 35-59.
[17] S. Rüdian, J. Quandt, K. Hahn and N. Pinkwart, "Automatic Feedback for Open Writing Tasks: Is this text appropriate for this lecture?," in *DELFI 2020 - Die 18. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V.*, 2020, pp. 265-276.
[18] E. Brill, "A Simple Rule-Based Part of Speech Tagger," DTIC, 1992.
[19] P. G. Otero and I. G. López, "A grammatical formalism based on patterns of Part of Speech tags," John Benjamins Publishing Company, 2011.
[20] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval. Vol. 161175*, 1994.
[21] ScriptFoundry, "snakespell.py 1.01," 03 03 2001. [Online]. Available: https://web.archive.org/web/20040503101753/ https://scriptfoundry.com/modules/snakespell/.
[22] M. Näther, "An In-Depth Comparison of 14 Spelling Correction Tools on a Common Benchmark," in *Proceedings of the 12th Language Resources and Evaluation Conference*, France, European Language Resources Association, 2020, pp. 1849-1857.
[23] M. Buckland and F. Gey, "The relationship between recall and precision," John Wiley & Sons, Inc., 1994.
[24] S. R. Jones, "Was there a Hawthorne effect?," in *American Journal of sociology 98(3)*, 1992, pp. 451-468.
[25] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," in *Bulletin of Mathematical Biology, Volume 48*, Elsevier Ltd, 1986, pp. 603-616.
[26] I. Mani, T. R. MacMillan, S. Luperfoy, E. Lusher and S. Laskowski, "Identifying unknown proper names in newswire text," in *In Acquisition of Lexical Knowledge from Text*, 1993.
[27] I. Sheikh, I. Illina, D. Fohr and G. Linares, "OOV proper name retrieval using topic and lexical context models," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 5291-5295.
[28] D. Willis, "Rules, patterns and words: Grammar and lexis in English language teaching," Ernst Klett Sprachen, 2003.
[29] S. Rüdian and N. Pinkwart, "Towards an Automatic Q&A Generation for Online Courses - A Pipeline Based Approach," in *Artificial Intelligence in Education (AIED 2019)*, Chicago, Springer, 2019, pp. 237-241.
[30] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"2018.
[31] L. Floridi, "Is semantic information meaningful data?," in *Philosophy and phenomenological research 70(2)*, 2005, pp. 351-370.
[32] T. Pedersen, S. Patwardhan and J. Michelizzi, "WordNet:: Similarity-Measuring the Relatedness of Concepts," in *AAAI (4)*, 2004, pp. 25-29.
[33] F. A. Sheikha and D. Inkpen, "Learning to classify documents according to formal and informal style," in *Linguistic Issues in Language T. 8*, 2012.