A Dataset for Oral Reading in Young English Readers

Madison Rose, Michael Bennie, Valeria Pagliai
Hatice Kubra Karakis, Qian Shen, Xinyi Tai, Walter L. Leite
Zoey Liu
University of Florida
{rose.m, liu.ying}@ufl.edu

Abstract

Among English child speech corpora, very few focus on oral reading. Existing resources such as the CMU Kids Corpus (Ellis Weismer et al., 2013) face limitations in the lack of grade-appropriate, curriculum-aligned reading texts, the annotation scope and quality, and most crucially, comprehensive annotation scheme for characterization of children's reading errors. This study presents a multi-layered, fully manually annotated corpus of oral reading from 63 1st-3rd grade students residing in the U.S. who grow up hearing and speaking English. Additionally, we contribute methodologically rigorous annotation guidelines that categorize 10 reading error categories and 26 sublevel error labels. Using Storiza, a digital reading platform supported by GPT-4o-mini (OpenAI, 2024), children read stories on topics of their own interest, while the system records their speech and logs their interactions with embedded digital supports. Each recording is paired with detailed demographic and educational metadata and subjected to linguistic annotations, including: (1) sentence- and word-level time alignment; (2) orthographic and phonemic transcriptions; (3) reading errors.

1 Introduction

Child-read speech corpora can serve as critical resources for studying speech and language learning (Schwanenflugel et al., 2004; Godde et al., 2020), facilitating automated analysis of oral reading fluency (Bolaños et al., 2013) and building speech foundation models for automatic speech recognition (ASR) (Fan et al., 2024; Block Medin et al., 2024). In addition, prior research has shown that speech production error patterns in early child language can not only inform characterization of grammatical development (Smith, 1933) and provide insights into language acquisition theories (Locke, 1980), but also guide the advancement of spoken language technologies in early literacy education (Booth et al., 2020; Attia et al., 2025); children's oral reading errors in particular can be further leveraged to build learning analytic models and intelligent tutoring systems that in turn provide customized educational support and intervention for reading development (Paige, 2020; Mostow, 2016).

While there are a few child speech corpora in English, they face several limitations. First, except for the CMU Kids Corpus (Eskenazi et al., 1997), other corpora do not necessarily focus on children's oral reading of short stories or passages, but rather spontaneous speech (e.g., the MyST Child Corpus (Pradhan et al., 2024)). While the CSLU Kids Corpus (Shobaki et al., 2000) attends to child-read speech, it only contains data of children reading prompted stimuli including simple words and sentences, which are not aligned to any particular reading curriculum.

Second, existing datasets (see Table 1) lack manual time alignment at the utterance- and word-level, therefore the quality of the orthographic transcriptions is not always clear. Specifically, it is unknown whether the available audio transcripts of children's reading production were always manually conducted or verified (e.g., the CMU Kids Corpus). This also holds true for the available phonetic or phonemic transcriptions in the corpora.

Lastly, current English child speech corpora lack detailed and systematic annotations of oral reading errors that capture patterns such as disfluency or phonological errors that can signal delay in language development and specific learning disabilities, such as dyslexia. While prior work has attempted to characterize children's reading errors (Gothi et al., 2024; Smith et al., 2025), most of them focus on whether a produced word is correct or not, lacking information about specific error categories. Others attend to narrow-range disfluency errors at the word-level, such as word omission or insertion, with little to no detail at other linguistic levels, such as phonological and grammatical

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 588–600

Name	Participants	Size	Error Labels	Phonemic Trans.
CMU Kids Corpus (Eskenazi et al., 1997)	76 children; ages 6–11	5,180 utterances, ∼9 hours	Yes; Stutter, Repeat, Incorrect, Skip, Correct	Yes
MyST Child Corpus (Pradhan et al., 2024)	4th–5th grade	230K utterances, about 400 hours	No	No
CID Children's Speech Corpus (Lee et al., 1999)	436 children, ages 5–17, and 56 adults	24,152 audios	No	Yes
PF-STAR (Batliner et al., 2005)	611 children, ages 4–14	63 hours	No	Yes
CSLU Kids Corpus (OGI) (Shobaki et al., 2000)	1,100 children, Kindergarten to 10th grade	1,017 files, approx. 8–10 min per speaker	No, but they mention false start	Yes
NCTE Dataset (Kane et al., 2022)	2,128 children, 4th–5th grade	5,235 hours of recordings	No	No
Storiza Corpus	∼63	∼9hrs	Yes	Yes

Table 1: Overview of child (read) speech corpora vs. the Storiza Corpus in English early child language.

(Gopal, 2018; Huang and Staub, 2021).

As a step to address these limitations, we curated a dataset of oral reading collected from 63 English-speaking children reading age-appropriate stories. The dataset consists of 320 recordings (one recording per story), totaling ∼9 hours and 33,612 words. Each story was generated, along with illustrations, by Storiza, a reading application supported by GPT-4o-mini (OpenAI, 2024). The stories are aligned with a phonics curriculum (UFLI) used in all 50 states of the U.S. (as well as in many other English-speaking countries) (Lane and Contesse, 2022). As children interact with Storiza, their interaction logs, which contains mouse clicks, text highlights, and reading-progress traces, are also documented.

With the audio recordings, we perform **fully manual, multi-layered** linguistic annotations. The corpus is time-aligned at the sentence- and word-level, with orthographic and phonemic transcripts based on the International Phonetic Alphabet (IPA). In addition, we design comprehensive annotation guidelines that capture oral reading errors, including 10 error categories with 26 sublevel error labels. Hereafter, we refer to our corpus as the Storiza Corpus. Albeit with considerations for the sensitivity of child speech, we have received parent consent to release 239 recordings (74.69%), which will be deposited through the Linguistic Data Consortium. Lastly, we make the following resources publicly accessible to the research community, including texts of the original stories, time-alignments, manual transcriptions, error annotations of the full dataset, as well as our annotation scheme to motivate future corpus design.[1]

[1] https://osf.io/v9hw4/overview?view_only=1517f470096a418f8f1b018a2a511cfe

In the remainder of this paper, § 2 presents relevant child-read speech corpora; § 3 and § 4 illustrate our data collection process and annotation scheme; we conduct descriptive corpus analysis in § 5; and we conclude with outlines for future research directions enabled by our dataset in § 6.

2 Related Work

Child speech corpora remain scarce, prompting efforts to address this limitation through data augmentation and the use of adult speech data (Li et al., 2024; Booth et al., 2020). Furthermore, the existing corpora range from spontaneous speech to scripted texts that are not curriculum-aligned, making oral reading stories even more rare. Currently, only a handful of child speech corpora are available, and their objectives and focus do not always align. For example, the NCTE Dataset (Kane et al., 2022) and the MyST Child Corpus (Pradhan et al., 2024) both collected speech data from children of 4th and 5th grade and include considerable amounts of spoken material, but without phonemic transcriptions. In contrast, the CID Children's Speech Corpus (Lee et al., 1999) and PF-STAR (Batliner et al., 2005) consider phonemic transcriptions but do not report detailed labeling of speech production errors. See Table 1 for a brief summary of the corpora.

The CMU Kids Corpus (Eskenazi et al., 1997) and the CSLU Kids Corpus (Shobaki et al., 2000) provide different general categorizations of reading error production, yet with varying extents of details. By comparison, our annotation guidelines are more thorough. For example, both incorporate *False Start*, which is expanded to at least two different kinds of disfluency errors in our corpus (*Self-Correction* and *Stutter*). Similarly with regards to

the error *Word Substitutions* present in the CMU Kids Corpus, we additionally consider a detailed *Ortographic* error category with five sub-labels to differentiate this phenomenon. In general, while all of these corpora are innovative and robust, our focus is to expand existing categorizations of reading errors to allow a comprehensive understanding of the characteristics of child speech.

3 Data Collection and Preprocessing

We collected all recordings using Storiza, a web-based reading application aligned with a validated reading curriculum (UFLI) adopted in all 50 states of the U.S. (Lane and Contesse, 2022); the goal of this app is to help elementary school students improve their oral reading fluency and comprehension by allowing children to leverage their own interests and cultural backgrounds to create and illustrate stories. For reading practice, a child and their parent first select a targeted lesson of a particular grade level from the curriculum, and the child either speaks or types out any topic of their interest. The reading app then employs the GPT-4o mini (OpenAI, 2024) and FLUX1-dev (Black Forest Labs) models to generate a story and illustration, respectively, given the child's topic and the selected lesson. Once the story is created, the reading app shows a recording button next to each story.

We obtained Institutional Review Board authorization to recruit 120 parents and their 1st to 3rd grade children from a private Facebook Group mostly consisting of teachers and reading tutors familiar with the curriculum. In total, 103 parents with children residing in the U.S. signed the consent form. We asked parents to accompany their children to record one story per week for four weeks and not to help the child during the reading. After removing ineligibles and recordings of poor quality, we arrived at 320 story recordings (Kindergarten-level stories: 20; 1st grade: 122; 2nd grade: 132; 3rd grade: 46), provided by 63 children (1st grade: 26; 2nd grade: 24; 3rd grade: 13) with typical reading development, including 37 females, 25 males, and one that did not report gender. The corpus totals ∼9 hours; on average, each story is 1min29s, with the shortest being 7.03s and the longest 5min36s.

4 Annotation Guidelines and Process

The Storiza Corpus was annotated by eleven carefully selected undergraduate and graduate linguistics students at the University of Florida with prior training in phonetics who were compensated for their work. The annotation process spanned from June to December in 2025, consisting of two consecutive stages: the *sentence-level* followed by the *word-level*, using Label Studio as the annotation interface.[2] We developed and refined annotation guidelines throughout the annotation process; these detail operational instructions for the annotation interface, the annotation scheme, transcription conventions, and idiosyncratic case specifications.[3]

4.1 Sentence-level annotations

With each story recording, annotators first manually segmented it into individual utterances. As every recording is child-read speech, there is a gold-standard story text that the child read (but might end up producing something different). For each segmented utterance, annotators labeled the *intended sentence* according to the gold-standard text. In cases where applicable, annotators can deploy *one or multiple* of the following labels to describe certain deviation: (1) *Run-On*, when the current utterance is produced rapidly in succession with the previous utterance, with no inter-sentence pause nor intonational indicators of a new sentence's beginning; (2) *Repetition*, when the current sentence is a repetition of a previous utterance; (3) *Non-child speech*, e.g., speech produced by the parent(s).

4.2 Word-level annotations

Based on the sentence-level time-alignments, we automatically segmented each story recording into individual audio files with one utterance each. At the word-level stage, annotators manually time-aligned each utterance into its individual words, and provided orthographic and IPA transcription. In addition, we characterize whether a word can be considered as *Correct* production, along with 10 distinct error categories to cover different linguistic features (Table 2), including: *Disfluency*, *Phonological*, *Orthographic*, *Grammatical*, and *Visual tracking*, all of which we describe in detail in this section. The other five error categories are *Structural*, *Self-Response*, *Unintelligible*, *Whispering*, and *Other*, which we discuss in Appendix A. Certain error categories contain multiple sub-labels to

[2] https://labelstud.io/

[3] We provide more details for comparing specific error categories and sub-labels between our corpus along with the other two existing child read speech corpora, the CMU Kids Corpus and the CSLU Kids Corpus, in Appendix B.

Category	Sub-label	Description	Example
Correct	-	Word read correctly	*Target*: "dog" → *Child*: "dog"
Disfluency	Interjection	Inserts filler word	*Child*: " Um , the dog..."
	Word Repetition	Repeats entire word	*Child*: "golden golden retriever"
	Self-Correction	Self-corrects error	*Target*: *going* → *Child*: "go es... going"
	Parent Correction	Initially misreads a word then corrects it with parent's help	*Target*: "going" → *Child*: " goes (going (Parent Aid)) going"
	Parent Aid	Word produced by parents to help the child	
	Stutter	Repeats part of word	*Child*: " fo- forever"
	Prolongation	Extends sound duration	*Child*: "He rrrrr r mom"
	Broken Word	Pauses within word	*Child*: "s ... top" or "c ... a ... t"
Orthographic	Letter Reversal	Substitutes visually similar letter	*Target*: "ca p e" → *Child*: "ca b e" (p/d confusion)
	Left-Right Tracking	Switches letter order	*Target*: " s a w " → *Child*: " w a s "
	Phonological Sub.	Substitutes with a phonetically similar *real word*	*Target*: " hou se" → *Child*: " hor se"
	Contextual Sub.	Substitutes intended word with a contextually appropriate *real word*	*Target*: " small " → *Child*: " little "
	Unrelated Sub.	Substitutes intended word with a *real word* that bears no phonetic or contextual similarity	*Target*: " went " → *Child*: " wow "
Phonological	Consonant Sub.	Replace one consonant with another	*Target*: " th ink" / θ ɪŋk/ → *Child*: " f ink" / f ɪŋk/
	Vowel Sub.	Replace one vowel with another	*Target*: "b ea t" /b i t/ → *Child*: "b e t" /b ɛ t/
	Consonant Omission	Omits one consonant	*Target*: "b l ack" /b l æk/ → *Child*: "back" /bæk/
	Vowel Omission	Omits one vowel	*Target*: " a bout" / ə baʊt/ → *Child*: "bout" /baʊt/
	Consonant Insertion	Adds extra consonant	*Target*: "big" /bɪg/ → *Child*: "b l ig" /b l ɪg/
	Vowel Insertion	Adds extra vowel	*Target*: "trip" /tɹɪp/ → *Child*: "t e rip" /t ɛ rɪp/
	Misplaced Stress	Misplace stress on the intended word	*Target*: "contented" → *Child*: " CON tented"
Grammatical	Substitution	Substitutes with wrong grammatical morpheme/function word	*Target*: " those " → *Child*: " the "
Visual Tracking	Skip Line	Loses place and skips a phrase/clause	*Target*: "I saw a cat, a bat, and a mouse " → *Child*: "I saw a cat and a mouse "
	Backtrack	Re-reads previous phrase/clause	*Target*: " I saw a cat " → *Child*: " I saw a cat a bat I saw a cat "
	Wrong Order	Reads clauses out of order	*Target*: "I saw a big red truck" → *Child*: "I saw a red big truck"
Structural	Word Omission	Skips entire word	*Target*: " climbed the tree" → *Child*: "[...] the tree"
	Word Insertion	Adds word not in text	*Target*: "went home" → *Child*: "went then home"
Self-Response	-	Reacts to story content	Text: "Are you ready?" → *Child*: " Yes! Are you ready?"
Unintelligible	-	Unrecognizable speech due to mumbling or background noise	
Whispering	-	Whispers (under the breath) word/phrase	
Other	-	Other types of errors not captured by the categories above	

Table 2: Word-level oral reading annotation taxonomy for 10 error categories and 26 sub-labels; '-' in "Sub-label" means there are no further error labels contained in that particular error category (e.g., "Correct"). Cyan highlights indicate deleted/substituted content from target, green highlights indicate child's different/added production.

capture further specifications. For example, *Phonological Errors* are distinguished by whether they involve vowels or consonants and by the operation affecting them (e.g., deletion). A single word may receive more than one error category and sub-labels should they be applicable.

For a word to receive the *Correct* label, it should be devoid of any reading error specified in the annotation scheme. The word production was considered in the context of how it would naturally be produced in read speech (i.e., considering natural phonological phenomena common in spoken language). For example, annotators frequently noted elision, where /t/ or /d/ may be omitted when centered in a three-consonant cluster, such as in "stopped the" (/stɔːpt ðə/ –> /stɔːp ðə/) (Rattanasak, 2025), and this was annotated as *Correct*. Additionally, the word "and" had broad criteria, as it often saw an omitted final /d/ and/or vowel reduction due to linking with prior and/or following words.

If a word-level error is present, possible annotation fields include: *intended word* (the word expected to be produced), *produced word* (what was actually said), and *IPA transcription* (phonemic transcription). Although other child speech corpora have relied on different phonemic transcription means, such as Worldbet (Hieronymus, 1993) or TIMIT (Zue and Seneff, 1996), we chose the IPA largely due to our annotators' existing familiarity with this system and for its fine phonetic detail and contrastive sound distinctions in English.

4.3 Disfluency Errors

Disfluency is an all-encompassing category which covers all events that disrupt the flow of speech. Its sub-labels consist of: *Interjection*, *Word Repetition*, *Self-Correction*, *Parent Correction*, *Parent Aid*, *Stutter*, *Prolongation*, and *Broken Word*. Among these, *Interjection* perhaps is the most straightforward to determine, which involves the insertion of a word that does not contribute to the syntactic structure of the sentence, such as "oh" or "um."

The label *Word Repetition* is given to the repeated instance(s) of a word (e.g., if a child were to say "her her her" then the last two "her" are labeled as repetitions). For a word to be considered repeated, it must have been read correctly;

otherwise, the repeated attempt will count as *Self-Correction* (i.e. *repair* (Levelt, 1983; Alexander et al., 1997)), which occurs when the child misreads a word or clause and tries to correct themselves without prompt. If a child needs to read "kite" and produces: 1. /k.kæt/ 2. /kæt/ 3. /kit/, 1 and 2 will be labeled *Self-Correction* as they indicate the point where the child tries to repair their error. 3. does not receive the label given that the child does not attempt to repair it. For *Self-Corrected* phrases and clauses, the label is given on the last word at the point of restart, regardless of if the word is read correctly. If the child should read "the big red house" but instead reads: "the red house, the big red house," the first uttered "house" is the point at which the child attempts to repair their omission of "big"; therefore, it receives the *Self-Correction* label. All words read correctly in the restarted attempt receive the *Word Repetition* label (in this case, the second "the," "red," and "house").

Parent Correction operates in a similar manner to *Self-Correction*, but the correction is prompted by a parent/non-child speaker rather than the child. So, the child incorrectly produces a word (marked as *Parent Correction*), the parent provides the correct version (marked as *Parent Aid*), and the child re-attempts the word with the parent's assistance. If an utterance were to go as: 1. I (*Child*) 2. see (*Child*) 3. saw (*Parent*) 4. saw (*Child*), the child's production of "see" is marked as *Parent Correction*. The parent's production of "saw" is marked as *Parent Aid*, while the child's final utterance of "saw" is *Correct*.

For *Parent Aid*, a parent may also assist a child in indirect manners, e.g., saying "keep reading." It is also possible for *Parent Aid* to occur without *Parent Correction*, e.g., when a child misreads a word but successfully self-corrects before the parent offers their aid. This would look like: 1. see (*Child*) 2. saw (*Child*) 3. saw (*Parent*) 4. saw (*Child*). In this case, "see" is marked as *Self-Correction*, and *Parent Correction* is not used. The child's repetition of "saw" in 4. is marked as *Word Repetition*.

Several sub-labels, including *Stutter*, *Prolongation*, and *Broken Word*, require special characters for their notation in both the IPA transcription and the produced word. For *Stutter*, a word is segmented with its stuttered portion included, and this portion is marked with a period following it (".). Although the period typically signifies syllabic divisions in IPA, we did not distinguish syllabic boundaries in our transcriptions. For example, if a child repeats the /k/ in "cat", this is transcribed in IPA as /k.kæt/ and the produced word is "c.cat."

Prolongation occurs when a child extends a speech sound beyond its normal duration. This can be an indicator of uncertainty or done for emphatic effect. All prolongations are transcribed with a colon (":"), following the IPA convention of denoting prolonged phonemes. For example, if a child says "she" and drags out /i/, it is transcribed as /ʃi:/. Annotators were encouraged to take an intuitive approach to assigning this label. The "prolonged" sound in question should jump out as abnormally long, and its context must be considered. For instance, some children tended to speak more slowly overall, which may contribute to some sounds appearing prolonged at a surface level. Lastly, we use the *Broken Word* sub-label when the child abruptly pauses or breaks within a word. This also includes instances where the child sounds out a word phoneme by phoneme. The vertical bar ("|") is used to indicate at what point(s) the word is broken. If a child spells out the word "cat" phoneme by phoneme, it would be transcribed as /k|æ|t/ with the produced word annotated as "c|a|t".

4.4 Phonological Errors

The following category is that of *Phonological Errors*, whose sub-labels include: *Consonant/Vowel Substitution*, *Consonant/Vowel Insertion*, *Consonant/Vowel Omission*, and *Misplaced Stress*. For *Consonant/Vowel Substitution*, the child attempts to read the intended word yet replaces one or more consonants/vowels with another consonant/vowel. It is important to note that only consonants may be substituted with another consonant to be considered as *Consonant Substitution*, and vowels with vowels to be considered as *Vowel Substitution*.

In regards to *Consonant/Vowel Omission,* a consonant/vowel has been dropped. If a child is intended to read the word "black" (/blæk/) but produces an utterance more similar to "back" (/bæk/), we would mark this as *Consonant Omission*, the /l/ having been omitted. With C*onsonant/Vowel Insertion*, an additional consonant/vowel has been added to the target word.

Misplaced Stress occurs when the child has reassigned primary stress in a word, producing a word that is nonexistent in English (e.g., "CONtented" instead of "conTENTed"). It is important to note that for the purposes of *Phonological Errors*, diphthongs (e.g., /oʊ/, /aɪ/, etc) and affricates (e.g., /t͡ʃ/

and /d͡ʒ/) are considered as one vowel/consonant segment respectively, as operations often affect the entire unit rather than the individual segments inside of it.

4.5 Orthographic Errors

Orthographic Errors encapsulate errors resulting from full-word substitutions, mix-ups of visually similar letters, or mis-ordered letter sequences. This category's sub-labels are as follows: *Letter Reversal Substitution*, *Left-Right Tracking Substitution*, *Phonological Substitution*, *Contextual Substitution*, and *Unrelated Substitution*.

A word receives the *Letter Reversal Substitution* label when a child has substituted a letter in the intended word with a similar looking word, such as "bogs" instead of "dogs." For *Left-Right Tracking Substitution*, a word receives this label when the order of two letters in the intended word has been swapped. For example, "t" and "p" exchange positions as a child substitutes "spots" for "stops."

The following substitutions: *Phonological*, *Contextual*, and *Unrelated*, involve the full substitution of the intended word with another real English word. For *Phonological Orthographic Substitutions*, the child has substituted the intended word with a new word that either sounds similar (e.g., "beat" /bit/ and "bait" /eɪt/) or shares similar letters (e.g., "horse" and "house"). With *Contextual Orthographic Substitution*, the child replaces a real word that suits the context of the intended word. The child may employ synonyms such as "small" for "little" or can replace function words, e.g., substituting "by" for "with." Lastly, *Unrelated Orthographic Substitutions* occur when the replacement word has no evident phonetic or semantic similarity with the target (e.g., "with" for "cat").

4.6 Grammatical Errors

The *Grammatical Errors* category simply includes *Grammatical Substitutions.* In this instance, a child either substitutes a function word for another, e.g., "the" for "those," or uses the wrong grammatical inflection, e.g., "ran" for "run".

4.7 Visual Tracking Errors

This error category relates to all instances where a child has lost their place while reading. This category can offer insights into how children parse lines whilst reading, and while our dataset does not end up containing a significant amount of data in this category (Figure 1), future corpora may use the scheme proposed here to further investigate this category. Its sub-labels include: *Skip Line*, *Backtrack*, and *Wrong Order*.

Skip Line is used when a child skips an entire two-word-or-larger phrase or clause, e.g., "that I saw" in "The cat that I saw plays with a ball" is omitted. *Backtrack* occurs when the child rereads a clause or two-word-or-larger phrase from the sentence. Every word in the backtracked portion is marked with the *Backtrack* label. One example is: "The fluffy orange cat saw the dog the fluffy orange cat," where the repetition of the phrase "the fluffy orange cat" constitutes *Backtrack* having occurred. *Wrong Order* is used when words within a clause or two-word-or-larger phrase swap places, e.g., "big red house" becomes "red big house". This label cannot be induced by word shifts caused by *Word Omissions*, *Word Insertions*, or restarts prompted by *Self-Correction* attempts.

4.8 Annotator Agreement

Considering that we have 11 annotators in total, to identify and mitigate potential annotation bias, we randomly selected approximately 10% of annotated word-level tasks from each annotator as a balance for cross-annotation; this resulted in 361 cross-annotated tasks out of the 3,023 in total (∼12%). Among all the annotators, we chose the two most experienced to perform cross-annotations.

To calculate agreement score for manual time alignments of individual words, which are continuous values, we used temporal intersection over union (IoU). Given two alignments for the same word, (t_{i1}, t_{i2}) and (t_{w1}, t_{w2}), we computed their corresponding duration, t_i and t_w, then the alignment overlap, $t_{overlap}$. We measured the IoU score as follows in Eq (1). The overall agreement for manual time alignments is 0.91 (out of 1).

$$IoU = t_{overlap}/(t_i + t_w - t_{overlap}) \quad (1)$$

With regards to agreement scores for *intended words*, *produced words*, *IPA*, words labeled as *Correct*, error categories and sub-labels (on average), we used Krippendorff's alpha (α). Since a single word may contain different reading errors, for these cases, two annotators are only considered as in agreement if their annotations include the exact same combinations of error sub-labels. As shown in Table 3, the overall agreement scores are reasonably strong, especially considering the annotation task difficulty and that our annotation scheme is quite comprehensive. (Distribution statistics for

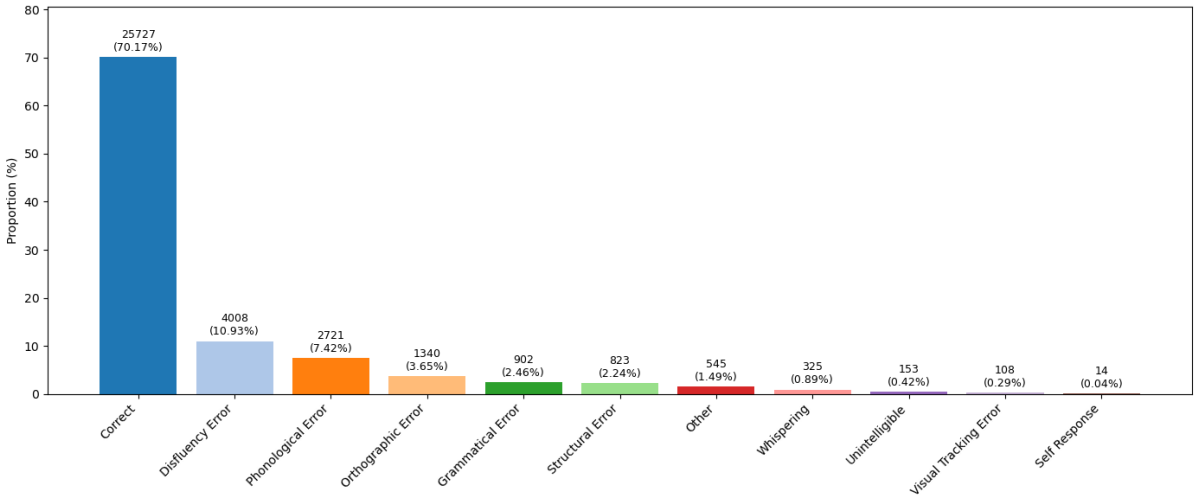

Figure 1: Distribution of Correct Word Production and 10 Error Categories. Note that the sum of the number of occurrence for each error category here exceeds the total number of words in our dataset, since a single word may receive multiple error categories and sub-labels.

cross-annotations and detailed agreement scores for each error category and sub-label are presented in Appendix D).

Intended word	Produced word	IPA	*Correct*	Error categories	Error sub-labels
0.97	0.84	0.87	0.95	0.90	0.87

Table 3: Overall agreement scores.

5 Descriptive Analysis

Overall, annotators segmented the 320 stories into 3,116 sentences, out of which 3,024 are readings of the intended sentences, 208 are *Run-On Productions*, 24 are *Sentence Repetition*, and 6 are *Non-Child Speech*. The average sentence length is 10 words, with duration averaging 9.48s, ranging from 0.6s to 2min3s.

Figure 1 illustrates the relative distribution of correct word production and the 10 error categories at the word-level. Based on our annotations, most of children's production in reading is considered *Correct* (71.29%). Among the error categories, *Disfluency* appears to be the most frequent (11.11%), followed by *Phonological* (7.54%), *Grammatical* (4.04%), and *Orthographic* (3.71%) errors. With regards to *Structural* errors, of all the produced words in our dataset (N=33,612), 0.67% (N =244) involve individual word insertion; on the other hand, 1.58% words (N=579) from the original texts of the stories were omitted. Across all actual grade levels of the children (not the grade level of the stories), this distribution generally holds. As grade level increases, the proportion of correct word production increases (from 67.37% in 1st grade to 72.80% and 75.01% in 2nd and 3rd grade). Among all error categories, *Disfluency* errors remain the most prevalent regardless of the grade levels (1st: 13.45%; 2nd: 9.55%; 3rd: 10.38%).

	General	1st	2nd	3rd
Disfluency				
Self-Correction	9.67%	11.10%	8.39%	9.67%
Prolongation	5.69%	7.37%	4.91%	3.90%
Broken Word	4.95%	7.27%	2.90%	4.93%
Phonological				
Consonant Omission	13.08%	11.68%	13.06%	16.31%
Consonant Substitution	9.83%	9.89%	9.83%	9.67%
Vowel Substitution	8.83%	9.42%	8.81%	7.52%

Table 4: Distribution of top three most frequent sub-labels in *Disfluency*, and *Phonological* errors.

Now we turn to examining the distributions of sub-labels within the two most frequent error categories, *Disfluency* and *Phonological* (see also Table 4 and full distributions of all error labels in Appendix C; note here to allow frequency comparisons between sub-labels from different error categories, the proportion for a given sub-label is measured relative to the total number of sub-labels in all error categories instead of the specific error category). For *Disfluency*, children seem to self-correct their production (10.06%) or prolong certain phonemes (5.93%) more often compared to other phenomena that reflects disfluency in speech; this trend is observed at all grade levels. A majority of the sub-label errors from this category see a decline in frequency as grade level progresses, with the exception of *Stutter*, *Word Repetition*, and

Interjection. *Stutter* saw no patterns of decline nor increase across each grade level. By contrast, *Word Repetition* and *Interjection* occur more frequently at higher grade levels, with 3rd graders seeing the highest for each sub-label. The increase of occurrence for these two sub-labels suggest employment of learned reading strategies when children encounter new or uncertain sounds. For example, *Interjections* provide children with a momentary pause to process the next word, formulate a strategy, and consequently attempt it. Similarly, when oral fluency is disrupted either by a within-text or external difficulty, *Word Repetitions* aids the child in relocating their place in the text.

As for *Phonological* errors, the frequencies of each sub-labels remain relatively stable across grades. For all grade levels, operations with consonants were always more frequent than their respective vowel counterparts, with consonant omissions and substitutions being the most frequent. This may be a result of typically having more consonants per syllable than vowels. For this error category, there does not appear to be any clear-cut trends of frequency decrease for any sub-label; however, we see a slight increase in *Misplaced Stress* and *Consonant Omission* along with the grade level. Among all categories, *Phonological* and *Disfluency* errors tend to co-occur most frequently, with a gradual increase in occurrence as grade level progresses. This perhaps is not surprising since some sub-labels that capture disfluency such as *Prolongation* and (sometimes) *Self-Correction* entail a phonological misproduction as well.

6 Future Research Directions

We foresee the Storiza Corpus enabling fruitful research in several directions. First, the availability of the audio data, coupled with multi-layered annotations such as produced words and reading errors, offer test cases for linguists and developmentalists interested characterizing the contexts in which different types of errors occur and how structural complexity at different levels possibly affect the occurrence of these errors. One might also be curious to explore linguistic differences between child-read speech in contrast to their spontaneous spoken speech, in order to probe, for instance, how speaking styles impact production errors. The manual time alignments in our corpus allow building forced aligners tailored for child-read speech that can benefit relevant work, which can alleviate annotation labor in the future and benefit these research endeavors (Mahr et al., 2021).

Others might engage in studying the phonological properties of child-read speech such as changes in vowel quality (Peterson, 1961), as well as how acoustic features are leveraged during oral reading (Schwanenflugel et al., 2004). Such insights may be particularly relevant to studies of first language acquisition, such as prosodic acquisition (Godde et al., 2020) or learning of consonant cluster productions (McLeod et al., 2001). Previous research has investigated how children (and adults) acquire novel phonotactic constraints through the review of speech errors, finding that children start picking up on them almost immediately (Smalle et al., 2017). This line of work, however, has remained largely experimental, e.g., using controlled learning paradigms with carefully curated stimuli, rather than relying on continuous speech in more natural settings. This makes our data which details child-read speech errors even more enlightening to facilitate this research direction.

Aside from linguistic analysis, our dataset can be readily employed for development of speech processing technologies that can be utilized in educational applications such as stealth assessment of oral fluency (Beigman Klebanov et al., 2023). One such technology is ASR. While there have been a number of relevant studies (Fan et al., 2024; Sukhadia and Chowdhury, 2024), they mostly use two child speech corpora (MyST (Pradhan et al., 2024) and CSLU (Shobaki et al., 2000)). These datasets consist of spontaneous speech or short isolated prompts, which differ from oral reading data involving continuous story narration. With the corpus presented here, one can fine-tune state-of-the-art ASR systems targeted towards child oral reading specifically. Data augmentation methods or transfer learning from parent speech (Li et al., 2024; Booth et al., 2020) can be also be adopted to improve model performance. Considering also that in cases of *Disfluency* errors, such as *Prolongation* and *Broken Word*, we include special characters for both the IPA transcription and the produced word, researchers can potentially design different training schemes so that the resulting ASR models learn to directly generate transcriptions with these characters at both the orthographic and phonemic levels. This can help with manual transcriptions along with identification of *Disfluency* errors.

The task of labeling speech errors, or miscue detection in reading has in fact received attention

in previous literature (Smith et al., 2025), though not to the same extent as ASR for child speech. Most studies used the CMU Kids Corpus, and due to the lack of thorough annotations for reading errors, they tend to focus on a very narrow range of speech errors that usually concern just omission, substitution, and insertion at the word level (e.g., Shankar et al. (2025)). Prior work has noted the challenge of this task (Black et al., 2007), taking into account the scarcity of relevant corpora and caveats with their annotation quality. For instance, with one private dataset of oral reading and the CMU Kids Corpus, Smith et al. (2025) found their error labeling models to be negatively impacted by automatic transcriptions present in these corpora. These limitations highlight the need for larger, high-quality child oral reading datasets with consistent annotation guidelines that allow more refined speech modeling and more accurate error prediction. We hope our dataset can contribute to mitigating these limitations, and that the research directions outlined here offer concrete suggestions to move forward in the future.

Limitations

At the expense of considerable manual annotation efforts and care involved in the development of our corpus, its resulting size may be deemed relatively restricted compared to some other child speech corpora (not necessarily pertaining to oral reading). Acknowledging this limitation, the resulting size of the dataset does not lag far behind the CMU Kids Corpus. We hope that future work can leverage the combination of these corpora for their experiments.

In subscribing to such fine-grained guidelines, errors naturally arise. Determining intention of a child's utterance can be quite challenging, and annotators may approach this challenge in different ways. For example, a child's substitution of an individual phoneme can produce an unintended real English word. If a child mispronounces the final /θ/ of "with" as /f/, this can correspond to either a non-English word /wɪf/, or the word "whiff". In these cases annotators try to take contexts into consideration, and some may deem unlikely that the child intends to substitute the word "whiff" for "with" and that this constitutes a *Phonological* error. Other times where contexts may not be helpful for disambiguation, in that the child's intention is never completely transparent, then annotators' assumptions can be inaccurate.

A final consideration lies in the quality of the stories generated by GPT-4o-mini. Although stories are generated with prompts that explicitly consider a child's grade level, target structures, and unique interests, some stories could be generated above the child's grade level or contain unusual content. Future efforts may attempt to further refine the story generation process and ensure that stories align more closely with the child's needs and with educational standards.

References

Dianne Alexander, Amy Wetherby, and Barry Prizant. 1997. The emergence of repair strategies in infants and toddlers. In *Seminars in speech and language*, volume 18, pages 197–212. © 1997 by Thieme Medical Publishers, Inc.

Ahmed Adel Attia, Dorottya Demszky, Tolúlpe Ògúnremí, Jing Liu, and Carol Espy-Wilson. 2025. Cpt-boosted wav2vec2.0: Towards noise robust speech recognition for classroom environments. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong. 2005. The PF_STAR children's speech corpus. In *Proc. Interspeech 2005*, pages 2761–2764.

Beata Beigman Klebanov, Michael Suhan, Zuowei Wang, and Tenaha O'reilly. 2023. A dynamic model of lexical experience for tracking of oral reading fluency. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 567–575, Toronto, Canada. Association for Computational Linguistics.

Matthew Black, Joseph Tepperman, Sungbok Lee, Patti Price, and Shrikanth Narayanan. 2007. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In *Interspeech 2007*, pages 206–209.

Black Forest Labs. FLUX.1 [dev]. Hugging Face model card (open weights). Retrieved 2026-02-18.

Lucas Block Medin, Thomas Pellegrini, and Lucile Gelin. 2024. Self-Supervised Models for Phoneme Recognition: Applications in Children's Speech for Reading Learning. In *Interspeech 2024*, pages 5168–5172.

Daniel Bolaños, Ron Cole, Wayne Ward, Gerald Tindal, Jan Hasbrouck, and Paula Schwanenflugel. 2013. Human and automated assessment of oral reading fluency. *Journal of educational psychology*, 105(4):1142.

Eric Booth, Jake Carns, Casey Kennington, and Nader Rafla. 2020. Evaluating and improving child-directed automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6340–6345, Marseille, France. European Language Resources Association.

Jerome D'Agostino, Robert Kelly, and Emily Rodgers. 2019. Self-corrections and the reading progress of struggling beginning readers. *Reading Psychology*, 40(6):525–550.

Susan Ellis Weismer, Courtney Venker, Julia Evans, and Maura Moyle. 2013. Fast mapping in late-talking toddlers. *Applied Psycholinguistics*, 34:69–89.

Maxine Eskenazi, Jack Mostow, and David Graff. 1997. The CMU Kids Corpus (LDC97S63).

Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models. In *Interspeech 2024*, pages 5173–5177.

Erika Godde, Marie-Line Bosse, and Gérard Bailly. 2020. A review of reading prosody acquisition and development. *Reading and Writing*, 33(2):399–426.

Revathi Gopal. 2018. Miscue analysis: A glimpse into the reading process. *Studies in English Language and Education*, 5(1):12–24.

Raj Gothi, Rahul Kumar, Mildred Pereira, Nagesh Nayak, and Preeti Rao. 2024. A dataset and two-pass system for reading miscue detection. In *Proc. Interspeech*, volume 2024, pages 4014–4018.

James Hieronymus. 1993. ASCII phonetic symbols for the world's languages: Worldbet. *Journal of the International Phonetic Association*, 23(3):12–54.

Kuan-Jung Huang and Adrian Staub. 2021. Why do readers fail to notice word transpositions, omissions, and repetitions? a review of recent evidence and theory. *Language and Linguistics Compass*, 15(7):e12434.

Thomas Kane, Heather Hill, and Douglas Staiger. 2022. National center for teacher effectiveness main study. ICPSR36095-v4.

Terri Lander and ST Metzler. 1997. The CSLU labeling guide. *Center for Spoken Language Understanding, Oregon Graduate Institute*.

Holly Lane and Victoria Contesse. 2022. UFLI foundations: An explicit and systematic phonics program. Ventris Learning.

Lisa LaSalle and Edward Conture. 1995. Disfluency clusters of children who stutter: Relation of stutterings to self-repairs. *Journal of Speech, Language, and Hearing Research*, 38(5):965–977.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Willem Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.

Jialu Li, Mark Hasegawa-Johnson, and Nancy L. McElwain. 2024. Analysis of self-supervised speech models on children's speech and infant vocalizations. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 550–554.

John Locke. 1980. The prediction of child speech errors: Implications for a theory of acquisition. In *Child phonology*, pages 193–209. Elsevier.

Tristan Mahr, Visar Berisha, Kan Kawabata, Julie Liss, and Katherine Hustad. 2021. Performance of forced-alignment algorithms on children's speech. *Journal of Speech, Language, and Hearing Research*, 64(6S):2213–2222.

Sharynne McLeod, Jan Van Doorn, and Vicki Reed. 2001. Normal acquisition of consonant clusters. *American Journal of Speech-Language Pathology*, 10(2):99–110.

Jack Mostow. 2016. Project listen's reading tutor. In Scott A. Crossley and Danielle S. McNamara, editors, *Adaptive Educational Technologies for Literacy Instruction*, pages 263–267. Routledge, New York, NY. Taylor & Francis.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

David Paige. 2020. Reading fluency: A brief history, the importance of supporting processes, and the role of assessment. *Online Submission*.

Gordon E Peterson. 1961. Parameters of vowel quality. *Journal of Speech and Hearing Research*, 4(1):10–29.

Suman Pradhan, Ronald Cole, and Wayne Ward. 2024. My science tutor (MyST)–a large corpus of children's conversational speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12040–12045, Torino, Italia. ELRA and ICCL.

Sonthaya Rattanasak. 2025. Phonological processes in english connected speech: implications for L2 speech learning and communication. *Cogent Education*, 12:1–16.

Paula J Schwanenflugel, Anne Marie Hamilton, Melanie R Kuhn, Joseph M Wisenbaker, and Steven A Stahl. 2004. Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of educational psychology*, 96(1):119.

Natarajan Balaji Shankar, Zilai Wang, Kaiyuan Zhang, Mohan Shi, and Abeer Alwan. 2025. CHSER: A dataset and case study on generative speech error correction for child ASR. *Preprint*, arXiv:2505.18463.

Khaldoun Shobaki, John-Paul Hosom, and Ronald A. Cole. 2000. The OGI kids' speech corpus and recognizers. In *Proc. 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages vol. 4, 258–261.

Eleonore HM Smalle, Merel Muylle, Arnaud Szmalec, and Wouter Duyck. 2017. The different time course of phonotactic constraint learning in children and adults: Evidence from speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(11):1821.

Griffin Dietz Smith, Dianna Yee, Jennifer King Chen, and Leah Findlater. 2025. Prompting Whisper for improved verbatim transcription and end-to-end miscue detection. In *Proc. Interspeech 2025*, pages 1943–1947.

Madorah Smith. 1933. Grammatical errors in the speech of pre-school children. *Child Development*, 4(2):183–190.

Vrunda N. Sukhadia and Shammur Absar Chowdhury. 2024. Children's Speech Recognition through Discrete Token Enhancement. In *Interspeech 2024*, pages 5143–5147.

Victor Zue and Stephanie Seneff. 1996. Transcription and alignment of the TIMIT database. In *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*, pages 515–525. Elsevier.

A All Other Error Categories

Here we characterize the remaining five error categories in our annotation scheme: *Structural*, *Self-Response*, *Unintelligible*, *Whispering*, and *Other*.

Structural Errors include *Word Insertion*s and *Word Omissions*; however, only *Word Insertions* are directly assigned to a word segment. With *Word Insertions*, a child inserts one or more words that were not intended to be read as part of the story, e.g., "I saw the big red house" when the target was "I saw the big house." The annotators provided the produced word and its corresponding IPA transcription; although, they do not provide the intended word. With *Word Omissions*, a child fails to read one or more words from the story. As one cannot directly "segment" the time period when a word was omitted, annotators were directed to leave a comment detailing the word omission in a trigram format. The trigram format provides the context of the omitted word, lending clarity as to which instance of a word is deleted (useful for commonly repeated function words). For example, if the child omits "big" in the sentence "The cat plays with a big red ball," the annotator is directed to leave a comment as follows: *Word omission: "big" in "with a big."*

The *Unintelligible* category corresponds to segments where the intended word or utterance is unrecognizable due to reduced articulation, mumbling, any type of distortion, etc. Annotators could attempt to transcribe what was audible, though were not explicitly expected to do so. By comparison, the *Whisper* error category is assigned to words whispered by the child, though their production is intelligible, i.e., annotators can hear what the child was actually saying.

Self-Response is a category given to words that a child utters as a genuine reaction to the content of the text. If a child reads a sentence and adds, "Wow!" at the end, this word is labeled as a *Self-Response*. This also encapsulates when children answer rhetorical questions within the sentence.

The *Other* category is a catch-all for all remaining cases. It is used for any non-child speech that is not directly attempting to aid the child, for cut-off audio at the beginning or end of a recording, and for any idiosyncrasies (such as overlapping speech). For all idiosyncrasies, annotators were explicitly requested to leave a comment detailing the instance and the segment to which it applied.

B Comparison to Existing Corpora

The CMU Kids Corpus (Eskenazi et al., 1997), the CSLU Kids Corpus (Shobaki et al., 2000) and the Storiza Corpus have all supplied innovative, robust approaches to annotating children's speech. The CSLU Kids Speech is strong in its careful details of auditory data, including thorough descriptions of non-speech and speech phenomena alike; although, a direct comparison is unfair given their annotation scheme was not developed with reading error identification in mind. The CMU Kids Corpus was developed with reading analysis in mind, and supplies numerous helpful labels to accomplish this task. These two corpora also provide fine-grained instructions for detailing non-speech sounds such as coughs, humming, sneezing, and clicking.

Our corpus does not provide such speech details; however, our strength primarily lies in our comprehensive coverage of reading errors. We expand on existing labels from the CMU Kids Corpus and the CSLU Kids Corpus, pinpointing specific reading

phenomena, such as *Stutter* (LaSalle and Conture, 1995) or *Self-Correction* (D'Agostino et al., 2019). Further, the Storiza Corpus was designed with ease of data accessibility and readability in mind. As such, errors are reported on in precise, thorough manners that are equally easy to view and interpret by readers. In this section, we detail more specifics of individual error categories and sub-labels in our dataset in contrast to those from the aforementioned two corpora.

Our corpus is unique in furnishing interaction data between children and non-child speakers with the sub-labels *Parent Aid* and *Parent Correction*, to better understand the impact of non-child speakers' assistance attempts on the child's oral fluency. The other corpora may not supply this data simply because of their relatively strict (positively speaking) recording conditions that prevented non-child speakers from assisting. Although the CSLU Labeling Guidelines (Lander and Metzler, 1997) provide labels for annotating all background speech, we expand on this by directly recording its interaction with the speaker. From this, one may explore at which point parents provide feedback, the types of corrective feedback provided, and the efficacy of each type (and how the child responds to each).

While *Broken Word* and *Prolongation* may be indirectly noted with other sub-labels such as phone insertion, deletion, or noise in the other two corpora, our corpus's use of these specific sub-labels aids in separating the nuances of each phone insertion or the origin of within-word pauses/non-speech noise. *Broken Word* is particularly useful in that it allows us to identify instances where a child attempts to "sound out" a word before speaking it, which can provide insights into how and when children attempt new or unfamiliar words.

Our corpus is not the first to notate *Interjections*. The CSLU Kids Speech Corpus thoroughly addresses *Interjections* and any sort of filler words not just in English, but across multiple languages. The CMU Kids Corpus, on the other hand, does not explicitly label *Interjections*. In addition, our annotation scheme does uniquely notate *Self-Response*, and while there was not much data provided for this category in our dataset, it addresses the child's potential interaction with the text and may indicate their motivation and interest in the text's content.

Both our corpus and the CMU Kids Corpus detail *Word Repetitions*. All three corpora provide the *Whisper* label, and the CSLU Kids Speech and our corpus provide explicit *Unintelligible* notations. Both the CSLU Kids Speech and our corpus note *Grammatical Errors*; however, we do so in different manners. The former provides *Grammatical Errors* in the lens of language transfer causing this error, while the latter supplies this label as a result of reading errors and first language development.

As for *Word Insertions* and *Word Omissions*, both our corpus and the CMU Kids Corpus recognize and supply data for these phenomena. Our corpus expands on the aforementioned labels by providing directly the specific word that received these categories for ease of access and readability. However, the CMU Kids Corpus uniquely supplies information on whether or not the *Word Insertion* (or *Substitution*) includes a word contained within their pronouncing dictionary, which can be helpful in identifying unseen words and explaining any transcription gaps.

Further, with the *Word Substitutions* label present in the CMU Kids Corpus, our dataset comments on the specific nature of the substitutions with our *Orthographic Category Errors*. We detail whether a substitution has phonemic, orthographic, or semantic relevance to its target. For partial substitutions, our labels designate whether the child has misordered letters or has visually confused them with another letter. This could be particularly helpful in future efforts to study automatic detection of reading disfluencies such as dyslexia.

While the CMU Kids Corpus details *Phonological Category Errors* as we do, we expand on these labels by including *Consonant* and *Vowel* labels. We also supply annotations for when multiple *Phonological Errors* have occurred on one word, which is provided but not explicitly stated in the other corpus. When this occurs, annotators have provided detailed documentation of each *Phonological Error*, its *Consonant/Vowel* status, and the affected phones in IPA.

C Descriptive Statistics for Error Sub-labels

Here we present the full distributions of error sub-labels for different error categories where applicable (Table 5).

D Descriptive Statistics for Cross-annotations and Detailed Agreement Scores

This section presents descriptive distributions of correct words, error categories (Table 6) and indi-

	General	1st	2nd	3rd
Disfluency				
Self-Correction	9.67%	11.10%	8.39%	9.67%
Prolongation	5.69%	7.37%	4.91%	3.90%
Broken Word	4.95%	7.27%	2.90%	4.93%
Word Repetition	4.71%	3.13%	5.34%	6.64%
Parental Aid	3.35%	5.60%	1.75%	2.35%
Stutter	2.66%	2.78%	2.54%	2.67%
Parent Correction	1.04%	1.65%	0.51%	1.03%
Interjection	0.98%	0.32%	0.91%	2.67%
Phonological				
Consonant Omission	13.08%	11.68%	13.06%	16.31%
Consonant Substitution	9.83%	9.89%	9.83%	9.67%
Vowel Substitution	8.83%	9.42%	8.81%	7.52%
Vowel Omission	5.20%	4.69%	5.37%	5.89%
Consonant Insertion	4.97%	4.55%	5.54%	4.46%
Vowel Insertion	2.73%	2.70%	3.08%	1.91%
Misplaced Stress	0.09%	0.04%	0.09%	0.20%
Orthographic				
Phonological Substitution	5.16%	4.83%	6.01%	3.74%
Contextual Substitution	1.85%	1.54%	2.31%	1.35%
Unrelated Substitution	1.85%	1.86%	2.00%	1.43%
Left Right Left Right Tracking Substitution	0.61%	0.51%	0.76%	0.44%
Letter Reversal Substitution	0.25%	0.35%	0.15%	0.28%
Grammatical				
Grammatical Substitution	6.18%	4.82%	6.38%	8.79%
Structural				
Word Insertion	1.66%	0.89%	2.17%	2.07%
Word Omission	3.94%	2.42%	6.21%	1.59%
Visual Tracking				
Wrong Order	0.38%	0.30%	0.53%	0.20%
Skip Line	0.21%	0.16%	0.33%	0.04%
Backtrack	0.14%	0.12%	0.11%	0.28%

Table 5: Distribution of sub-labels for different error categories.

vidual error sub-labels (Table 7). Additionally, we provide average agreement scores for specific error category and sub-labels. Note that we removed error categories and sub-labels that appeared for less than 1% in our cross-annotations.

	Distribution proportion	Krippendorff's α
Correct	72.47%	0.95
Disfluency Error	12.17%	0.90
Phonological Error	9.20%	0.84
Orthographic Error	3.78%	0.87
Grammatical Error	2.39%	0.86

Table 6: Distribution of Correct words and error categories, and their agreement scores; error categories occurring for less than 1% in the cross-annotations were removed from computation.

	Distribution proportion	Krippendorff's α
Stutter	2.62%	0.85
Word Repetition	6.39%	0.87
Prolongation	8.36%	0.85
Broken Word	8.29%	0.86
Self-Correction	13.89%	0.90
Consonant Substitution	10.30%	0.77
Vowel Substitution	7.90%	0.76
Consonant Omission	13.10%	0.74
Vowel Omission	4.74%	0.69
Consonant Insertion	4.63%	0.74
Vowel Insertion	3.77%	0.79
Phonological Substitution	6.35%	0.78
Unrelated Substitution	2.37%	0.79
Substitution	7.29%	0.86

Table 7: Distribution of error sub-labels, and their agreement scores; error sub-labels occurring for less than 1% in the cross-annotations were removed from computation.