MICROSOFT READING PROGRESS AS CAPT TOOL

MAREK MOLENDA
University of Łódź, Poland
marek.molenda@uni.lodz.pl

IZABELA GRABARCZYK
University of Łódź, Poland
izabela.grabarczyk@uni.lodz.pl

Abstract
The paper explores the accuracy of feedback provided to non-native learners of English by a pronunciation module included in Microsoft Reading Progress. We compared pronunciation assessment offered by Reading Progress against two university pronunciation teachers. Recordings from students of English who aim for native-like pronunciation were assessed independently by Reading Progress and the human raters. The output was standardized as negative binary feedback assigned to orthographic words, which matches the Microsoft format.

Our results indicate that Reading Progress is not yet ready to be used as a CAPT tool. Inter-rater reliability analysis showed a *moderate* level of agreement for all raters and a *good* level of agreement upon eliminating feedback from Reading Progress. Meanwhile, the qualitative analysis revealed certain problems, notably false positives, i.e., words pronounced within the boundaries of academic pronunciation standards, but still marked as incorrect by the digital rater.

We recommend that EFL teachers and researchers approach the current version of Reading Progress with caution, especially as regards automated feedback. However, its design may still be useful for manual feedback. Given Microsoft declarations that Reading Progress would be developed to include more accents, it has the potential to evolve into a fully-functional CAPT tool for EFL pedagogy and research.

Keywords: CAPT, EFL, ASR

1. Introduction

Since the beginning of the global pandemic caused by the SARS-Cov-2 virus, many educational institutions were forced to close their premises and transfer to “Emergency Remote teaching” (Hodges et al. 2020). In response, developers of educational platforms intensified their work on research and development of new affordances. One of such updates, which is potentially useful to pronunciation teachers and researchers, is Microsoft Reading Progress, implemented as a part of the MS Teams suite (https://education.microsoft.com/en-us/resource/50b18238). This feature promises to provide automated assessment and feedback of learners’ reading performance, including their pronunciation. Therefore, we decided to compare its output against the analysis of the same recordings conducted by human raters.

2. Theoretical and methodological background

2.1 Computer-Assisted Pronunciation Training

MS Reading Progress falls into the category of tools which are referred to as Computer-Assisted Pronunciation Training or CAPT. Unlike Computer-Assisted Language Learning (CALL), which was split into CALL proper, MALL (Mobile-Assisted Language Learning; Yaman & Ekmekçi, 2016) and RALL (Robot-Assisted Language Learning; Han 2012) in response to rapid technological changes, CAPT encompasses all commonly used digital tools for pronunciation instruction. Therefore, modern CAPT encompasses a wide variety of tools which can serve a number of pedagogical goals (Henrichsen 2021).

CAPT affordances requested by teachers and researchers to a large degree reflect educators' hopes for effective technology-assisted language instruction in general. These aspects include self-paced practice, individualized instruction, individualized feedback, authentic materials and elements of gamification (Levis 2007; Henrichsen 2021). In addition to these general requests, there exist domain-specific affordances, namely shifting some attention to suprasegmentals, the ability to accommodate acceptable variation in one's pronunciation, and learner exposure to various pronunciation models.

The assessment of CAPT solutions might be conducted on several levels; one of the most fundamental questions is the choice between the *nativeness principle* and *intelligibility principle* (Levis 2005). The former concept stems from the belief that correct pronunciation ought to be as close as possible to a selected native model. In contrast, the intelligibility principle posits that the appropriate way of teaching pronunciation is to focus on comprehensibility of one's speech to both native and non-native users of a given language (ibid.: 372). While choosing the right model depends on several easy-to-determine factors (e.g., the aim of pronunciation classes, formal requirements, etc.), it is considerably more difficult to determine which CAPT solution is optimal for a selected model. Although Rogerson-Revell remarks (with regret) that most CAPT is geared towards nativeness (2021: 192), the execution of this goal is still considered to be imperfect. For instance, Davey (2017) mentions the need to teach exaggerated (as opposed to native-like) "pitch, volume and speed" in order to satisfy the pronunciation assessment algorithm.

Another way of classifying CAPT tools focuses on the pedagogical value of the feedback given to learners. There is a general consensus that in this context *targeted feedback* is preferable to *binary feedback* (Bajorek 2017). In the case of the latter, the learner receives binary information on whether a certain sound was produced correctly. In contrast, targeted feedback normally contains information about the nature of the problem and clues as to what needs to be modified to improve one's performance. Among the tools that are able to provide targeted feedback, one interesting sub-category is the speech-to-text dictation software, which was not designed as a CAPT tool, but it may be nonetheless used in such capacity. As described in Molenda et al. (2018), dictation software based on Automated Speech Recognition (ASR) might show the learner what a native speaker would actually hear and understand. In this case, there is no explicit feedback; instead, the information about one's pronunciation deficiencies needs to be induced by comparing the target message with the output transcribed by the computer (see also Henrichsen 2021: 186).

Since the use of CAPT is a relatively recent advancement – in comparison to pronunciation training understood as a component of foreign language pedagogy – there is still a considerable amount of uncertainty as to which affordances of digital tools form the basis of a pedagogically-sound methodology. For instance, Pennington and Rogerson-Revell (2019) remark that many modern CAPT solutions are based on drilling, thus regressing to the audiolingual method instead of promoting communicative and phonological competence. The aforementioned debate between the nativeness principle and the intelligibility principle also contributes to contradictory opinions concerning digital solutions for pronunciation training: Henrichsen (2021) provides positive comments on studies which show the increased accuracy in non-native speech recognition in various software suites, while Bajorek (2017) stresses the importance of thresholds for "correct" pronunciation by stating that, "If the standard is too low, learners might believe that their utterances are sufficient…" (36). The existence of these and similar debates seems to indicate that the field of CAPT is not yet fully mature and stabilized. We believe that a part of the effort to develop a comprehensive methodology is related to the constant evaluation of new affordances, with the aim of building a robust knowledge base of digital tools used in pronunciation instruction.

2.2 ASR vis-à-vis human raters

One of means of assessing the quality of CAPT solutions is to compare the quality of their output with the feedback provided to the learners by human raters (Bernstein et al. 1990). This method is used to determine the *accuracy* of a given tool, and it differs from research which aims to present long-term effects on learners' pronunciation after using a CAPT application or program (see Mahdi & Al Khateeb 2019, for a meta-analysis of such studies). Instead, the accuracy-oriented assessment of CAPT tools can be likened to a review of a pedagogical resource before it is given to learners to use, as is common with dictionaries or textbooks. An accuracy-oriented review is also necessary in order to determine whether a given tool is ready to be deployed in research projects concerning language education. Given the aforementioned contradictory trends in a relatively young field of CAPT, there still remains some potential to produce ineffective or even harmful solutions (cf. Molenda et al. 2018).

In early experiments conducted by Coniam (1999), reading passages were used in order to determine the accuracy of ASR software. The assessment method used the data on the number of words, clauses and other units that were correctly retrieved by the computer. The results showed that the text read in a non-native accent of the Cantonese subject was not satisfyingly decoded by the software. In a similar study, conducted with the same piece of software (Dragon Naturally Speaking), Derwing et al. (2000) found that the ASR was considerably less successful "in interpreting accented speech" than the human raters (597). In addition, the research showed that there was no correlation between "software's recognition scores and the listeners' intelligibility scores", and the results of Pearson correlations clearly indicated that recognition of non-native speech was not related to human judgement (ibid.:598).

A different approach was adopted in Cucchinarini et al. (2000); unlike Coniam (1999) and Derwing et al. (2000), the researchers decided to use a dedicated piece of software capable of providing more specialized feedback. The results indicate that the most important "human" criterion for assessing pronunciation is segmental quality, "which (…) can be predicted most poorly on the basis of (…) machine scores" (ibid.: 118). Similarly, Kim (2006) compared pronunciation scores by specialized software and human raters in a reading-aloud test. The correlation between the two methods of rating was described by the author as "mediocre" (327), at 0.56 ($p<0.1$).

Agarwal and Chakraborty (2019) list a number of studies related to the measuring of CAPT accuracy. Chen and Jang (2015) studied the accuracy of their own ASR model against human raters to find that there was ≈15% mispronunciation detection, while Wang and Lee (2015) used HMM/GMM in a two-pass Viterbi decoding architecture to discover a ≈28% mispronunciation detection error rate. In another study based upon a two-pass framework, Qian et al. (2016), discovered ≈15% error in mispronunciation detection and ≈17% error in mispronunciation diagnosis. In 2017, Li et al. investigated the use of deep neural networks for mispronunciation detection, discovering ≈5% false positives and ≈31% false negatives in mispronunciation detection. More recently, Nushi and Sadeghi (2021) described ELSA Speak, remarking that the focus on the American variety of English might remain problematic for learners who pronounce their utterances with other accents, even if they are native-like. Finally, Souza and Gottardi (2022) examined the ability of dictation tools MS Word and VoiceNotebook to correctly transcribe foreign accented speech. Despite high overall intelligibility, they discovered vowel and consonant substitution, which could lead to communication breakdowns.

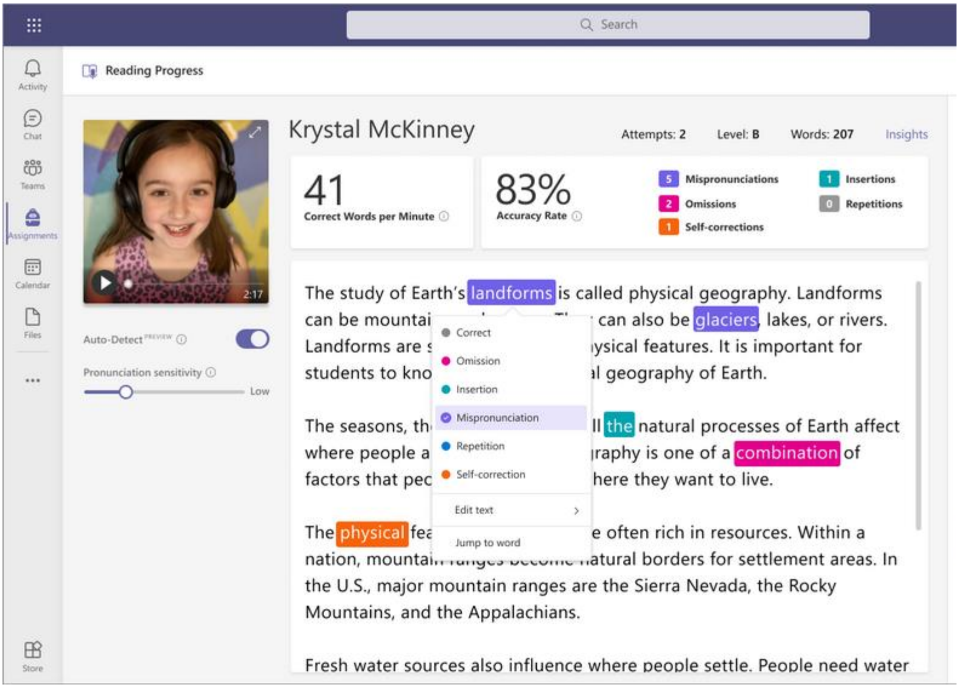
The design of the studies mentioned in this section illustrates choices and assumptions which have to be made by researchers who want to investigate the relationship between human ratings and ASR-based CAPT tools. In the studies which compare the feedback provided to students by human raters (HRs) and by digital solutions, the human is normally the standard for pronunciation assessment, so the tool which comes close to replicating raters' scores is considered to be more reliable. Secondly, there remains the choice between general-purpose ASR and some specialized software. With the rapid development of speech recognition systems, such as digital assistants by Google, Microsoft, Apple, Samsung or Amazon, it seems that the former option might offer better quality due to the sheer amount of data that these and similar companies can use to perfect their machine learning (Molenda et al. 2018). However, specialized software is more likely to provide quality targeted feedback to the learner. The feedback provided by specialized software might come in two forms (Derwing 2000: 594). It may be either implicit targeted feedback, when the system misrepresents the word or phrase read by the student, or the explicit targeted feedback, which is manifested by outputting the correct word while still marking it as mispronounced by means of highlighting, underlining, changing font, etc. It should be noted that the authors are not aware of any ASR-based solutions which would highlight the words/phrases which are particularly well-pronounced. Therefore, all the feedback we refer to is to be considered corrective in its nature.

Our study, to the best of our knowledge, is the first one to examine the accuracy of a newly-introduced Microsoft Reading Progress in the context of CAPT. Given the fact that Microsoft – a long-time provider of speech recognition technology (e.g. Souza and Gottardi 2022) – decided to introduce a tool which explicitly targets pronunciation problems, it was deemed necessary to assess its performance in the context of potential future studies as well as pedagogical use.

2.3 ASR-based CAPT in MS Reading Progress

Microsoft announced Reading Progress Module as a part of their 2021 MS Teams update (Ray 2021). Especially promising was the Pronunciation section which marks pronunciation errors at three levels of sensitivity: "less sensitive", "default" and "more sensitive". Technically, this solution has the potential to become one of the tools preferred by the teachers for a number of reasons – firstly, it combines the capacity of cloud-based ASR with direct pedagogical focus on pronunciation; secondly, it is available free of charge to any applicable school (which translates to all state schools and state universities in the Polish context); thirdly it can offer pronunciation feedback in 36 languages (Getting started with Reading Progress in Teams, n.d.).

However, it ought to be noted that Reading Progress is not per se a pronunciation training tool. Instead, it aims at building oral reading fluency by focusing on eliminating undesirable performance phenomena such as omissions, insertions, repetitions, self-corrections or the aforementioned mispronunciations. All these criteria are aggregated to Accuracy Rate (expressed in percentage terms), which is coupled with the number of correct words per minute to form the basis for assessing learners' oral reading performance (c.f. Figure 1). Therefore, it may be assumed that Reading Progress is mostly a tool developed for native speakers of a given language. This claim is corroborated by the choice of scientific papers presented by Microsoft as a theoretical background for their innovation (Research Related to Reading Progress, n.d.); all the sources listed on the website refer to the training offered to the context in which English is the language of instruction (e.g., Wexler et al. 2007). However, it is worth emphasizing that Reading Progress is described by Microsoft in the context of the US educational market, and US students do not need to be native speakers of English, using a native-like accent. On the contrary, the proportion of English learners in US public schools was reported, as of 2018, to be 10% nationwide, with up to 21% in California (Bialnik, Scheller & Walker 2018).



Krystal McKinney

Attempts: 2 Level: B Words: 207

41 Correct Words per Minute

83% Accuracy Rate

5 Mispronunciations
2 Omissions
1 Self-corrections
1 Insertions
0 Repetitions

Auto-Detect PREVIEW

Pronunciation sensitivity Low

The study of Earth's landforms is called physical geography. Landforms can be mountai can also be glaciers, lakes, or rivers. Landforms are s ysical features. It is important for students to kno l geography of Earth.

The seasons, th ll the natural processes of Earth affect where people a raphy is one of a combination of factors that pec here they want to live.

The physical fea e often rich in resources. Within a nation, mountai natural borders for settlement areas. In the U.S., major mountain ranges are the Sierra Nevada, the Rocky Mountains, and the Appalachians.

Fresh water sources also influence where people settle. People need water

Figure 1: A teacher view of a Reading Progress task. Note that the value on the pronunciation sensitivity slider (located under the camera feed box) can be re-set even after collecting students' works. Source: Microsoft Corporation

Even if one assumes that the needs of non-native speakers were not considered while designing Reading Progress, the tool might still be useful to learners of English at the university level, as in this context the native-like pronunciation is often one of target outcomes of foreign language instruction. As regards the claim that the tool was developed solely for the youngest learners – i.e. the ones who are still in the process of mastering basic reading skill – it can be refuted on the basis of the fact that Reading Progress offers a library of readings which includes tasks at the level of secondary education. In addition, some research papers cited by Microsoft explicitly target skills development in middle school learners (Rasinski et al. 2009; Paige et al. 2021).

While the creators of Reading Progress do not reveal the exact specifications of their pronunciation assessment module, it is suggested that standard native non-accented speech is the model against which the system compares learners' input. The following passage from the help page also suggests that in the future the tool should become more inclusive in terms of non-standard accents:

> Note: Pronunciation detection for each language is generalized based on common pronunciation and may not recognize accents and dialects well. This is just a starting point, and we are working to ensure those with accents and dialects are included. Use your discretion to mark errors manually when the speech detection does not meet the needs of your student. (Getting started with Reading Progress in Teams, n.d.)

Therefore, it remains unclear whether Reading Progress by Microsoft can be used as a pronunciation training tool for non-native speakers and whether its use might be beneficial for the teachers and learners alike.

3. The study

The aim of our study was twofold; firstly, it was deemed necessary to focus on the quality of feedback provided by Reading Progress. Specifically, we wanted to determine whether this feedback may be a credible source of information for non-native students of EFL. Secondly, it was decided to propose pedagogical recommendations in order to help teachers use the aforementioned tool in the most effective way. Therefore two research questions were posed:

Q1: What is the accuracy of pronunciation-oriented feedback provided by MS Reading Progress to non-native learners of English as a foreign language?
Q2: What are research and pedagogical implications concerning the use of MS Reading Progress?

In the analysis of feedback quality, it was decided to focus on false positives, i.e., instances of feedback which erroneously suggests learners' mispronunciations. We believe that the false positives are potentially more harmful than having one's errors remain unnoticed (false negatives). While the latter problem may result in the lack of progress and subsequent fossilization, the former one can lead to artificially-induced and largely avoidable back-sliding (cf. Han 2004). In addition, false positives may cause frustration in the learner, as they create the impression of unreasonably high pronunciation standards, which may be perceived as "disheartening" (Bajorek 2017: 36).

The reference point for machine-generated feedback was the human feedback provided by academic pronunciation teachers. Both our independent raters were non-native English pronunciation instructors employed at the Institute of English Studies (the University of Łódź). They both had had at least two years of experience in teaching and assessing students' pronunciation during various phonetics-oriented courses.

3.1 Participants

The subjects were 20 BA students of English (16 F and 4 M) who had completed two semesters of *Practical Phonetics*. *Practical Phonetics* course is aimed at raising students' awareness of the importance of pronunciation in English. Students learn about the articulation and discrimination of sounds in English and practice their pronunciation with the use of various techniques and activities. The course also aims at developing the students' ability to read and write using transcribed text (IPA transcription). Among some of the core coursebooks for this course are *Practical Phonetics and Phonology* by Peter Roach, *Ship or sheep* by Ann Baker, as well *English Phonetics for Poles* by Włodzimierz Sobkowiak, or *English Pronunciation in Use series.*

All students were enrolled, at the time of the experiment, in the first semester of a second-level pronunciation course *Prosody and discourse*. At this stage of their English Philology studies, students are expected to be familiar with standard (RP or Modern British English and General American in this case) pronunciation of English vowels and consonants, word stress, sentence stress, aspects of connected speech, etc., and display a certain level of consistency in their pronunciation patterns.

3.2 Data collection

The text to be read by the subjects (see: Appendix 1) was selected from the integrated Teams database offered by a reading-oriented educational platform ReadWorks ("Retrospective Study of Read Works Use and Effectiveness Web Analytics and Survey Findings", 2014). The length of the passage was 489 words, and the difficulty level was categorized as "secondary", which indicates that it was created for US students in the 11-14 age bracket. The text was reviewed by pronunciation teachers to ensure that it did not include any vocabulary that the subjects would find too difficult; it was concluded that despite some specialized words, such as "sulfur" or "dioxide", no items would pose significant problems in terms of comprehension and oral production.

3.3 Procedure

The reading passage was distributed to the subjects as an MS Team Reading Progress Assignment. The feedback provided by MS Teams app was not, however, made accessible to the learners. Instead, they received targeted feedback from the teachers in the form of comments and suggestions. Thus, Teams feedback was only used for research purposes. It was compared against feedback from independent human raters who were specifically asked to mark mispronunciations and other relevant inaccuracies in the same way in which it is done in Reading Progress.

The output from Teams as well as human raters were operationalized in the form of binary numerical data. Words without marking were assigned the value of 0, while the marked words were coded as 1. Since all feedback was provided on the level of orthographic words, it was possible to align the data in one table, as presented in Table 1. The full database is available online (see Appendix 2).

Table 1: Sample structure of the aligned data.

Rater	Word 1	Word 2	Word 3
Human rater 1 (HR 1)	0	0	1
Human rater 2 (HR 2)	0	0	1
Reading Progress	0	1	0

Fleiss' Kappa was used to determine the inter-rater reliability (Koo & Lee 2016; Grant et al. 2017). In addition, qualitative analysis was conducted in order to describe and classify false positives.

4. Results

Since each orthographic word in the text could theoretically be marked as mispronounced, each rater had 9674 opportunities to use the "mispronounced" label. This number multiplied by the number of raters gives the total of 29,292 possible contexts for the use of the aforementioned label. Out of all the context possible, the total number of mispronunciations marked by all the raters was 1413 (≈ 5%). For 8762 words, all the pronunciation was deemed satisfactory by all raters. All three raters agreed that a given word should be marked as mispronounced in 87 cases; two-raters' agreement was recorded 327 times, while the instances where a given word was marked by one rater amount to 498 cases. In terms of the total number of words marked as mispronounced, Reading Progress achieved the lowest score of 305, followed by HR 2 (441) and HR 1

(667). A more detailed breakdown of agreement as to which word was mispronounced is presented in Chart 1:

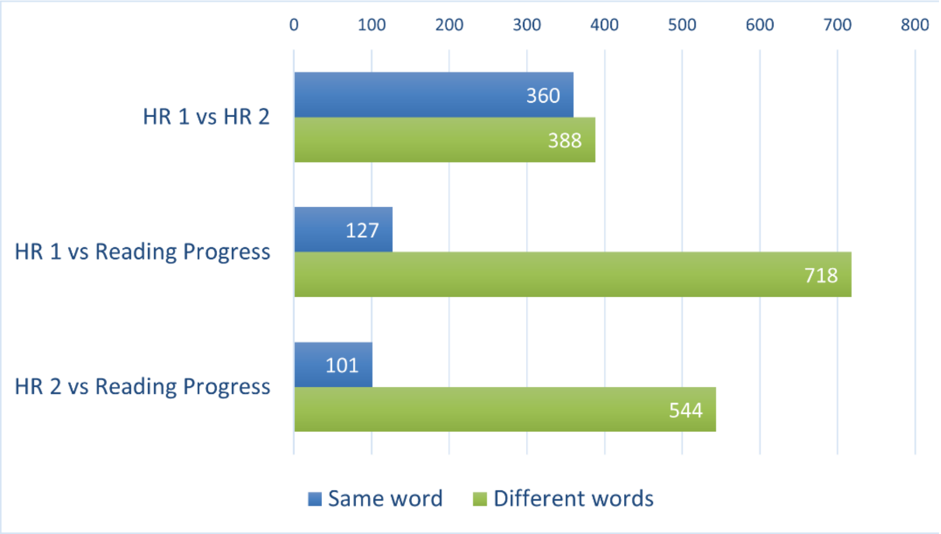


Figure 2: The breakdown of agreement on mispronounced words.

The data presented suggest that while there is no perfect agreement between the human raters, the difference of agreement-disagreement between pronunciation teachers and Reading Progress might be nearly seven-fold in absolute terms. While the variation in human raters can be partly explained by the restrictive format of the feedback (division of speech into orthographic words), the difference observed between raters and the pronunciation module in this case shows that the application in question falls short of achieving the human rater standard. This observation is further confirmed by the results of the Fleiss' Kappa test; the model which includes all raters shows a *moderate* level of inter-rater agreement (according to the classification proposed by Landis and Koch 1977), with κ = 0.43, p < 0.05. The level of agreement between human raters, however, is noticeably higher (κ = 0.64, p < 0.05), and it should be labelled as *good*.

The number of false positives produced by Reading Progress amounted to 129 occurrences. The average number of false positives per student was *M = 8.3*, *SD = 7.36*, *Mdn = 5.5*. The number ranged from 1 false positive (S14, S15) to 26 (S3, S11). The proportion of false positives to all errors reported by MS Reading Progress also varied considerably across the participants. The lowest recorded value was 23% (S5), while the highest was 100% (S10). The average proportion was 59%, with *SD = 19.91%* and *Med = 59%*. The distribution of the values indicating the aforementioned proportion (normalized as per cent) is presented in Figure 3.

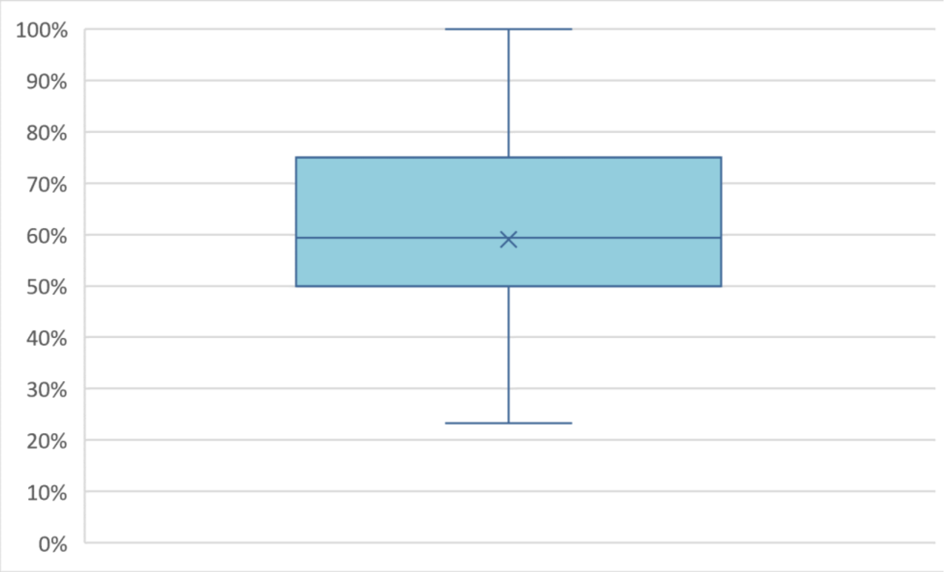


Figure 3: The distribution of values indicating the proportion of false positives to all errors recorded by Reading Progress expressed in percentage terms.

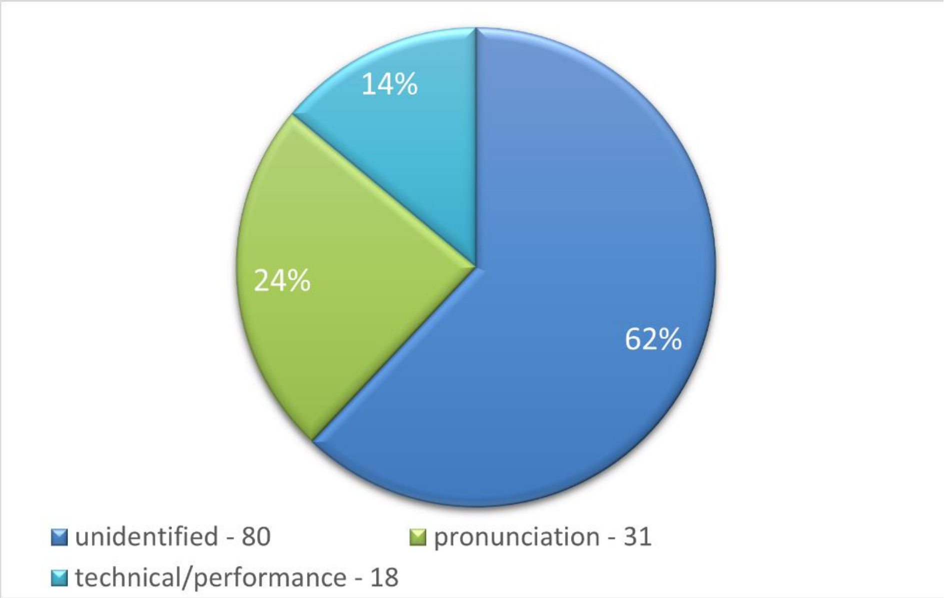


Figure 4: Classification of false positives

For the majority of false positives, the raters were not able to determine the source of the problem. In some cases, slight deviations from the target pronunciation were observed, while performance effects (i.e., very fast pace or self-corrections) and technical issues were suspected in relatively fewer cases. The distribution of false positives across these categories is presented in Figure 4.

While the source of 62% of all false positives was unexplained, the composition of the remaining 38%, presented in Table 2, provides insight into problems that might be encountered while using MS Reading Progress.

Table 2: Suspected pronunciation problems that triggered MS Reading Progress to return false positives, ordered by frequency of occurrence.

Suspected problem	**Number of occurrences**	**Classified as…**
word pronounced very quickly/unclear	13	Technical/performance issue
sound quality (vowel length, consonant quality, final consonant devoicing)	13	Pronunciation issue
glottal stop	7	Non-standard native-like pronunciation
recording quality	6	Technical/performance issue
stress	6	Pronunciation issue
self-correction	2	Technical/performance issue
difficult word (sulphuric)	2	Pronunciation issue

The data indicate that technical aspects might have some impact on pronunciation assessment in Teams. It appears that the system lacks the capacity to correctly process fast speech in certain cases; moreover, the quality of the recording seems to be an important factor as well. It should be noted that Reading Progress (as of 2021) does not provide warnings about poor sound quality to users, which is a standard feature of Teams videoconferencing tools.

In terms of pronunciation issues, it appears that some problems can be attributed to the use of certain native-like forms, such as the realization of /t/ as a glottal stop in words like *might*, in syllable-final position followed by a vowel or a pause, or as an unexploded variant in words like *first*. This was evident in the case of one student with a particularly high number of false positives (26) marked by Reading Progress, who consequently applied elements of connected speech and native-like features of pronunciation.

Another group of problematic words included those with possible, although slight, divergence from the standard in terms of sound quality (13). For example, vowel length (6 instances) in high-frequency function words such as *these* /ð**i**ːz/, but also in lexical words such as *carbon* /ˈk**ɑ**ːbən/ or *particles* /ˈp**ɑ**ːtɪkəlz/. In addition, potential issues were identified in the case of the quality of fricatives or fricative clusters (5) in words like *with* /wɪ**ð**/ or *earth's* /ɜː**θs**/, where the realization of voiceless and voiced dental fricative was closer to a dental stop. This group of words also includes two instances of possible final consonant devoicing (2) in *gases* /ˈɡæsɪ**z**/. As mentioned above, these only displayed a slight divergence from the standard and could be attributed rather to the speed with which they were pronounced.

There were also six instances of word stress related problems in multisyllabic words like *contaminate* /kənˈtæmɪneɪt/, *distances* /ˈdɪstənsɪz/ or *dioxide* /daɪˈɒksaɪd/, where the main stress was not sufficiently prominent or shifted slightly to a neighboring syllable. It was especially clear in the case of the word *dioxide*, which proved to be quite problematic to the participants in general, and which was the second most frequently marked false positive (19 instances), but also frequently mispronounced by students overall.

5. Analysis

The results of Fleiss' Kappa indicate that Reading Progress is not consistent with human raters. More worrying is the fact that this inconsistency can be attributed to pedagogically-harmful false positives. On average, almost 60% of all words marked for mispronunciations by Reading Progress were not considered errors by human raters. Given the fact that university criteria for students' pronunciation are relatively strict – as they should represent the highest possible standards for non-native speakers within the state education system – such a high proportion required further investigation.

Out of all false positives, over 60% were unexplained; however, the remaining cases provide an insight into possible issues that need to be addressed in Reading Progress. The first relevant factor is the quality of the recording. Since there is no low-quality warning, the learners might be convinced that their performance is being recorded correctly, which does not have to be the case. Secondly, high speech rate might play an important part in the process of generating false positives. This is corroborated by the fact that in addition to recorded failures to correctly understand speech produced at a high rate, certain phenomena attributed to fast speech (e.g., slight final consonant devoicing) were identified as possible culprits for false positives. This finding is especially troubling, as a higher rate of speech is generally considered to be related to more native-like performance (Morill et al. 2015).

Non-standard native-like speech was also an issue. The glottal stop, although sub-standard in certain English accents, is not considered to be a non-native feature, and hence its use is not penalized by our academic assessment system. This raises the question of which accent/accents were considered to be the standard for Reading Progress. The system does not specify the variety of English which it uses (as it is done in other MS products, such as MS Office spellcheckers), so a pronunciation teacher would be justified in assuming that multiple native accents were included in the program. On the contrary, it seems that the system does not recognize certain native-like forms.

Finally, there remains the question of idiolect and individual variation in every language user and their pronunciation. We agree that instances described in this category are most likely to have been misjudged by human raters. However, given all the aforementioned problems, including over 60% of unidentified problems, it is possible that at least in some cases the algorithm used is not yet finely tuned to decide on error gravity in terms of mispronunciations. One supporting piece of evidence for this explanation might be the fact that teachers have access to pronunciation sensitivity tools in MS Reading Progress. In the case of this research, the sensitivity was set to "High", as no other setting produced relevant results – i.e., student works did not show any pronunciation errors at all.

6. Conclusions and implications

The data collected during our research seem to suggest that the accuracy of MS Reading Progress is not sufficient for it to be used with foreign language students as a source of reliable machine-generated feedback (**Q1**). In view of our findings, the aforementioned warning by Microsoft about the inability to "recognize accents and dialects well" should be interpreted as a clear indication that the system is not yet ready to fully support pronunciation training, especially in the case of non-native speakers – even if they are advanced learners who strive for a "standard" accent, such as RP or GA.

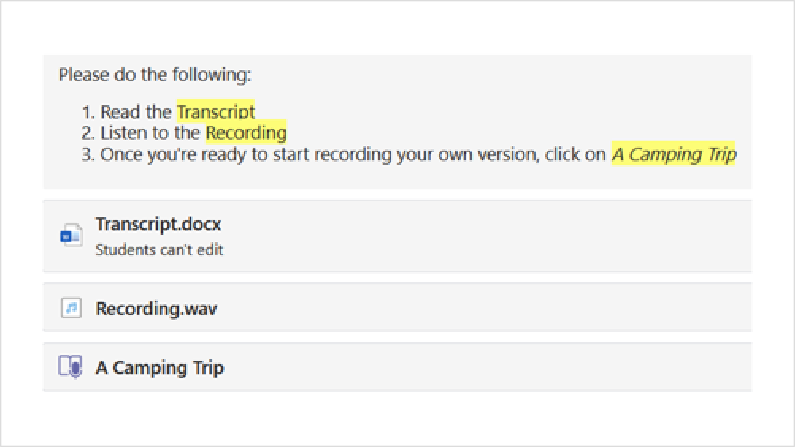
However, the implications for teachers and researchers (**Q2**) should not be limited to negative aspects of Reading Progress, especially if one considers its potential for future improvements. Since it is based on a widely adopted ASR which has access to substantial language samples, and since it is being developed by one of the biggest IT companies, it might still be fine-tuned to deal with accuracy issues. Especially encouraging is the prospect of automated feedback which can be given at three sensitivity levels. It appears that this function has not been given much attention in other CAPT software, while teachers of non-native speakers might be especially in need of graded sensitivity so as not to discourage the beginners.

In terms of design, a noteworthy feature of Reading Progress is the alignment of the text with the recording, which provides instant access to a selected fragment without the need to constantly re-position the slider on the timeline. Owing to this feature, pronunciation instructors do not need to take notes with time signatures; instead, they can manually highlight specific words, which is faster and more convenient for all parties involved in the process of pronunciation training. However, the reliance on orthographic words seems to be limiting for all aspects of reading-aloud skills assessment. It is particularly true in the case of pronunciation, where a number of orthographic words can be pronounced as one phonological word (Dixon & Aikhenvald 2002). Other significant drawbacks are the lack of pronunciation model and the ability to provide solely binary feedback.

While the former problem might be partly amended by uploading a recording with model pronunciation along with the reading-aloud task in MS Teams, the latter one would require further expansion of system functionality to offer custom-made labels, preferably with attachable notes including teacher-generated feedback wherever necessary.

From pronunciation instructors' perspective, the capacity of MS Reading Progress as a CAPT tool still remains limited. As of December 2021, its main advantages are the convenience of automated collecting and storing all assigned recordings in dedicated cloud folders, the alignment of text input and speech output, and the ability to provide manual feedback by means of labels (with Auto Detect deactivated). We recommend that for now it be used mostly in such a setup. However, it is likely that with the inclusion of more accents and dialects the quality of the system increases to the level at which it will be pedagogically sound to use automated pronunciation assessment module. If this is the case, pronunciation sensitivity will be one of the features that sets Reading Progress apart from a number of other solutions.

By contrast, there are certain problems which are less likely to be addressed by Microsoft in the upcoming future. As it was mentioned, it is possible to provide a transcript and a recording of the text to ensure that students are familiar with model pronunciation (see Figure 5), but this solution still seems to be worse than being able to listen to individual words or sentences upon request.



Please do the following:

1. Read the Transcript
2. Listen to the Recording
3. Once you're ready to start recording your own version, click on *A Camping Trip*

Transcript.docx
Students can't edit

Recording.wav

A Camping Trip

Figure 5: Reading task with a transcript and a recording of model pronunciation attached.

Additionally, the manual feedback system seems to be relatively limited in comparison with pronunciation teachers' needs. It is possible to add some general notes to the assignment, but a comment-based system, similar to the one featured in the Office suite, would be more convenient to use for both learners and teachers. In sum, the most valuable feature of Reading Progress may be its automated feedback – provided that it “matures” enough to offer valuable information for language learners.

In terms of implications for research projects, Reading Progress can prove a reliable and time-saving tool for providing manual feedback for individual orthographic words. However, researchers who intend to test the accuracy of the automated pronunciation assessment module are advised to proceed with caution. We believe that in such a case the current quality of automated feedback necessitates pedagogical intervention, should Reading Progress be used with actual language learners. Nonetheless, such research ought to be conducted in the future, especially if Microsoft issues major updates to the pronunciation module.

References

Bajorek, Joan. 2017. L2 pronunciation tools: The unrealized potential of prominent computer-assisted language learning software, *Issues and Trends in Educational Technology, 5*(2).

Baker, Ann. 2007. *Ship or sheep? An intermediate pronunciation course* (3rd ed.). Cambridge University Press.

Bernstein, Jared, Cohen, Michael, Murveit, Hy, Rtischev, Dimitry and Weintraub, Mitchel. 1990. Automatic evaluation and training in English pronunciation. *ICSLP*, 1185-1188.

Bialik, Kristen, Scheller, Alissa and Walker, Kristi. 2018. *6 facts about English language learners in U.S. public schools*. Pew Research Center. Retrieved April 13, 2022, from https://www.pewresearch.org/fact-tank/2018/10/25/6-facts-about-english-language-learners-in-u-s-public-schools/

Chen, Liang-Yu and Jang, Jyh -Shing Roger. 2015. Automatic pronunciation scoring with score combination by learning to rank and class-normalized DP-based quantization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(11), 1737–1749.

Coniam, David. 1999. Voice recognition software accuracy with second language speakers of English. *System, 27*(1), 49–64. https://doi.org/10.1016/S0346-251X(98)00049-9

Cucchiarini, Catia, Strik, Helmer and Boves, Lou. 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication, 30*(2-3), 109–119. https://doi.org/10.1016/S0167-6393(99)00040-0

Davey, Melissa. 2017, August 9. Outsmarting the computer: The secret to passing Australia’s English-proficiency test. *The Guardian*.

Derwing, Tracy M., Munro, Murray J., and Carbonaro, M. 2000. Does popular speech recognition software work with ESL speech? *TESOL Quarterly, 34*(3), 592–603. https://doi.org/10.2307/3587748

Dixon, Robert M. W. and Aikhenvald, Alexandra Y. (eds.) 2002. Word: A cross-linguistic typology. *Phonology, 20,* 425–429. https://doi.org/10.1017/S0952675704210119

Getting started with Reading Progress in Teams. (n.d.). Microsoft. Retrieved December 19, 2021, from https://support.microsoft.com/en-us/topic/getting-started-with-reading-progress-in-teams-7617c11c-d685-4cb7-8b75-3917b297c407?storagetype=live#ID0EDD=Educators

Grant, Malcolm J., Button, Cathryn M. and Snook, Brent. 2017. An evaluation of interrater reliability measures on binary tasks using d-prime. *Applied Psychological Measurement, 41*(4), 264-276. https://doi.org/10.1177/0146621616684584

Han, Jeonghye. 2012. Emerging technologies: Robot-assisted language learning. *Language Learning & Technology, 16*(3), 1–9.

Han, ZhaoHong. 2004. Fossilization: Five central issues. *International Journal of Applied Linguistics, 14*(2), 212–242. https://doi.org/10.1111/j.1473-4192.2004.00060.x

Henrichsen, Lynn E. 2021. An illustrated taxonomy of online CAPT resources, *RELC Journal*, *52*(1), 179–188. https://doi.org/10.1177/0033688220954560

Hodges, Charles B., Moore, Stephanie, Lockee, Barb B., Trust, Torrey and Bond, M. Aron. 2020. *The difference between emergency remote teaching and online learning.* Educase. https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

In-Seok Kim. 2006. Automatic speech recognition: Reliability and pedagogical implications for teaching pronunciation. *Journal of Educational Technology & Society, 9*(1), 322–334. http://www.jstor.org/stable/jeductechsoci.9.1.322

Koo, Terry K. and Li, Mae Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

Landis, J. Richard and Koch, Gary G. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. PMID: 843571.

Levis, John M. 2005. Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377. https://doi.org/10.2307/3588485

Levis, John M. 2007. Computer technology in teaching and researching pronunciation, *Annual Review of Applied Linguistics, 27*, 184–202. https://doi.org/10.1017/S0267190508070098

Li, Kun, Qioan, Xiaojun and Meng, Helen. 2017. Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, *25*(1), 193-207. https://doi.org/10.1109/TASLP.2016.2621675

Research related to Reading Progress. (n.d.). Microsoft. Retrieved December 19, 2021, from https://education.microsoft.com/en-us/resource/9d24e8eb

Marks, Johathan, Hancock, Mark, Hewings, Martin and Donna, Silvie. 2012. *English pronunciation in use* (Vol. 1-3). Cambridge University Press.

Mahdi, Hassan Saleh and Al Khateeb, Ahmed Abdulateef. 2019, The effectiveness of computer-assisted pronunciation training: A meta-analysis. *Review of Education, 7*, 733-753. https://doi.org/10.1002/rev3.3165

Molenda, Marek, Adamczyk, Michał and Rybińska, Paulina. 2018. Beyond the CAPT – Automatic Speech Recognition in pronunciation training. In J. Pitura & S. Sauro (Eds.), *CALL for mobility*, 32–153. Peter Lang. https://doi.org/10.3726/b13451

Morril, Tuuli, Baese-Berk, Melissa., Heffner, Christopher and Dilley, Laura. 2015. Interactions between distal speech rate, linguistic knowledge, and speech environment. Psychonomic Bulletin & Review, 22(5), 1451–1457. https://doi.org/10.3758/s13423-015-0820-9

Nushi, Musa and Sadeghi, Mahsa. 2021. A critical review of ELSA: A pronunciation app. *Computer Assisted Language Learning Electronic Journal*, *22*(3), 287-302.

Paige, David D., Rasinski, Timothy V. and Magpuri-Lavell, Theresa. 2012. Is fluent, expressive reading important for high school readers? *Journal of Adolescent & Adult Literacy, 56*(1), 67–76. https://doi.org/10.1002/JAAL.00103

Pennington, Martha C. and Rogerson-Revell, Pamela. 2019. *English pronunciation teaching and research*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-47677-7

Qian, Xiaojun and Meng, Helen. 2017. A two-pass framework for mispronunciation detection and diagnosis for computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(6), 1020–1028.

Rasinski, Timothy, Rikli, Andrew, & Johnston, Susan. 2009. Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research and Instruction, 48*(4), 350–361. https://doi.org/10.1080/19388070802468715

Ray, Susanna. 2021. *Students have a new, less stressful way to improve their reading – and it's easier for teachers, too*. Microsoft. https://news.microsoft.com/features/reading-progress/

Retrospective study of ReadWorks use and effectiveness Web analytics and survey findings. (2014). Rockman et al.

Roach, Peter. 2009. *English phonetics and phonology: A practical course* (4th ed.). Cambridge University Press.

Rogerson-Revell, Pamela M. 2021. Computer-Assisted Pronunciation Training (CAPT): Current issues and future directions, *RELC Journal, 52*(1), 189–205. https://doi.org/10.1177/0033688220977406

Sobkowiak, Włodzimierz. 2004. *English phonetics for Poles* (2nd ed.). Wydawnictwo Poznańskie.

Souza, Hanna Kivistö and Gottardi, William. 2022. How well can ASR technology understand foreign-accented speech?. *SciELO Preprints*. https://doi.org/10.1590/010318138668782v61n32022

Yaman, İsmail and Ekmekçi, Emrah. 2016. Shift from CALL to MALL? *Participatory Educational Research, special issue 2016-IV,* 25-32.

Wexler, Jade, Vaughn, Sharon, Edmonds, Meaghan and Reutebuch, Colleen Klein. 2008. A synthesis of fluency interventions for secondary struggling readers. *Reading and Writing, 21*(4), 317–347. https://doi.org/10.1007/s11145-007-9085-7

Appendices

Appendix 1 Reading Assignment

Reading Assignment offered by MS Reading Progress is a sample from a larger text that can be found in the ReadWorks library (log-in is required): https://www.readworks.org/article/All-About-Weather/a30f13bf-de1f-4c21-909d-f8c4264e9db6#!articleTab:content/contentSection:a7df6033-091d-405c-ac7a-024f341d69f6/

Appendix 2 Transcribed Data

Data can be accessed online at: [https://uniwersytetlodzki-my.sharepoint.com/:x:/g/personal/marek_molenda_filologia_uni_lodz_pl/EZqsoTI9tzZNk5teGIDvTNUBcss_MQgOycT4VTQSIRLoyg?e=BZErq4]. Each tab represents data gathered for one student. Words which were assessed to have been mispronounced were marked with “1”, while “0” represents words which were assessed to have been pronounced correctly.