The LetsRead Corpus of Portuguese Children Reading Aloud for Performance Evaluation

Jorge Proença[1,2], Dirce Celorico[1], Sara Candeias[3], Carla Lopes[1,4], Fernando Perdigão[1,2]
[1] Instituto de Telecomunicações, Coimbra, Portugal
[2] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[3] Microsoft Language Development Centre, Lisbon, Portugal
[4] Polytechninc Institute of Leiria, Leiria, Portugal
IT, Department of Electrical and Computer Engineering, University of Coimbra - Pole II, 3030-290, Coimbra, Portugal
E-mail: {jproenca,dircelorico}@co.it.pt, t-sacand@microsoft.com, {calopes ,fp}@co.it.pt

Abstract

This paper introduces the LetsRead Corpus of European Portuguese read speech from 6 to 10 years old children. The motivation for the creation of this corpus stems from the inexistence of databases with recordings of reading tasks of Portuguese children with different performance levels and including all the common reading aloud disfluencies. It is also essential to develop techniques to fulfill the main objective of the LetsRead project: to automatically evaluate the reading performance of children through the analysis of reading tasks. The collected data amounts to 20 hours of speech from 284 children from private and public Portuguese schools, with each child carrying out two tasks: reading sentences and reading a list of pseudowords, both with varying levels of difficulty throughout the school grades. In this paper, the design of the reading tasks presented to children is described, as well as the collection procedure. Manually annotated data is analyzed according to disfluencies and reading performance. The considered word difficulty parameter is also confirmed to be suitable for the pseudoword reading tasks.

Keywords: Children's speech, Reading disfluencies, European Portuguese

1. Introduction

To evaluate the reading aloud performance of children, human assessment is usually involved, where a teacher or tutor has to take time to individually determine the performance in terms of fluency (speed, accuracy and expression). The automatic estimation of reading ability can be an important alternative or complement to the usual methods, and can improve other applications such as e-learning. However, there are no computer assisted applications for European Portuguese (EP) that automatically evaluate the reading aloud performance of children. Techniques must be developed to analyze audio recordings of read utterances by children and detect the deviations to the intended correct reading (disfluencies). Even for other languages, this automatic evaluation is a developing topic, and the focus is on reading of isolated words instead of larger sentences (Black et al., 2011; Duchateau et al., 2007).

To carry out the goals of the LetsRead project[1], we found it necessary to create a large new speech corpus of EP children's speech with utterances of reading tasks that are rich in the common disfluencies that children commit while reading. There are some children's speech databases for EP, such as Speecon with rich sentences (Speecon Consortium, 2005); ChildCAST (Lopes, 2012; Lopes et al., 2012) with picture naming; the Contents for Next Generation (CNG) Corpus targeting interactive games (Hämäläinen et al., 2013) and (Santos, 2014; Santos et al., 2014) with child-adult interaction. However, these databases do not present the required samples of disfluent reading speech. Since children's speech has different characteristics from adult speech (such as fundamental frequency, formant frequency variability, vowel duration variability, etc.) (Hämäläinen et al., 2014a; Lee et al., 1999), special care is needed to adapt or create robust acoustic models that target children (Hämäläinen et al., 2014b; Potamianos and Narayanan, 2003). The overall goal of the LetsRead project is to automatically evaluate the reading aloud performance of Portuguese children from around 6 to 10 years old (1st-4th grades) through the analysis of a small number of reading tasks.

This paper introduces the LetsRead corpus of EP children reading aloud. Section 2 describes the preparation of the reading tasks presented to children, entailing careful selection and distribution of material according to difficulty. Section 3 relates the data collection procedure and section 4 presents the corpus along with analyses of disfluencies and reading performance from annotated data.

2. Design of Reading Tasks

The Portuguese government has defined certain Curricular Goals (CG) with qualitative and quantitative objectives per grade for reading aloud (Buescu et al., 2015). Some of these objectives include target reading speed of words per minute on different tasks. With the analysis of curricular goals in mind, the reading of sentences and pseudowords was the target material to be collected. The pseudoword reading task provides an objective analysis of reading skills. However, sentences are the main focus of this database, where plenty of reading disfluencies can be collected to measure the overall reading performance of a child. Each child will be presented with a reading task that consists of reading aloud twenty sentences and ten pseudowords. Forty reading tasks were established (10 per grade) to balance repetition and diversity of the data. At a later stage, these

[1] The LetsRead project: http://lsi.co.it.pt/spl/letsread.htm

were shortened to 5 tasks per grade, to reinforce repetition of the same samples.

2.1 Sentences: Selection and Difficulty Level

A large set of sentences was extracted from child's tales and school books of the level of the target group (6-10 years old, 1st-4th grades). Twenty sentences were included in each reading task (for a recording session of one child). The first concern for distributing sentences along the grades was to maintain a good representation of all phones, so that acoustic models with significant quality can be built with the data. The other main concerns were to maintain the same average difficulty within a grade (with a rising average difficulty from 1st to 4th grades) and to have sentences of varying difficulty in a task (overlapping distributions of difficulty for the grades). A parameter of difficulty was developed to classify sentences according to phonological constraints. Although it would be ideal to also relate a word's difficulty to its age-of-acquisition or familiarity, not all words of the proposed reading tasks were present in available lexical databases such as ESCOLEX (Soares et al., 2014), and it was not possible to consider such features. The proposed parameter of difficulty is based on the method described in (Mendonça et al., 2014) where sentences are evaluated in terms of phonetic complexity and variety. All words were split into syllables and to each syllable a difficulty level was assigned, determined from these rules: the length of the syllable; the dubious pronunciation of some graphemes (e.g. <mãe> [mˈɐ̃j̃][2] and <bem> [bˈɐ̃j̃]); the presence of consonant clusters (e.g. <prever> [pɾəvˈeɾ] or <florescer> [fluɾəʃsˈeɾ]) and vocalic encounters (<candeeiro> [kɐ̃diˈɐjɾu] or <veem> [vˈeɐ̃j̃]). Since each syllable has a given minimum difficulty, the length of the sentence also contributes to difficulty.

2.2 Pseudowords Creation

Pseudowords represent non-existing or non-sense words which can be used to evaluate morphological and phonemic awareness. A novel method for the creation of pseudowords was developed. Existing tools such as Wuggy (Keuleers and Brysbaert, 2010) take as input existing words and output pseudowords that differ in one or two syllables to the original words. This creates pronounceable words that are similar to existing words (such as <sapado> from <sapato>). The proposed method creates pseudowords without the starting point of valid words, though still maintaining full pronounceability. It should create unfamiliar words and the difficult of reading them should be slightly higher. The aim was to create pseudowords of two, three and four syllables. First, the most frequent syllables in each position for words with those number of syllables were extracted from a large lexicon of European Portuguese, CETEMPúblico (Rocha and Santos, 2000; Santos and Rocha, 2001). Then, words of the three lengths are created randomly from a set of the most frequent syllables. Words that have syllabic encounters that do not respect pronounceability rules are deleted, as are words that exist in the lexicon. The difficulty score for a pseudoword is calculated by the same method described above for sentences. The distribution of the pseudowords along the reading tasks also promotes varying difficulty and a rising average difficulty along the grades.

3. Data Collection

The corpus of children reading aloud was collected at 2 private and 9 public schools in urban centres and periphery areas of the central Portugal region, with children that attend primary school, aged 6 to 10 years old. A specific application was developed in which the sentences are displayed in a large font size on a computer screen simultaneously with the start of recording. The recording environment and a screenshot of the application can be seen in Figure 1 and Figure 2. Children are asked to read aloud a set of 20 sentences and 10 individual pseudowords. A lapel Lavalier microphone (Shure WL93) was used as the main recording device, accompanied by a standard table top PC microphone as backup (Plantronics Audio 10). The recordings were performed in chosen school classrooms that exhibited low reverberation and noise.

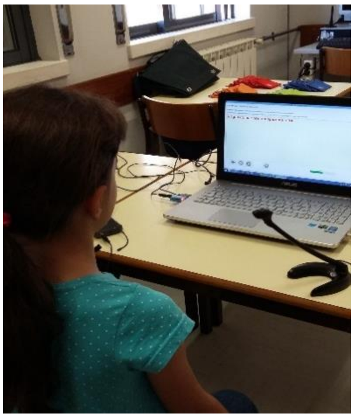
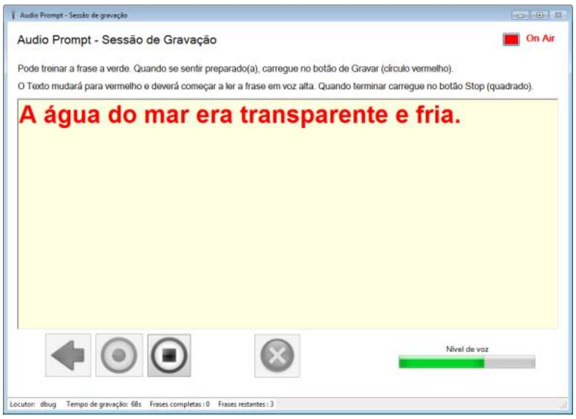


Figure 1: Example of the recording environment.



Audio Prompt - Sessão de gravação

Audio Prompt - Sessão de Gravação

Pode treinar a frase a verde. Quando se sentir preparado(a), carregue no botão de Gravar (círculo vermelho).

O Texto mudará para vermelho e deverá começar a ler a frase em voz alta. Quando terminar carregue no botão Stop (quadrado).

A água do mar era transparente e fria.

Locutor: dbug Tempo de gravação: 68s Frases completas : 0 Frases restantes : 3

Figure 2: Example of the recording software

[2] Phonetic symbols in this document are in the notation of the International Phonetic Alphabet (IPA), https://www.internationalphoneticassociation.org/content/full-ipa-chart

4. Corpus

The collected database consists of around 20 hours of recorded speech from 284 children, 147 female and 137 male, distributed from the 1st to the 4th grade with 68, 88, 76 and 52 children, respectively. A set of 104 children's speech utterances has been fully and manually annotated considering several disfluency types. These 104 children (46 male and 58 female) are equally distributed along the 4 grades (26 per grade) and their utterances amount to approximately 5 hours and 30 minutes of speech. For the remaining data, an automatic alignment was performed according to the method described in (Proença et al., 2015), allowing repetitions and false starts.

4.1 Disfluencies

The annotated speech exhibits a large variety of disfluencies that represent the most common types of errors in reading aloud of children. Based on previous work (Candeias et al., 2013), the rules for the annotation and labeling procedure were defined and several types of disfluency were identified:

- PRE – False starts that are followed by the attempted correction (pre-corrections).
- SUB – Substitution or severe mispronunciation of a word.
- PHO – Small mispronunciation of a word, usually with a change in one phone.
- REP – Repetition of a word.
- INS – An inserted word that is not part of the original sentence.
- DEL – The word was not pronounced (deleted).
- CUT – The word is cut, usually in the initial or final syllable, but not corrected later.
- EXT (:) – Phone Extension.
- PAU (…) – Intra-word pause, when a word is pronounced syllable by syllable and silence occurs in between.

Silence and noise events such as breathing, labial and background noise were also annotated. Extensions and intra-word pauses may occur simultaneously with others disfluencies and are marked with [:] and […] inside phonetic transcriptions. The number of occurrences for each type of disfluency for sentence and pseudoword reading tasks and their percentage of total uttered words in the database are presented in Table 1.

Some interesting phenomena can be observed, such as the defined false start disfluency type being the most common for sentences, whereas mispronunciations are more common for pseudowords. This occurs as there are less attempts to correct unknown words. Table 2 shows the disfluency statistics for sentence reading tasks for the four grades.

It can be observed that 1st grade children are the ones that exhibit more intra word pauses and extensions (due to slower reading) and 4th grade children have more insertions and deletions (due to faster reading). Overall, the percentage of words in the data that present any type of disfluency is approximately 13%.

Tags	Sentences	Pseudowords
	Total	**Total**
PRE	1156 (5.13%)	173 (16.63%)
SUB	754 (3.34%)	156 (15.00%)
PHO	729 (3.23%)	233 (22.40%)
REP	501 (2.22%)	1 (0.09%)
INS	179 (0.79%)	17 (1.63%)
DEL	85 (0.38%)	3 (0.29%)
CUT	82 (0.36%)	2 (0.19%)
...	686 (3.04%)	236 (22.69%)
:	472 (2.09%)	136 (13.07%)

Table 1: Disfluencies in sentences and pseudowords (number of events and % of total uttered words).

Tags	Sentences			
	1st grade	**2nd grade**	**3rd grade**	**4th grade**
PRE	295 (7.44%)	278 (5.72%)	281 (4.41%)	302 (4.10%)
SUB	182 (4.59%)	149 (3.07%)	215 (3.38%)	208 (2.83%)
PHO	214 (5.40%)	169 (3.48%)	203 (3.19%)	143 (1.94%)
REP	122 (3.08%)	89 (1.83%)	129 (2.03%)	161 (2.19%)
INS	30 (0.76%)	42 (0.86%)	42 (0.66%)	65 (0.88%)
DEL	5 (0.13%)	14 (0.29%)	16 (0.25%)	50 (0.68%)
CUT	11 (0.28%)	15 (0.31%)	29 (0.46%)	27 (0.37%)
...	256 (6.46%)	145 (2.98%)	212 (3.33%)	73 (0.99%)
:	179 (4.52%)	126 (2.59%)	102 (1.60%)	65 (0.88%)

Table 2: Disfluencies in sentences for the four grades (number of events and % of total uttered words).

4.2 Reading Performance

With annotated data, a simple analysis of the reading performance of each individual child can be done. A common metric is to evaluate reading speed considering only correctly read words, which is defined as Words Correct per Minute (WCPM) (Hasbrouck and Tindal, 2006). The average values of WCPM per grade of 80 children at the end of school year are shown in Table 3 for sentence reading tasks and Table 4 for pseudoword reading tasks, side-by-side with the target curricular goals. A large inter-grade overlap of the distributions is observed, showing a variability in reading performance of different children, although the average does increase per grade. For sentence reading, the difference from average WCPM to curricular goals increases in absolute terms along the grades, and these lower WCPM values may be explained by the difficulty of the reading tasks. It can be concluded that the suggested increase of difficulty along the grades could be too steep to directly evaluate CG as intended, and, for overall reading ability evaluation, this difficulty needs to be taken into account. For pseudowords, although there are no CG for the third and fourth grades, WCPM values are significantly lower than CG, suggesting that the created pseudowords (based on joining common syllables and not on existing words) are of high difficulty.

	Words in Sentences		
Grade	WCPM	CG	WCPM-CG
1st	59.7±18.1	55	+8.5%
2nd	85.2±22.9	90	-5.3%
3rd	97.1±23.5	110	-11.7%
4th	104.1±23.0	125	-16.7%

Table 3: Per grade Mean and Standard Deviation of Words Correct per Minute (WCPM), Curricular Goals (CG) and relative difference of WCPM to CG, for sentence reading tasks.

	Pseudowords		
Grade	WCPM	CG	WCPM-CG
1st	18.8±8.0	25	-24.8%
2nd	26.7±8.4	35	-23.7%
3rd	26.1±6.5	-	
4th	34.9±9.6	-	

Table 4: Per grade Mean and Standard Deviation of Words Correct per Minute (WCPM), Curricular Goals (CG) and relative difference of WCPM to CG, for pseudoword reading tasks.

To analyze if the difficulty parameter used is suitable, we can compare the number of disfluencies given by children with the computed difficulty. Figure 3 shows a histogram-like distribution of the average number of disfluencies given per pseudoword through intervals of the difficulty parameter.

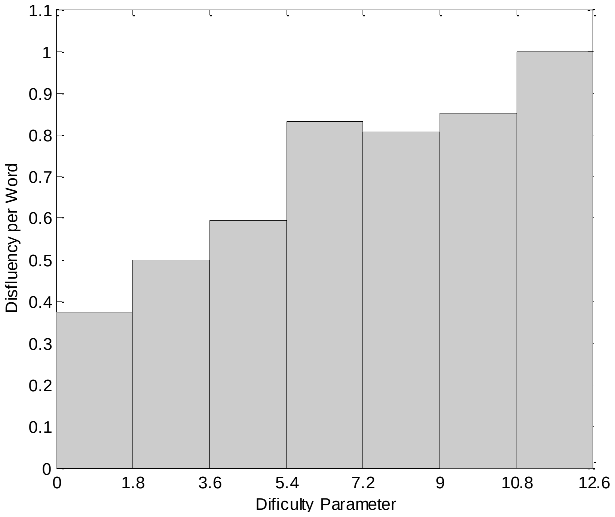


Figure 3: Average number of disfluencies per pseudoword for different difficulty parameter intervals.

An increase of disfluencies with higher difficulties is evident, showing that the parameter seems to be adequate. Due to the annotated data not presenting many repeated samples of the same tasks (resulting in, e.g., some pseudowords of high difficulty presenting zero disfluencies on three utterances), the correlation between difficulty and average number of disfluencies cannot be accurately computed. For future work, in order to improve the parameter of difficulty and take it into consideration when evaluating reading ability, the weight of the several sub-parameters of difficulty level calculation can be optimized by fitting them to the collected data.

5. Conclusion

A database of Portuguese children reading aloud was collected, covering a large variation of reading performance levels of 6-10 years old children. It also contains significant occurrences of the most common disfluency types in reading aloud. Besides a direct study of the data, as executed in this paper, the main purpose of the corpus is to develop automatic methods for evaluating reading performance.

The corpus is made available at ftp://193.136.94.104 with audio files, meta-data and descriptive documentation. It falls under the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0). The included automatic alignment of utterances is likely to be improved and revised for the foreseeable future.

6. Acknowledgements

This work was supported in part by Fundação para a Ciência e Tecnologia under the projects UID/EEA/50008/2013 (pluriannual funding in the scope of the LETSREAD project), and Marie Curie Action IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP). Jorge Proença is supported by the SFRH/BD/97204/2013 FCT Grant. We would like to thank *João de Deus*, *Bissaya Barreto* and *EBI de Pereira* school associations and CASPAE parent's association for collaborating in the database collection.

7. Bibliographical References

Black, M.P., Tepperman, J., and Narayanan, S.S. (2011). Automatic Prediction of Children's Reading Ability for High-Level Literacy Assessment. Trans. Audio, Speech and Lang. Proc. *19*, pp. 1015–1028.

Buescu, H.C., Morais, J., Rocha, M.R., and Magalhães, V.F. (2015). Programa e Metas Curriculares de Portugês do Ensino Básico (Ministério da Educação e Ciência).

Candeias, S., Celorico, D., Proença, J., Veiga, A., and Perdigão, F. (2013). HESITA(tions) in Portuguese: a database. In ISCA, Interspeech Satellite Workshop on Disfluency in Spontaneous Speech - DiSS, (KTH Royal Institute of Technology, Stockholm, Sweden), pp. 13–16.

Duchateau, J., Cleuren, L., hamme, H.V., and Ghesquière, P. (2007). Automatic assessment of children's reading level. In Proc. Interspeech, (Antwerp, Belgium: ISCA), pp. 1210–1213.

Hämäläinen, A., Rodrigues, S., Júdice, A., Silva, S.M., Calado, A., Pinto, F.M., and Dias, M.S. (2013). The CNG

Corpus of European Portuguese Children's Speech. In Text, Speech, and Dialogue, I. Habernal, and V. Matoušek, eds. (Springer Berlin Heidelberg), pp. 544–551.

Hämäläinen, A., Cho, H., Candeias, S., Pellegrini, T., Abad, A., Tjalve, M., Trancoso, I., and Dias, M. (2014a). Automatically Recognising European Portuguese Children's Speech: Pronunciation Patterns Revealed by an Analysis of ASR Errors. In Proc. International Conf. on Computational Processing of Portuguese - PROPOR, (São Paulo, Brazil), pp. 1–11.

Hämäläinen, A., Candeias, S., Cho, H., Meinedo, H., Abad, A., Pellegrini, T., Tjalve, M., Trancoso, I., and Dias, M.S. (2014b). Correlating ASR Errors with Developmental Changes in Speech Production: A Study of 3-10-Year-Old European Portuguese Children's Speech. In Proc. WOCCI 2014 – Workshop on Child Computer Interaction, (Singapore), pp. 7–11.

Hasbrouck, J., and Tindal, G.A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. The Reading Teacher *59*, pp. 636–644.

Keuleers, E., and Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. Behav Res Methods *42*, pp. 627–633.

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

Lopes, C., Veiga, A., and Perdigão, F. (2012). A European Portuguese Children Speech Database for Computer Aided Speech Therapy. In Computational Processing of the Portuguese Language, H. Caseli, A. Villavicencio, A. Teixeira, and F. Perdigão, eds. (Springer Berlin Heidelberg), pp. 368–374.

Mendonça, G., Candeias, S., Perdigao, F., Shulby, C., Toniazzo, R., Klautau, A., and Aluisio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. In Proc. of the International Telecommunications Symposium (ITS), (São Paulo, Brazil), pp. 1–5.

Potamianos, A., and Narayanan, S. (2003). Robust recognition of children's speech. IEEE Transactions on Speech and Audio Processing *11*, pp. 603–616.

Proença, J., Celorico, D., Candeias, S., Lopes, C., and Perdigão, F. (2015). Children's Reading Aloud Performance: a Database and Automatic Detection of Disfluencies. In ISCA - Conf. of the International Speech Communication Association - INTERSPEECH, (Dresden, Germany), pp. 1655–1659.

Rocha, P., and Santos, D. (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In PROPOR, pp. 131–140.

Santos, A.L., Généreux, M., Cardoso, A., Agostinho, C., and Abalada, S. (2014). A Corpus of European Portuguese Child and Child-directed Speech. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), (Reykjavik, Iceland), pp. 1488–1491.

Soares, A.P., Medeiros, J.C., Simões, A., Machado, J., Costa, A., Iriarte, Á., de Almeida, J.J., Pinheiro, A.P., and Comesaña, M. (2014). ESCOLEX: a grade-level lexical database from European Portuguese elementary to middle school textbooks. Behav Res Methods *46*, pp. 240–253.

8. Language Resource References

Lopes, C. (2012). ChildCAST - European Portuguese Children Speech Database for Computer Aided Speech Therapy. Available at http://lsi.co.it.pt/spl/childCAST/ChildCAST_v3.zip.

Santos, A.L. (2014). Corpus Santos - European Portuguese. ISLRN 532-620-702-768-3. http://www.clul.ul.pt/en/resources/546.

Santos, D., and Rocha, P. (2001). CETEMpublico. LDC2001T62. ISLRN 544-982-311-455-3.

Speecon Consortium (2005). Portuguese Speecon Database. ELRA-S0180, ISLRN 824-839-200-501-4.