From Miscue to Evidence of Difficulty: Analysis of Automatically Detected Miscues in Oral Reading for Feedback Potential

Beata Beigman Klebanov, Michael Suhan, Zuowei Wang, Tenaha O'Reilly
ETS, Princeton NJ, USA
{bbeigmanklebanov,msuhan,zwang,toreilly}@ets.org

Abstract

This research is situated in the space between an existing NLP capability and its use(s) in an educational context. We analyze oral reading data collected with a deployed automated speech analysis software and consider how the results of automated speech analysis can be interpreted and used to inform the ideation and design of a new feature – feedback to learners and teachers. Our analysis shows how the details of the system's performance and the details of the context of use both significantly impact the ideation process.

1 Introduction

Reading a text fluently – accurately and with a good speed – is evidence of development of foundational reading skills, such as decoding and word recognition (Sabatini et al., 2019). Research suggests that the development of oral reading fluency is an essential bridge from decoding to comprehension of text (Pikulski and Chard, 2005). Instructional approaches to foster fluency include modeling fluent reading to the developing reader; repeated reading (re-reading passages multiple times); and engaging students in wide independent reading (Ardoin et al., 2016; Hudson et al., 2020; Pikulski and Chard, 2005; Wexler et al., 2008). Extensive reading was also shown to support fluency development in students of English as a foreign language (Huffman, 2014; Suk, 2017).

Given the importance of fluency for reading development, we built Relay Reader™, an app[1] where readers can practice by taking turns reading out loud from full-length stories with skilled audiobook narrators (Madnani et al., 2019). The narrator reads a passage while the user follows in the text; then the user reads the next passage aloud, and so on. Users can set the word counts of their own and narrator turns between 70 and 200 words.[2] The app has been available since 2020 to readers-in-the-wild, initially with one story (an English translation of Collodi's *The Adventures of Pinocchio*) and gradually expanding to 26 stories, from a 460-word fable to a 120K-word novel. In parallel, the app has been used for independent reading ('Drop Everything And Read') in school and summer camp contexts.

During the development of the app, we conducted a needs assessment with teachers which showed that obtaining estimates of students' oral reading fluency and accuracy was the top priority, followed by being able to see students' specific difficulties in reading through miscue analysis (Kannan et al., 2019). Accordingly, a speech analysis system was developed and is currently used to provide fluency information to teachers. Fluency is measured as words read correctly per minute; hence the system transcribes the audio and compares to the passage text in order to provide fluency estimates. As a byproduct, the system produces an alignment between the transcript and the passage from which the miscues can be easily recovered.

The goal of this study is to explore the potential of using these miscues for feedback, guided by the following **research questions**: (1) What is the extent of miscues in the data? (2) How are miscues distributed in reading passages? (3) How does the extended reading context come into play? (4) How reliable is miscue detection?

The main contribution of our work is the exploration of the space between a deployed NLP capability and its use case. We show how the analysis of the data collected through the system can support ideation of using the system in a new way – for feedback, specifically regarding frequency and content of such feedback. More generally, in the context of using NLP for building educational applications, we zoom in on the process of **ideating**

[1] The app is available freely on https://relayreader.org/.

[2] These are approximate since turn transitions happen on paragraph breaks only.

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 459–469
June 20, 2024 ©2024 Association for Computational Linguistics

a new feature in an existing ecosystem, and show how the analysis of existing data can inform the ideation process.

2 Related work

Research on automated speech recognition (ASR) for young readers suggests that misreadings and slow reading constitute significant challenges (Gelin et al., 2021; Wu et al., 2019); focus on subword units (Hagen et al., 2007) and data augmentation with synthetically generated mistakes (Gelin et al., 2023) are some of the approaches proposed to improve identification of misreadings. The technical challenges notwithstanding, ASR has long been used for feedback in automated reading tutors. The Reading Tutor from Project LISTEN, an influential early system that entered classrooms in the 1990s, displayed the text one sentence at a time. As the student read, the system interrupted if a word was read incorrectly and not self-corrected by underlining the incorrect word and occasionally “coughing" to get the student’s attention (Mostow and Aist, 1999). Lalilo is another reading tutor for early elementary students. Students record themselves reading a word, phrase, or sentence; their recording is played back, followed by a fluent model of the sentence. The reader gets feedback when the system is confident that it was correct (‘Perfect’) or incorrect (’Try again’); if uncertain, the student is asked whether their recording matched the fluent one and is encouraged with ‘Good job!’ (Hembise et al., 2021). BookBuddy is a chat bot that converses with young readers about the story they are reading by answering their questions, quizzing them, and automatically evaluating their spoken answers (Ruan et al., 2019). The Charlesbridge Reading Fluency program ‘listens’ as a student is reading, and when a child misreads or struggles with a word, the machine models it and asks the child to repeat it and continue reading; problem words are marked in a separate report for review and practice (Adams, 2013). The virtual reading tutor Marni tracks the student while reading aloud by moving the cursor to each word as it is spoken (Cole et al., 2007). A reading tutor for Dutch supports reading individual words, word lists, and short stories; for the latter, the student is asked to reread the sentences where they read incorrectly one or more words, as detected by ASR software (Bai et al., 2020).

In general, prior work on automated fluency support tends to focus on very young learners (K-2) and on an early stage of fluency development, using words, sentences, or, at most, very short stories, and on helping the student get every item right. In contrast, Relay Reader is targeting a *more advanced* stage of fluency, with a focus on *immersive extended* reading. In this context, it may not be very important to get every word right, especially if it comes at the cost of breaking the flow of reading. Still, detailed speech analysis data similar to that available in reading tutors can be obtained and can therefore be used for stakeholder feedback. This work is a preliminary investigation towards designing miscue-based feedback appropriate to the extended reading application.

3 Data

The data for this study come from users-in-the-wild and from study participants in school and summer programs. Users-in-the-wild may choose to respond to a few demographic questions during app sign-up – who the target reader is (self, child, student, other) and whether the reader is a native speaker of English. Non-native speakers using the app themselves is the largest group, followed by native-speaking children. Study participants in schools and summer camps were predominantly upper elementary students (grades 3-5) in the North-East of the USA at schools and camps catering to majority African American and Hispanic students. Different books were added to the library at different times and received more or fewer readings, depending on study designs and reader interest.

We start with a subset of the data with reasonably complete readings, that is, recordings where at least 70% of the words of the passage were found in the automated transcription (reading accuracy $\geq$ 70%). The 70% cutoff helps filter out data that is unlikely to be useful for studying reading errors, for two reasons: (a) Low accuracies often correspond to cases where large stretches of the passage are left unread (skipped) or to very noisy recordings; feedback in such cases, if any, might have to focus on improving engagement in the reading activity or on improving the quality of recordings, rather than on mispronunciation of specific words. (b) The automated speech analysis is less reliable on low-accuracy recordings (Beigman Klebanov and Loukina, 2021). More information about the system cam be found in Loukina et al. (2019). The system produces fluency estimates that correlate with those obtained using human transcribed data at $r = 0.94$ for recordings above the 70% cutoff (Beigman Klebanov and Loukina, 2021).

The resulting dataset (**ByPassage**) consists of 9,432 recordings by 293 readers of 2,009 unique passages from 24 books. These recordings cover 7,511 word types (unique words) and 136,450 word tokens (all occurrences of the words). Table 1 shows descriptive statistics. Average passage length is 109 words (sd = 37.5) and average accuracy is 91.1% (sd = 8.2). The population distribution of the recordings is: 73.1% school, 9.5% summer camp, and 17.4% users-in-the-wild.

To detect miscues, we use the automated alignment of the recording to the text of the passage generated as part of the accuracy computation (Loukina et al., 2019); we consider as miscues all deletions of words in the passage and all substitutions of words in the passage with other words; insertions of words that were not in the passage were ignored.

4 Patterns of Miscue Occurrence

Reading accuracy, namely, the proportion of words read correctly out of all words in a passage, averaged 91.1%. That is, readings of about 9 in 100 words are miscues; this answers RQ1.

To answer RQ2, we investigate whether miscues tend to cluster together. To determine the proximity of errors to one another, we cluster errors occurring within five tokens of each other with the condition that tokens in a cluster must be part of the same paragraph. Thus, the sequence ECEECCCCECCE, where E stands for error and C for correct, will be considered as one cluster, since there is no stretch of more than four Cs in the sequence. We find that while 32.4% of errors occur singly, most errors are proximal to other errors (see Table 2). On average, clusters have 3.6 errors and span 4 tokens; see Table 3. Thus, errors tend to occur immediately next to each other; patterns like ECCCECCCE are uncommon. Since there are, on average, 9.8 errors per passage and these tend to occur in clusters of 3.6, an average passage would contain 2 or 3 error clusters. The following examples, from *Pinocchio* and *Hansel and Gretel*, respectively, show typical occurrences, with cluster boundaries enclosed in brackets and miscued words denoted in bold:

1. And growing angrier each moment, they went from words to blows, and **[finally began** to **scratch]** and bite and slap each other.
2. The man's **[heart]** smote him heavily, and he thought: "Surely it would be better to share the last **[bite with one's]** children!"

We observe that 5.2% of the errors occur in clusters of 11 errors or more (see Table 2), with the largest cluster consisting of 57 errors. Inspection of the largest cluster, which occurs after about four minutes of reading, reveals that the reader did not read aloud the final paragraph of a long passage.

5 Extended Reading

The app contains a mix of short and long stories, including novels, such as *The Adventures of Pinocchio* and *The Wonderful Wizard of Oz*, each with about 40K word tokens. A novel is different from a sequence of short stories that amount to a similar overall word count in that there tends to be continuity of characters, relationships, and settings throughout the story, with the corresponding repetition of key vocabulary. For example, the word *marionette*, a generally infrequent word, repeats 185 times in *Pinocchio*. Such frequency of occurrence, sometimes in narrator turns and sometimes in reader turns, would provide a lot of opportunities for readers to hear the model performance of the word as well as to practice reading it themselves. The interleaved reading activity itself thus constitutes a kind of feedback to the reader, albeit not immediate and indirect: Frequently occurring words may be self-corrected in subsequent encounters, perhaps making immediate corrective feedback to the reader unnecessarily intrusive.

For our next analysis, we use readings from readers who completed *Pinocchio*, the most read book in the app. For every word type in the book, we collect all its readings from those readers who misread it at least once; these are readers who have correction potential since they made a mistake on the word. Words with fewer than five such readers are discarded. The dataset **Pinocchio** has 19,763 readings of 631 word types read by 47 readers.

Each point in the plot in Figure 1 corresponds to a word type; the size of the dot corresponds to the total number of readers with correction potential. On the x-axis is the $\log_2$ total number of occurrences of the word in the book. On the y-axis is the proportion of readers with correction potential who had at least one correct reading of that word. We start with x = 1, since for x = 0 (one occurrence in the book) it is always the case that y = 0.

Statistic	% Correct	#Words Read Correctly	#Words Read Incorrectly	#Words in the Passage
Mean (SD)	91.1 (8.2)	99.3 (35.5)	9.8 (10.0)	109.1 (37.5)
Mode	100	90	0	99
[Min, Max]	[70.1, 100]	[4, 443]	[0, 74]	[5, 444]
[25%, 50%, 75%]	[85.9, 93.5, 98.0]	[79, 94, 114]	[2, 7, 15]	[90, 101, 124]

Table 1: Descriptive statistics of the reading passages (ByPassage dataset), N = 9,432.

#Errors	Freq.	%	Cumulative %
1	8,284	32.4	32.4
2	5,019	19.6	52.0
3	3,493	13.7	65.7
4	2,302	9.0	74.7
5	1,559	6.1	80.7
6	1,162	4.5	85.3
7	889	3.5	88.8
8	651	2.5	91.3
9	506	2.0	93.3
10	388	1.5	94.8
≥11	1,329	5.2	100

Table 2: Distribution of error clusters (N = 25,582) by number of errors in the cluster.

The Figure suggests that generally the more occurrences in the book, the higher the chances of readers figuring out the correct reading even without explicit corrective feedback. We observe that the area to the right of x = 4.32 (20 occurrences or more) and under y = 0.9 (<90% of readers with correction potential with at least one correct reading) is empty, with the exception of the word *would*. As a rough estimate, it seems that about 20 occurrences suffice for the word to largely stop being a problem. We checked this threshold on *The Wizard of Oz* data extracted similarly to the *Pinocchio* data (11,224 readings of 480 word types read by 36 readers) and found it violated by only two words.

These observations suggest that we may want to concentrate the explicit corrective feedback on words that do not occur frequently enough in the book to make self-correction through repeated exposure a near certainty. This would mean that the actual proportion of miscues that are candidates for explicit feedback to the reader may be lower than the 8.9% overall estimate. Removing words with at least 20 occurrences in a story from the list of candidates for explicit feedback for that story, we observe that the proportion of feedback-eligible miscues goes down from 8.9% of all word tokens to 3.3%, for the ByPassage dataset. For an average passage of 109 words, this would correspond to about 3.5 miscues eligible for correction per passage, on average, instead of 9.8. This reduction in the number of miscues eligible for correction is an affordance of the extended reading context; this finding, therefore, answers RQ3.

6 Reliability of miscue detection

Before designing feedback to readers or teachers based on automatically detected miscues, we estimate how reliably the system points out miscues (RQ4). In particular, our focal measure is precision of miscue detection – if a system declares an error, which would presumably trigger feedback, how often is there indeed an error?

We considered words with 50% or lower %Correct, reasoning that these would be likely loci for error flagging. There were 87 such words that were read by at least 10 readers each. We excluded 12 non-dictionary words that may not have a standard pronunciation.[3] Table 4 shows the statistics of the **ByMiscue** sample. These words are generally infrequent, occurring no more than 6 times in the corpus of 24 books. Table 5 lists the words, the number of readings and readers per word, and the titles of the books that included the words.

For every one of the 75 word types, we randomly sampled 3 readings where the machine classified the reading as 'correct' and 3 readings classified as 'incorrect'. In cases with fewer than 3 predicted 'correct's, we used all the instances the machine deemed 'correct' (2 or 1). There were 446 cases in total for the 75 words, of which 221 had the machine's prediction of 'correct' and 225 'incorrect'.

A trained linguist with experience in analysis

[3] The system used human-provided phonetic transcriptions for these words as 'correct' pronunciations during the recognition step, but deviation from that may not be clear-cut cases of miscues. These were the excluded words: 'I, 'this, E, h'm, pep-pe, tchee, zik, ziz-zy, zum, zuz-zy, pi-pi-pi, sha'n't.

Statistic	#Errors per cluster	Cluster span (#words)	#Clusters per passage
Mean (SD)	3.6 (3.6)	4.0 (4.2)	2.7 (2.1)
Mode	1	1	2
[Min, Max]	[1, 57]	[1, 58]	[0, 14]
[25%, 50%, 75%]	[1, 2, 5]	[1, 3, 5]	[1, 2, 4]

Table 3: Descriptive statistics of # errors and # consecutive word tokens in error clusters (cluster spans).

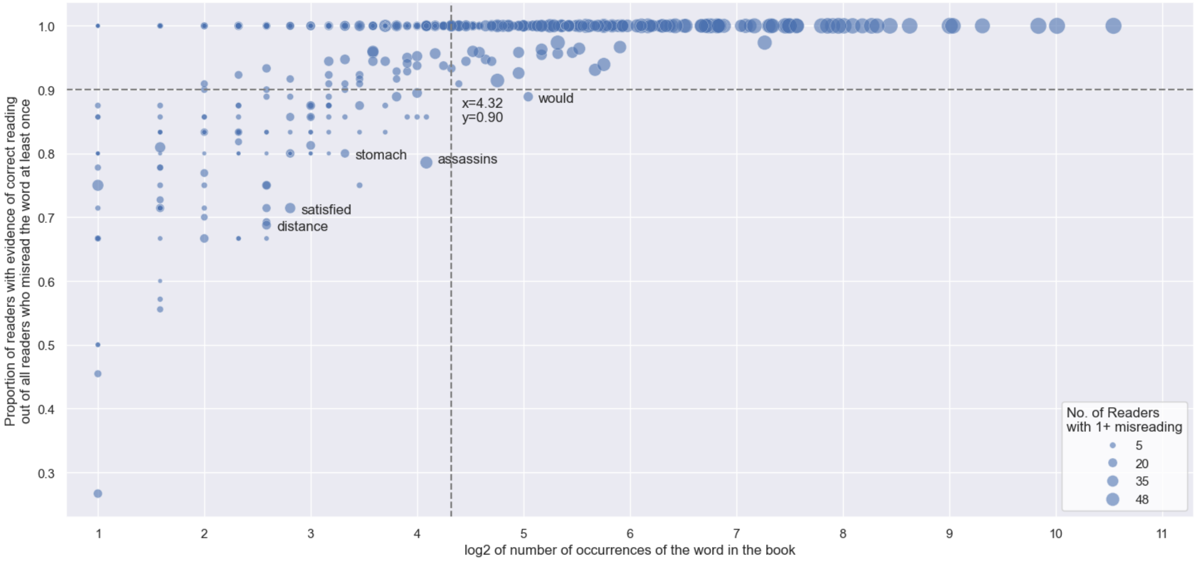

Figure 1: Plot of the relationship between the frequency of a word's occurrence in *Pinocchio* and the proportion of readers who provided a correct reading for the word out of all readers who misread the word at least once. n = 631.

Statistic	Readers	Readings	Tokens
Mean	19.00	21.01	1.95
Median	15	16	1
Mode	10	10	1
SD	10.52	13.40	1.27
Min	10	10	1
Max	55	66	6

Table 4: Descriptive statistics for the ByMiscue sample that covers 75 of the most misread word types.

of oral data (one of the authors of the paper) has listened to the 446 recordings of passages containing the target words, and marked the readings of the target word as 'correct' or 'incorrect'. Table 6 shows the human-machine confusion matrix. The human rater could not make a judgment for 10 instances; these all show as disagreements, equally split between the off-diagonal cells. For 'incorrect' classifications, machine precision was 0.66, recall was 0.65, and the F1 score was 0.66. Thus, about 1 in 3 'incorrect' classifications are false positives – predicting error where there was none.

While performing the annotation, we observed that even when the final execution of a word was correct, there were often indicators that the reader was having some difficulty, such as pausing right before or right after the word, making one or more mistakes leading to the word, or repeating part of the word (e.g., *a fresh convul-convulsion seized her*). The reader's difficulties may manifest in the acoustic signal and, in turn, make it more difficult for the machine to tell whether the reading was correct or incorrect.

We therefore considered a different construct for analysis – that of 'evidence of difficulty' vs. 'no evidence of difficulty' – for the human classification. All instances marked by the human as 'incorrect' in the previous round were labeled as 'evidence of difficulty' by default, whereas the 'correct' cases were further classified into cases with or without evidence of difficulty. Comparing human classification of 'evidence of difficulty' / 'no evidence of difficulty' to the machine's 'incorrect' / 'correct' classification (see Table 7), we found that in 80% of the cases where the machine declared an 'incor-

Word	% C.	R-ngs (R-rs)	Bks
inseparable	10	10(10)	G
scuttling	10	30(30)	P
aristocratic	20	14(14)	B
melodious	20	10(10)	G
zest	20	10(10)	B
pheasants	21	57(39)	P
intuitions	21	14(13)	B
caressed	23	56(55)	P
convulsion	25	12(12)	B
personified	27	11(11)	G
imperturbably	29	14(14)	B
whitened	30	10(10)	G
Pulcinella	31	39(36)	P
indigestion	32	66(53)	P
gold-piece	33	46(46)	P
certainty	33	12(12)	B
Turkish	33	12(12)	B
ventriloquist	33	15(15)	O
astonishment	35	20(20)	PBGR
deductions	36	14(13)	B
sewn	36	14(14)	O
distinguishing	36	11(11)	G
immoderately	36	11(11)	G
sagacity	36	11(11)	G
bedgraggled	37	30(30)	P
bewilderment	37	19(17)	BO
forbearance	40	10(10)	G
persecutors	40	25(25)	P
saucily	40	15(15)	O
magicians	40	10(10)	O
disconsolate	40	10(10)	G
studded	40	15(15)	O
assistance	41	17(17)	BG
caress	41	22(22)	P
carabeneers	41	22(22)	P
amusing	41	22(22)	P
spit	41	64(33)	P
convulsed	42	12(12)	B
excursion	42	12(12)	B
pauper	43	14(14)	B
maliciously	43	21(21)	H
writhed	43	21(13)	GB
slats	43	14(14)	O
partridges	44	27(27)	P
deceived	44	16(16)	O
satin	44	16(15)	NO
perspiration	44	25(24)	P
exquisite	45	31(21)	EG
singed	46	11(11)	O
digested	46	26(14)	P
mastiffs	46	22(22)	P
perpendicular	46	11(11)	G
pursuers	47	19(17)	P
courteously	47	32(30)	PE
sensibly	47	36(36)	P
ferocious	47	15(15)	B
Hippodamia	47	19(14)	G
mysterious	48	21(21)	P
jeeringly	48	21(21)	H
tinsmiths	48	29(14)	O
exhausted	48	31(25)	PAO
spectacle	49	63(52)	P
fancied	49	33(28)	PGR
reproached	50	22(22)	H
perplexity	50	16(16)	GO
brocaded	50	10(10)	O
countless	50	12(12)	O
crocuses	50	14(14)	B
disgustedly	50	22(22)	P
keenly	50	10(10)	P
distinctly	50	10(10)	G
severely	50	12(12)	O
mosquito	50	16(16)	P
stammering	50	24(21)	P
sprightliness	50	10(10)	G

Table 5: 75 most miscued words. %C.: % Correct readings. R-ings(R-rs): #Readings (#Readers). Bks: the source books, from Project Gutenberg: *The Adventures of Pinocchio* by Collodi (P), *the Wonderful Wizard of Oz* by Baum (O), *The Gorgon's Head* by Hawthorne (G), *The Adventure of the Speckled Band* by Conan Doyle (B), *Hansel & Gretel* by Lang (H), *The Necklace* by Maupassant (N), *The Emperor's New Clothes* by Lang (E), *Martin Guerre* by Dumas (A), *Pride & Prejudice* by Austen (R).

Human Machine	Correct	Incorrect
Correct	142	79
Incorrect	77	148

Table 6: Confusion matrix for correct/incorrect human vs machine classification.

Human	No Evidence	Evidence
Machine	of Difficulty	of Difficulty
Correct	106	115
Incorrect	44	181

Table 7: Confusion matrix where machine's correct/incorrect classification is compared the the human's no evidence of difficulty / evidence of difficulty classification.

rect' reading, the human annotator found 'evidence of difficulty' (precision = $\frac{181}{181+44}$ = 0.80); recall was 0.61, and the F1 score was 0.70. Thus, the machine's prediction of 'incorrect' is capturing the human construct of 'evidence of difficulty' with higher precision than the human construct of an 'incorrect' reading.

To confirm the reliability of these findings, a second annotator unrelated to the project with a master's degree in applied linguistics and prior experience annotating speech and oral reading data annotated a reliability sample of 90 randomly selected recordings out of the 446 (about 20%) for (1) correctness of the reading of the target word, and (2) for those items marked as correct, whether there is evidence of difficulty (Appendix A shows the annotation protocol). Cohen's κ between raters for the 3-way classification (incorrect, correct with evidence of difficulty, correct without evidence of difficulty) was 0.604; it was nearly the same (0.601) for a binary classification where 'incorrect' and 'correct with evidence of difficulty' were combined into a single 'evidence of difficulty' class and contrasted with the 'correct with no evidence of difficulty' class.

Using the 90 instances annotated by the second annotator, we also confirmed that the machine's precision was higher in detecting the second rater's 'evidence of difficulty' annotations than the second rater's 'incorrect' annotations (precision of 90% for 'evidence of difficulty' and 84.2% for 'incorrect'). The precision for the first annotator's data for the same subset of 90 instances was 84.2% vs 76.3% for the two constructs, respectively.

To summarize: Our analyses suggest that the machine's detection of a miscue corresponds more precisely to what a human listener would consider as a reading showing evidence of difficulty (80.4% precision) than to what a human listener would designate as a miscue (65.8% precision). This is because readers sometimes ended up reading the word correctly, perhaps after an initial stumble or a partial reading, or recovering from a misreading of a few words just before the current one; the machine often did not recognize these as correct readings.

7 Discussion: Implication for feedback ideation

Our analysis of the automatically transcribed read aloud data from an interleaved book reading app shows a substantial extent of reading difficulty in the readers: About 9% of all word readings in the eligible transcripts show evidence of difficulty based on an automated analysis. The actual extent of difficulty, as detected by a human listener, is likely to be higher, since, while the system shows fairly high precision (0.80) in detecting what a human listener would consider evidence of difficulty, it misses many such cases, since the recall stands at 0.61. Inspecting the patterns in about 9.5K recordings by 293 readers of 2K unique passages (excerpts from novels and short stories), we observed that reading difficulties tend to cluster in 3-4 consecutive words, suggesting that corrective feedback to the reader may need to contain a model performance of whole phrases rather than individual words.

Further, we examined the interleaved extended reading and listening itself as a kind of delayed (not immediate) and indirect feedback to the reader that does not require to break the flow of reading. We estimated that a word that occurs 20 times or more in the book is likely to have sufficient exposure in narrator and reader turns for 90% of the readers who misread it at least once to also produce at least one correct reading. Assuming that there is no urgency that the reader learn a particular word now instead of a few chapters later, we may want to forgo giving the reader direct feedback on misreadings of words that will almost certainly get fixed by the time the reader finishes the story, focusing instead on misreadings of words that do not get repeated very often in the story.

Finally, when designing the feedback based on automatically detected misreadings, it is important to keep in mind that, at least with the speech recognition technology currently implemented in the app and the type of data typical of this use case (no acoustic control of the environment, consumer-level devices and headsets), the detection of miscues is only 66% precise.[4] However, in 80% of the cases where the machine flags a miscue, there is evidence that the reader is having some difficulty – whether or not they produced a correct reading in the end. The ideation and design of feedback will need to reflect this shift in the construct. This finding also suggests that, in terms of learner modeling, automatically detected reading errors provide evidence not only on the knowledge dimension, but also on a behavioral dimension – miscues flagged by the system may provide a first-cut detection of loci where evidence of multiple attempts, self-corrections, pausing to consider the difficult word, and other behaviors related to the trait of perseverance may be found, upon further analysis.

As a first step in exploring feedback to the teacher based on evidence of difficulty, we created class-level heatmaps per paragraph for an ongoing reading of *Pinocchio* in a 4th grade classroom and sent the teacher the heatmaps for the paragraphs that were most difficult for the class, one per chapter. In an interview, the teacher described her use of the heatmap shown below. She told the students she was showing them a challenging passage and explained the darker red as standing for more readers having a difficulty. She told students that some of it was a bit of a tongue-twister for any reader (she said she would have had a hard time herself); she then praised the class for reading much of the passage well and for giving the more challenging part a go. The class also had a brief discussion of what *a gold-piece indigestion* meant. The teacher thus used the feedback not only for providing a correct reading of the miscue cluster "gold-piece indigestion" that occurred in many of her students' readings, but also for a brief but rich motivational, affective, and comprehension-related activity.

> Pinocchio ate **least** of all. He asked for a bite of bread and a few nuts and then hardly touched them. The poor fellow, with his mind on the Field of Wonders, **was suffering from** a **gold-piece indigestion**.

8 Conclusion

In this paper, we analyzed miscues detected by an automated speech analysis system deployed through a publicly available reading app where readers take turns reading books out loud with a pre-recorded skilled narrator. The impetus for considering miscues came from teachers' request to provide such data; however, it is not a-priori clear (a) what the extent of the target behavior is; (b) in what patterns it occurs that may suggest certain ways of designing feedback, (c) how the design of the reading activity may impact feedback, and (d) exactly what a flagged miscue is communicating.

Our analysis shows that miscuing, or, rather, readers experiencing some difficulty reading the word, even if they do pronounce it correctly in the end, is extensive – about nine in a hundred words and possibly more, since our automated system does not detect all such cases. Second, problems tend to cluster together, suggesting that corrective feedback may better be presented by modeling the reading of a phrase rather than of individual words. Third, in order to minimize the interference with the flow of reading, one may want to prioritize modeling misread words that do not occur very frequently in the book. We found that words with 20 or more repetition are very likely to be learned through the interleaved reading activity itself, without additional explicit feedback.

Finally, examining the reading instances flagged by the system as miscues, we found that these are not necessarily incorrectly read words but is closer to what a human listener would consider as evidence of some difficulty on the reader's part, whether or not the word came out correct in the end. This opens up the possibility of considering the automatically detected miscues as a first-cut detection of instances of reader struggles – not only those that manifest as an error but also those that show gearing-up or preparation (pause before), persistence (multiple attempts), or self-correction (successful final readings) – all providing evidence not only on the skill dimension, but also on important learner traits such as perseverance.

From the point of view of feedback development, our analyses suggest that when designing feedback to the reader, it may be incorrect to start

[4] It is interesting that the system nevertheless provides reliable fluency estimates – estimates of words read correctly *per minute*. It may suggest that the impact of crediting or not those words that came out correctly after some struggle is relatively small, considering that the struggle itself has taken time without emission of correct words, which is appropriately captured as detrimental to fluency.

from the vision of "give feedback on every miscue." First, there may be too many of them. Second, it would make sense to fuse feedback on multiple miscues since they tend to cluster together. Third, some of the miscued words have a verifiably high chance of getting fixed during the activity without explicit feedback. Finally, the content of the feedback would not actually be a correction of a miscue, because there may not have been an actual miscue – or misreading – to begin with; it may have been a successful reading following some struggle. This shift in the construct suggests feedback not only, or not necessarily, along the dimension of reading skill, but also learner traits such as perseverance. Our first trial with a teacher shows promise in that the teacher was able to use the evidence of difficulty feedback not only for a corrective purpose, but also for a motivational and affective one.

More broadly, our case study shows how a detailed examination of existing data from a new angle may provide new insights into the performance of the system and support ideation of a new use of the system to the benefit of the stakesholders.

9 Limitations

In this study, data was not separated by characteristics of readers that might impact the kind of mistakes they are making during reading. For example, we did not consider the possible effects of age, linguistic background, or learning disabilities, since we know relatively little about users-in-the-wild and about readers in informal contexts, beyond the general description provided earlier. In addition, it is possible that the automated system performs more accurately on data from certain kinds of readers than from others. For example, recordings from soft-spoken readers or readers with speech disorders may be more difficult to analyze accurately. Different kinds of performances may also be easier or harder to handle; we have anecdotal evidence that particularly creative performances – such as a reader singing the passage – might be difficult for automated analysis.

In the current study, we investigated relatively large-scale patterns in order to identify important considerations for feedback ideation; specific designs will need to be informed by more nuanced analyses of use cases, user populations, and personal reading histories of the users of the application.

10 Ethics statement

Data collections at all the school and summer camp sites were approved by our institution's IRB. The users-in-the-wild agree to the Terms of Use (`https://relayreader.org/terms`) during sign-up into Relay Reader, including the following statement that appears on the Terms of Use summary page displayed prominently during sign-up: "ETS collects voice recordings and other data from users of the App. The recordings and usage data are used in an anonymized manner in connection with ETS research," followed by a link where more information about the research can be obtained. If the application is being installed by parents or teachers for their children and students, respectively, the following statement (that also appears in the Terms of Use summary) additionally applies: "If I am downloading this App for use by my child or student, I have the authority to permit ETS to collect the recordings and usage data as described in the Terms of Use." Our organization's Privacy Policy is linked from relayreader.org and is available here: `https://www.ets.org//legal/privacy.html`.

The data is oral reading data of stories in the Relay Reader app and process data from the app. As such, it is not expected to contain content such as the reader's name, thoughts or opinions, and, indeed, this has not been observed in the data inspected in detail (ByMiscue sample). We have not taken additional steps to check whether the data that was collected contains any information that names or uniquely identifies individual people or offensive content. The data collected by the app is securely stored and managed in accordance with our organization's Privacy Policy.

All the stories and narrations used in the App are either in the public domain (in which case the texts are sourced from Project Gutenberg and the narrations are sourced from LibriVox.org, a collection of volunteer public domain recordings of public domain books), or licensed from the copyright holders.

Acknowledgements

We would like to thank Dr. Burcu Arslan for her insightful comments on an early draft of this paper.

References

Marilyn Jager Adams. 2013. The promise of automatic speech recognition for fostering literacy growth in children and adults. In *International handbook of literacy and technology*, pages 109–128. Routledge.

Scott Ardoin, Katherine Binder, Tori Foster, and Andrea Zawoyski. 2016. Repeated versus wide reading: A randomized control design study examining the impact of fluency interventions on underlying reading behavior. *Journal of School Psychology*, 59:13–38.

Yu Bai, Ferdy Hubers, Catia Cucchiarini, and Helmer Strik. 2020. ASR-based evaluation and feedback for individualized reading practice. In *Proceedings of INTERSPEECH*, pages 3870–3874.

Beata Beigman Klebanov and Anastassia Loukina. 2021. Exploiting structured error to improve automated scoring of oral reading fluency. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, pages 76–81.

Ronald Cole, Barbara Wise, and Sarel van Vuuren. 2007. How Marni teaches children to read. *Educational Technology*, 47(1):14–18.

Lucile Gelin, Morgane Daniel, Thomas Pellegrini, and Julien Pinquier. 2023. Comparing phoneme recognition systems on the detection and diagnosis of reading mistakes for young children's oral reading evaluation. In *Proceedings of Speech and Language Technologies in Education (SLaTE)*, pages 6–10.

Lucile Gelin, Morgane Daniel, Julien Pinquier, and Thomas Pellegrini. 2021. End-to-end acoustic modelling for phone recognition of young readers. *Speech Communication*, 134:71–84.

Andreas Hagen, Bryan Pellom, and Ronald Cole. 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication*, 49(12):861–873.

Corentin Hembise, Lucile Gelin, and Morgane Daniel. 2021. Lalilo: A reading assistant for children featuring speech recognition-based reading mistake detection. In *Proceedings of INTERSPEECH, Show & Tell contribution*.

Alida Hudson, Poh Koh, Karol Moore, and Emily Binks-Cantrell. 2020. Fluency interventions for elementary students with reading difficulties: A synthesis of research from 2000–2019. *Education Sciences*, 10(3):52.

Jeffrey Huffman. 2014. Reading rate gains during a one-semester extensive reading course. *Reading in a Foreign Language*, 26(2):17–33.

Priya Kannan, Beata Beigman Klebanov, Shiyi Shao, Colleen Appel, and Rodolfo Long. 2019. Evaluating teachers' needs for ongoing feedback from a technology-based book reading intervention. In *Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME)*, Toronto, ON, Canada.

Anastassia Loukina, Beata Beigman Klebanov, Patrick L Lange, Yao Qian, Binod Gyawali, Nitin Madnani, Abhinav Misra, Klaus Zechner, Zuowei Wang, and John Sabatini. 2019. Automated estimation of oral reading fluency during summer camp e-book reading with My Turn To Read. In *Proceedings of INTERSPEECH*, pages 21–25.

Nitin Madnani, Beata Beigman Klebanov, Anastassia Loukina, Binod Gyawali, Patrick L Lange, John Sabatini, and Michael Flor. 2019. My Turn to Read: An interleaved e-book reading tool for developing and struggling readers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 141–146.

Jack Mostow and Gregory Aist. 1999. Giving help and praise in a reading tutor with imperfect listening — because automated speech recognition means never being able to say you're certain. *CALICO Journal*, 16(3):407–424.

John Pikulski and David Chard. 2005. Fluency: Bridge between decoding and reading comprehension. *The Reading Teacher*, 58(6):510–519.

Sherry Ruan, Angelica Willis, Qianyao Xu, Glenn M Davis, Liwei Jiang, Emma Brunskill, and James A Landay. 2019. BookBuddy: Turning digital materials into interactive foreign language lessons through a voice chatbot. In *Proceedings of Learning@Scale*, pages 1–4.

John Sabatini, Zuowei Wang, and Tenaha O'Reilly. 2019. Relating reading comprehension to oral reading performance in the NAEP fourth-grade special study of oral reading. *Reading Research Quarterly*, 54(2):253–271.

Namhee Suk. 2017. The effects of extensive reading on reading comprehension, reading rate, and vocabulary acquisition. *Reading Research Quarterly*, 52(1):73–89.

Jade Wexler, Sharon Vaughn, Meaghan Edmonds, and Colleen Klein Reutebuch. 2008. A synthesis of fluency interventions for secondary struggling readers. *Reading and Writing*, 21:317–347.

Fei Wu, Leibny Paola García-Perera, Daniel Povey, and Sanjeev Khudanpur. 2019. Advances in automatic speech recognition for child speech using factored time delay neural network. In *Proceedings of INTERSPEECH*, pages 1–5.

A Annotation protocol

Use the following coding scheme to classify how each word was read:

Correct A word was read correctly, without difficulty.

Correct with difficulty A word was read correctly, even if initially read incorrectly, or with other signs of difficulty.

Incorrect A word was read incorrectly, even if initially read correctly.

A.1 Features of incorrect reading

A word should be coded as **incorrect** if it has any of the following qualities.

Mispronunciation Part of the word is not pronounced as it should be expected. In the case of proper nouns, any reasonable phonetic pronunciation of the word is acceptable. Mispronunciations can include: (a) Pronouncing the wrong segment, e.g., saying SUN as SOON, saying NATION as NATE-EE-ON; (b) Inserting an extra syllable, e.g., saying NATION as NA-SHE-ON; (c) Omitting part of a word such as a segment, syllable, or suffix, e.g., saying WISH instead of WISHES, saying DESCRIBED as DESCRIDE, saying AMBIGUOUSLY as AMBIGUSELY; (d) Reversing the order of segments, e.g., saying ONIMOUS instead of OMINOUS; (e) Using the wrong lexical stress pattern, e.g., saying JAPANESE as JA**PAN**ESE.

Replacement The reader says a different word instead of the target, e.g., IMPERIAL instead of EMPIRICAL, AUTOMOBILE instead of AUTOMATIC.

Not blending The reader sounds out the individual segments in a word instead of blending them together.

Intra-word pausing The reader pauses for an extended period of time mid-word, especially at a point that is not near an inflectional suffix or in a way that reduces intelligibility. e.g., ELE ... PHANT, TER ... MINATE.

Subvocalization The reader makes noises that resemble the word, such as by pronouncing a few segments while grunting or mumbling the rest.

A.2 Features of correct reading with difficulty

If a word has none of the features of incorrect reading, it should be coded as **correct with difficulty** if any of the following occur in or around the word.

Pausing The reader **unnaturally** pauses before or after the word at a point where the pausing is expected to be caused by difficulty with the word, such as: (a) Immediately before or after the word; (b) At a phrasal or clausal boundary before the word, in a manner where it does not seem that the difficulty is associated with another word.

Errors near the word The reader reads one or more word incorrectly before or after the word. This may be in an adjacent word or up to 4 words before or after the word if the errors do not seem to be caused by another difficult word nearby. This classification would occur, for example, if "immoderately" were targeted for analysis, and the reader omitted "so" but read "immoderately" correctly when reading "In short, she is so **immoderately** wise people call her wisdom personified...".

Repetition The reader says the word multiple times and says the final attempt correctly. The initial attempts may be either correct or incorrect.

False start The reader says part of the word, stops, and says the word again from the beginning correctly, e.g., saying CONCERT as CONCONCERT.

Intonation The reader uses rising intonation on the word, as if asking a question, in a manner that expresses uncertainty about correctness of the reading.

Mumbling The reader is mumbling through the part where the word occurs, perhaps subvocalizing several words, but reads the target word correctly.