Avaliação de fluência leitora em língua portuguesa: primeira experiência com uso em larga escala de Inteligência Artificial

Caio C. Rocha[1], **Rômulo C. de Mello**[1], **Jairo F. de Souza**[1,2]

[1]LApIC Research Group – Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora – MG – Brazil

[2]Departamento de Ciência da Computação – UFJF
Juiz de Fora – MG – Brazil

{caiocedrola,romulomello,jairo.souza}@ice.ufjf.br

Abstract. *With recent advances in AI, the use of automatic systems for assessing learning has grown. This study presents the results of the first large-scale automatic assessment of reading fluency in Portuguese carried out with second-year elementary school students. The results were compared with a significant volume of manual corrections performed by trained correctors. Although the solution doesn't cover all the data produced by humans, the results show that automatic marking is compatible with the data produced by humans, constituting a viable and economical alternative for large-scale assessments.*

Resumo. *Com os recentes avanços da IA, tem crescido o uso de sistemas automáticos para avaliação de aprendizagem. Este estudo apresenta os resultados da primeira avaliação automática de fluência leitora em língua portuguesa realizada em larga escala com estudantes do segundo ano do ensino fundamental. Os resultados automáticos foram comparados com um volume expressivo de correções manuais realizado por corretores treinados. Embora a solução ainda não contemple todos os aspectos da avaliação, os resultados mostram que a correção automática é capaz de produzir dados compatíveis com os dados produzidos por corretores humanos, proporcionando uma alternativa viável e econômica para avaliações em larga escala.*

1. Introdução

A avaliação de fluência em leitura mede o nível de fluência de alunos em fase de aprendizado. Sistemas automáticos de avaliação de fluência reduzem tempo e custos [de Assis et al. 2022], proporcionando uma análise mais objetiva que a humana. Com os avanços da IA, o uso desses sistemas na educação tem crescido [Forero-Corba and Bennasar 2024, Zhai et al. 2021, Messer et al. 2024, González-Calatayud et al. 2021], auxiliando professores na adoção de uma metodologia de avaliação formativa e em boas práticas de avaliação [Jones 2005].

Contudo, muitas propostas de IA na educação encontram-se em fase de prototipação por dificuldades de validação devido à necessidade de um grande volume de dados [Memarian and Doleck 2023, Xu 2020]. Na avaliação de fluência em leitura oral, o investimento financeiro e de recursos humanos é elevado [Soares et al. 2018], pois os avaliadores precisam escutar áudios e anotar diversas métricas.

DOI: 10.5753/sbie.2024.244971

A avaliação de fluência oral do PARC é composta pela leitura de três itens: um texto narrativo breve, uma lista de 60 palavras e uma lista de 40 pseudo-palavras, seguindo critérios como extensão, tonicidade e complexidade silábica [Lemle 1987]. As pseudopalavras avaliam a decodificação grafêmica [Pinheiro and Vilhena 2022], oferecendo uma análise da consciência fonética e das correspondências entre letras e sons, desvinculada de conhecimentos prévios [Coscarelli 2002]. Neste projeto, alunos do 2º ano do Ensino Fundamental (2EF) são instruídos a ler todo o caderno avaliativo, guiados por um professor responsável, que grava 60 segundos da leitura de cada item em um aplicativo de celular. Os áudios, enviados aos servidores do projeto, podem conter ruídos de fundo, pois o ambiente de gravação não é controlado. Em seguida, são analisados por corretores treinados que informam as métricas definidas na Figura 1.

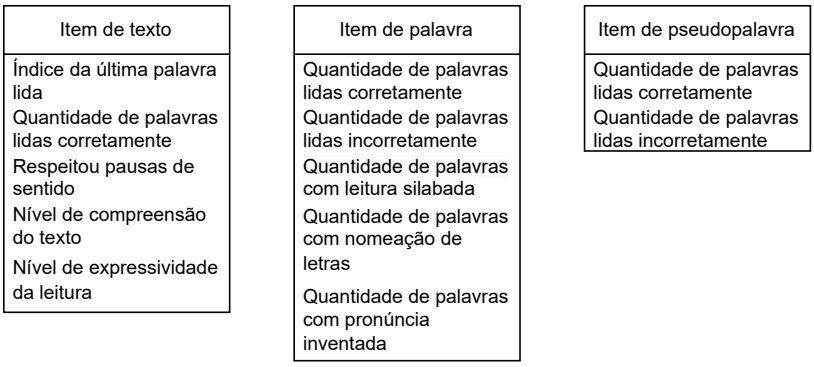


Figura 1. Diagrama de métricas coletadas por item avaliativo.

Embora a avaliação seja tradicionalmente manual, são apresentados neste artigo o resultado do uso de técnicas de reconhecimento de fala para apoiar a avaliação de 680.991 estudantes do 2EF de vários estados brasileiros. Esta é, até onde os autores têm conhecimento, a maior aplicação de correção automática para avaliação de alfabetização em língua portuguesa. Os resultados da avaliação são comparados com os resultados da avaliação manual realizada por corretores treinados. Embora, até o momento, apenas sejam retornados as métricas de Última Palavra Lida, Quantidade de Palavras Lidas e Quantidade de Palavras Lidas Incorretamente (Figura 1, os resultados mostram a qualidade da solução em um cenário real de avaliação de alunos em fase de alfabetização.

2. Revisão da literatura

A avaliação da fluência em leitura com tecnologias de reconhecimento automático de fala (ASR) tem sido explorada em diversos estudos. Analisamos como esses trabalhos se comparam em contexto, tecnologia utilizada e métodos de avaliação.

Os estudos de fluência com ASR abrangem contextos educacionais e linguísticos. [Yıldız et al. 2024] avaliaram a leitura de 120 alunos de escolas públicas na Turquia utilizando ASR. [RODRIGUES et al. 2023] focaram em alunos do Ensino Fundamental no Brasil, avaliando a leitura em português brasileiro. [Zhang et al. 2012] conduziram um estudo com 4.000 estudantes na China, avaliando a leitura de textos em inglês como segunda língua. [Bolaños et al. 2013] trabalharam com 313 alunos nos EUA, usando o sistema FLORA para avaliar a leitura em inglês. [Proença et al. 2017] investigaram a

fluência de crianças portuguesas utilizando frases e pseudopalavras em português europeu. [Bailly et al. 2022] fizeram um estudo com 442 crianças na França, avaliando a leitura em francês com um sistema ASR. [Bernstein and Cheng 2023] descreveram métodos para extração de conteúdo lexical de respostas faladas, destacando a importância de aspectos extralinguísticos na avaliação de fluência.

As tecnologias ASR variam em abordagem. [Yıldız et al. 2024] usaram um modelo de IA para turco, empregando técnicas de regressão para comparar pontuações com especialistas humanos. [RODRIGUES et al. 2023] usaram Wav2vec2, adaptado para reconhecimento de fala de crianças brasileiras. [Zhang et al. 2012] usaram SVM para avaliar textos em inglês. O sistema FLORA, descrito por [Bolaños et al. 2013], utilizou técnicas de *machine learning* para medir a precisão e a taxa de leitura, assim como a expressividade. [Proença et al. 2017] usaram o corpus LetsRead e modelos de *machine learning* para avaliar o português europeu. [Bailly et al. 2022] propuseram um framework que combinava características linguísticas e prosódicas para avaliar a fluência leitora.

Os métodos de avaliação também variam. [Yıldız et al. 2024] compararam as avaliações do sistema com especialistas humanos, mostrando concordâncias nas pontuações. [RODRIGUES et al. 2023] compararam os dados do pós-processamento com as avaliações humanas, destacando a precisão do sistema ASR. [Zhang et al. 2012] treinaram seu modelo com 3200 estudantes e testaram com 800, mostrando pequena diferença entre as pontuações automáticas e humanas. [Bolaños et al. 2013] avaliaram o sistema FLORA com 783 gravações, demonstrando que as pontuações geradas estavam muito próximas das dadas por avaliadores humanos. [Proença et al. 2017] usaram um *corpus* específico e modelos de *machine learning* para detectar erros de pronúncia e disfluências. [Bailly et al. 2022] validaram seu sistema ASR com 1063 gravações, mostrando que características prosódicas melhoraram a previsão de expressividade. [Bernstein and Cheng 2023] destacaram a aplicação de ASR em testes de segunda língua, abordando limites e oportunidades para melhorias.

Como em [RODRIGUES et al. 2023], o presente trabalho usa tecnologias atuais para processamento de fala em português, mas utiliza outras informações além da transcrição automática. Diferente dos outros, esta solução foi projetada para avaliação de falantes em idade de alfabetização na sua língua materna. O volume de dados utilizado permite uma análise mais confiável, oferecendo uma contribuição significativa para a literatura. Além disso, os áudios captados foram em ambiente não controlado e com microfones de baixa qualidade.

3. Métricas da avaliação de fluência leitora e correção automática

Os corretores humanos responsáveis por anotar as métricas definidas na Figura 1 determinam o perfil leitor de cada estudante. Os perfis são: pré-leitor, iniciante e fluente; determinados pela quantidade de palavras lidas (QPL) e pela quantidade de palavras lidas corretamente (QPC) de acordo com cada item lido (texto, palavra e pseudo). A Equação 1 especifica as regras utilizadas na classificação.

$$\begin{cases} \text{Fluente} \leftarrow QPC_{texto} > 65 \ \wedge \ \frac{QPC_{texto}}{QPL_{texto}} \geq 0,9 \\ \text{Iniciante} \leftarrow \neg\text{Fluente} \wedge QPC_{palavra} > 10 \ \wedge \ QPC_{pseudo} > 5 \\ \text{Pré-leitor} \leftarrow \neg\text{Fluente} \wedge \neg\text{Iniciante} \end{cases} \quad (1)$$

Além disso, o perfil de pré-leitor é subdividido em 6 níveis, que dependem da quantidade de palavras com leitura silabada, da quantidade de palavras com nomeação de letras e de informações mais subjetivas, como o ritmo da leitura e a quantidade de palavras com a pronúncia inventada.

A correção automatizada das leituras foi feita separadamente da avaliação manual, pois o objetivo da pesquisa é validar a viabilidade da correção automatizada em avaliações de fluência em larga escala. O fluxograma da Figura 2 mostra o funcionamento do sistema.

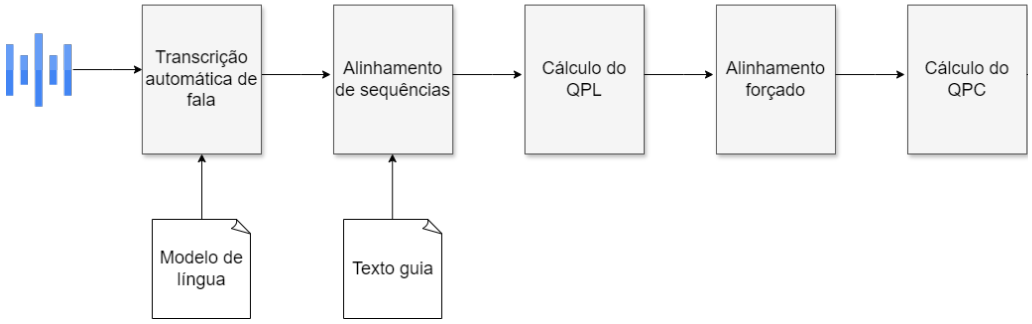


Figura 2. Fluxograma do sistema de avaliação automática.

Inicialmente, é utilizado um ajuste fino do modelo wav2vec 2.0 para o reconhecimento de fala do português brasileiro [Grosman 2022]. O modelo acústico foi utilizado em conjunto com um modelo de língua (LM) criado com unigramas e bigramas derivado de cada item do teste, com o papel de aproximar as transcrições ao texto guia lido pelo aluno.

A etapa de alinhamento de sequências entre as transcrições e os textos guia permite a determinação da Quantidade de Palavras Lidas (QPL) em cada leitura. O QPL indica o índice da última palavra lida corretamente pelo aluno, sendo uma métrica que reflete o progresso na leitura do texto. Para calcular o QPL, é realizado um alinhamento utilizando técnicas de busca aproximada, que permitem encontrar correspondências entre as palavras da transcrição e as do texto de referência, mesmo quando não há uma correspondência exata. Em casos onde a correspondência entre as sequências não é suficiente, ferramentas de alinhamento textual mais refinadas são empregadas, considerando palavras substituídas, deletadas ou inseridas.

Após calcular o QPL, realiza-se o alinhamento forçado entre os áudios e as sequências de texto, permitindo uma análise detalhada dos segmentos de áudio para identificar a pronúncia de cada palavra. Esse processo envolve a criação de uma matriz de probabilidades que representa as letras pronunciadas em cada instante de tempo. A matriz é gerada a partir da representação vetorial do áudio, onde a probabilidade de cada letra é estimada para cada frame de tempo. Essas probabilidades são comparadas com as sequências de letras esperadas, de acordo com o texto de referência. O alinhamento

das letras é realizado para encontrar a correspondência mais provável entre a pronúncia registrada e a sequência de letras esperada.

O cálculo da Quantidade de Palavras Corretamente Lidas (QPC) é feito a partir da análise dessas probabilidades ao longo do tempo. Para uma palavra ser considerada corretamente lida, é necessário que todas as letras previstas na sequência de referência estejam presentes e corretamente alinhadas na sequência predita pelo modelo. Se houver alguma discrepância, como letras ausentes ou substituídas, a leitura da palavra é classificada como incorreta. Esse método de cálculo permite uma avaliação mais precisa e detalhada da leitura de cada aluno.[1]. Por fim, com base nos resultados das métricas de QPL e QPC, os alunos foram classificados em diferentes perfis leitores, conforme descrito na Seção 3.

Como os limiares de classificação são rígidos, áudios com valores próximos desse limiar tem maior probabilidade de terem sido classificados incorretamente pela IA. Assim, propõe-se a adoção de uma zona cinzenta, na qual abrange áudios que deveriam ser corrigidos manualmente quando a solução estiver em ambiente de produção. Espera-se, neste caso, que a zona cinzenta contenha um mínimo de áudios, mas que filtre o máximo de áudios que possam ter sido classificados incorretamente. Após experimentos [Silva et al. 2022], as regras de filtro para a triagem de áudios na zona cinzenta são definidas na Equação 2 e na próxima seção é analisado o efeito do uso dessa abordagem.

$$\begin{cases} \text{Texto:} & 55 < QPC_{texto} \leq 65 \ \vee \ (QPC_{texto} > 40 \ \wedge \ 0,5 < \frac{QPC_{texto}}{QPL_{texto}} < 0,9) \\ \text{Palavra:} & 9 \leq QPC_{palavra} \leq 13 \\ \text{Pseudo:} & 4 \leq QPC_{pseudo} \leq 8 \end{cases} \tag{2}$$

4. Resultados

Os resultados comparativos entre a correção humana e da IA são discutidos em duas partes: a avaliação automática por item (Seção 4.1) e a classificação final (Seção 4.2).

4.1. Resultados por item avaliado

A classificação binária de cada item separa as leituras em *acima do limiar* e *abaixo do limiar*. As tabelas 1, 2, e 3 resumem os resultados.

As classificações de texto apresentaram o melhor desempenho, seguidas pelas classificações de palavras e pseudopalavras. Como esperado, a classificação de pseudopalavras foi mais difícil devido às pronúncias ambíguas e à maior complexidade do item. Ainda, a remoção de áudios na área cinzenta aumentou a acurácia, precisão e revocação.

A Figura 3 mostra as distribuições dos valores de QPC automáticos e manuais, indicando um desempenho esperado do sistema. No entanto, houve diferença na distribuição de leituras com QPC baixo, especialmente em $QPC_{texto} \in [0, 10]$, $QPC_{palavra} \in [0, 5]$, e $QPC_{pseudo} \in [0, 3]$, devido a pronúncias confusas e transcrições

[1]Por limitação de espaço, decidiu-se omitir neste artigo as estratégias computacionais desenvolvidas para correção automática e se limitar à análise dos resultados alcançados em uma avaliação em larga escala. Para mais informações sobre a tecnologia desenvolvida, veja [de Assis et al. 2022, Almeida Silva et al. 2021, Soares et al. 2018]

	Sem filtro	Com filtro
% de áudios filtrados		10,6
Acurácia	0,961	0,966
Precisão (acima do limiar)	0,750	0,750
Revocação (acima do limiar)	0,883	0,976
Precisão (abaixo do limiar)	0,988	0,997
Revocação (abaixo do limiar)	0,969	0,965

Tabela 1. Resultados da classificação do item de texto.

	Sem filtro	Com filtro
% de áudios filtrados		6,9
Acurácia	0,926	0,947
Precisão (acima do limiar)	0,918	0,942
Revocação (acima do limiar)	0,911	0,932
Precisão (abaixo do limiar)	0,932	0,950
Revocação (abaixo do limiar)	0,937	0,958

Tabela 2. Resultados da classificação do item de palavra.

	Sem filtro	Com filtro
% de áudios filtrados		11,5
Acurácia	0,902	0,935
Precisão (acima do limiar)	0,955	0,976
Revocação (acima do limiar)	0,857	0,896
Precisão (abaixo do limiar)	0,853	0,898
Revocação (abaixo do limiar)	0,954	0,976

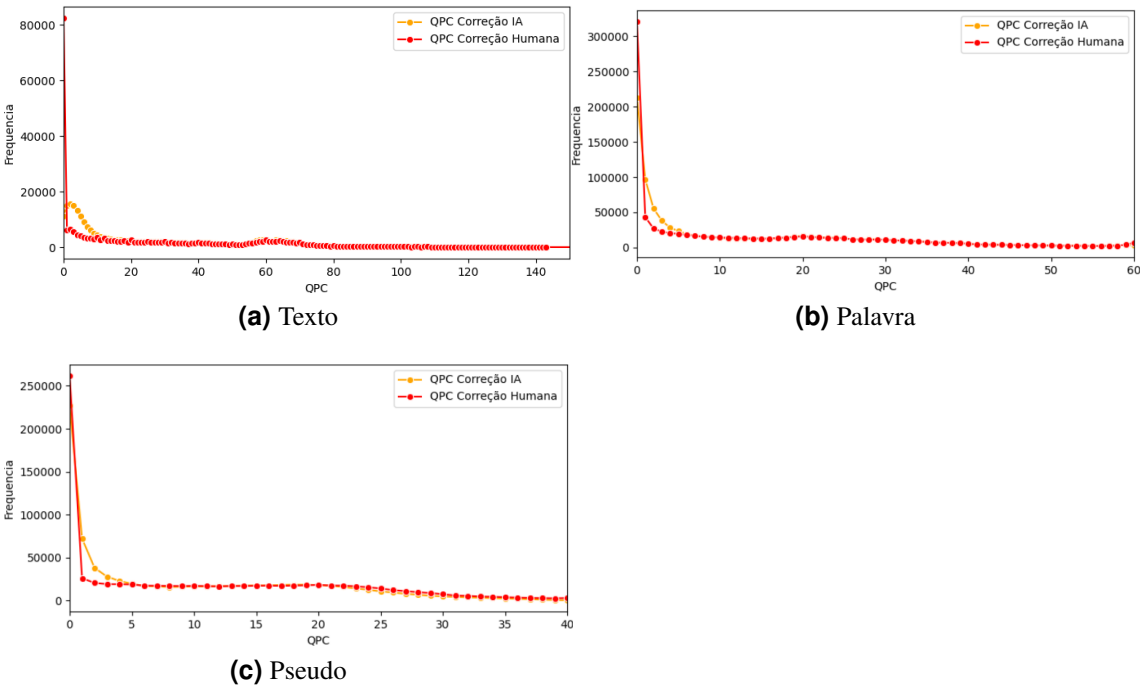
Tabela 3. Resultados da classificação do item de pseudo-palavras.

distantes do texto guia, dificultando a avaliação. Contudo, como esses valores não estão próximos aos limiares de classificação, não impactaram negativamente a classificação.

4.2. Resultados da classificação final

O comparativo entre as classificações nos três perfis leitores é apresentado na Tabela 4. O número de alunos classificados como pré-leitores, iniciantes e fluentes pela IA ficou muito próximo dos números obtidos pela avaliação manual, o que demonstra a precisão do sistema automático. No entanto, algumas variações ocorreram na classificação automática devido a certos alunos não cumprirem as regras específicas de classificação descritas na Equação 1. Isso significa que alguns alunos não puderam ser classificados automaticamente, pois os dados coletados não atenderam aos critérios estabelecidos para a classificação. O filtro aplicado permitiu que mais áudios fossem classificados, pois dados inicialmente indisponíveis foram substituídos pelos da avaliação manual, aumentando a acurácia ponderada com apenas 9% dos registros corrigidos manualmente.

As acurácias para todas as abordagens ficaram próximas. No entanto, a acurácia ponderada, que equilibra a acurácia entre pré-leitores, iniciantes e fluentes, mostrou que o desempenho caiu em um cenário equilibrado entre as três classes. Melhorias são ne-



(a) Texto **(b)** Palavra

(c) Pseudo

Figura 3. Distribuição do QPC automático e do QPC manual.

Tabela 4. Resultados da classificação final (PL = pré-leitores, I=iniciantes, F=fluentes)

Correção	#PL	#I	#F	*#alunos*	#registros filtrados	Acurácia	Acurácia ponderada
Manual	592.079 (86,94%)	65.037 (9,55%)	23.875 (3,51%)	680.991			
Automática (sem filtro)	560.458 (86,44%)	61.620 (9,50%)	26.330 (4,06%)	648.408		**97,07%**	**91,15%**
Automática (com filtro)	569.839 (86,46%)	60.625 (9,20%)	28.602 (4,34%)	659.066	121.129 (9,03%)	**97,94%**	**94,75%**

cessárias, especialmente na classificação dos *iniciantes*. Como mostra a Figura 4, as maiores confusões ocorreram entre *iniciantes* e *pré-leitores*, e entre *iniciantes* e *fluentes*. Como essas classes representam níveis de qualidade de leitura, é esperado um maior número de erros entre classes adjacentes.

Destaca-se o erro de classificação entre classes não adjacentes, como fluentes e pré-leitores. Uma verificação manual dos áudios revelou erros humanos nas correções, como como atribuição incorreta de valores, intervenções do avaliador, erros de aplicação (a criança não era quem lia) e cadastro (leitura não aplicada ou áudio errado). Dos 400 registros verificados (128 falsos pré-leitores e 272 falsos fluentes), aproximadamente 53% dos problemas ocorreram na correção de textos, 38% em palavras e 9% em pseudopalavras. Cerca de 86% foram causados por atribuição incorreta de valores, 7% por intervenções do professor durante a leitura (falas inesperadas) e 7% por erros de cadastro ou erro de aplicação (áudio contendo a leitura de um caderno diferente da referência).

PREDIÇÃO

REFERÊNCIA	PRÉ-LEITOR	INICIANTE	FLUENTE
PRÉ-LEITOR	565.884	4.183	272
INICIANTE	3.827	56.232	4.978
FLUENTE	128	210	23.352

Figura 4. Matriz de confusão da classificação final.

Esses resultados demonstram que, ao comparar com grandes bases sem um controle rígido de qualidade, os dados rotulados manualmente não podem ser considerados padrão ouro e a comparação com os dados gerados automaticamente deve ser encarada como o grau de concordância entre o dado humano e a predição do sistema, ou seja, nem toda discordância do sistema automático pode ser considerada como um erro. Destas 400 discordâncias, apenas 26 registros (6,50%) representam erros do sistema.

5. Conclusão

Este estudo demonstra a eficácia da tecnologia de reconhecimento automático de fala (ASR) na avaliação da fluência leitora de estudantes do Ensino Fundamental. A correção automática avaliou com precisão a quantidade de palavras lidas (QPL) e corretamente lidas (QPC), mostrando convergência com avaliações manuais de corretores treinados. Diferente de estudos anteriores com amostras menores, este estudo utilizou dados de 680.991 estudantes, proporcionando uma análise robusta e confiável.

Além da transcrição, o estudo utilizou informações adicionais, como alinhamento forçado e probabilidades atribuídas pelo modelo de ASR, proporcionando uma avaliação mais detalhada das habilidades de leitura. A implementação do sistema ASR tem potencial para reduzir o tempo, custos e subjetividade das avaliações manuais. A acurácia final de 97,94% e a alta concordância com as correções manuais confirmam a viabilidade do uso de ASR na avaliação de fluência em larga escala.

Por outro lado, uma avaliação de fluência leitora deve levar em consideração informações (ver Figura 1) ainda não contempladas neste projeto, as quais possuem desafios de pesquisa e necessitam de soluções não triviais. Por exemplo, há um alto nível de discordância entre corretores humanos quanto à percepção de silabação, o que está relacionado à presença de uma ou mais pausas entre sílabas, ao ritmo de leitura da criança e ao prolongamento de vogais. Assim, são próximos passos deste projeto fornecer soluções para cobrir as informações faltantes da avaliação de fluência leitora, como a identificação de silabações, nomeações de palavras, expressividade e compreensão da leitura.

Referências

Almeida Silva, W., Carchedi, L., Gomes Jr, J., Souza, J., Barrere, E., and Souza, J. (2021). A framework for large-scale automatic fluency assessment. *International Journal of Distance Education Technologies*, 19.

Bailly, G., Godde, E., Piat-Marchand, A.-L., and Bosse, M.-L. (2022). Automatic assessment of oral readings of young pupils. *Speech Communication*, 138:67–79.

Bernstein, J. C. and Cheng, J. (2023). Speech analysis in assessment. *Advancing Natural Language Processing in Educational Assessment*, pages 31–57.

Bolaños, D., Cole, R., Ward, W., Tindal, G., Hasbrouck, J., and Schwanenflugel, P. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology*, 105:1142.

Coscarelli, C. V. (2002). Entendendo a leitura. *Revista de estudos da linguagem*, 10(1):7–27.

de Assis, E., Ferreira, A. L., Silva, C., and de Souza, J. (2022). Classificação automática de áudios de leituras de pseudopalavras para avaliação em larga escala de fluência da leitura de crianças em fase de alfabetização. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 27–38, Porto Alegre, RS, Brasil. SBC.

Forero-Corba, W. and Bennasar, F. N. (2024). Techniques and applications of machine learning and artificial intelligence in education: a systematic review. *RIED-Revista Iberoamericana de Educación a Distancia*, 27(1).

González-Calatayud, V., Prendes-Espinosa, P., and Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12):5467.

Grosman, J. (2022). Fine-tuned XLS-R 1B model for speech recognition in Portuguese. `https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-portuguese`.

Jones, C. (2005). Teachers need help too: aiding the marking process through a human-computer collaborative approach. In *Human Centred Technology Workshop 2005*, page 56.

Lemle, M. (1987). *Guia teórico do alfabetizador*. Ed. Ática.

Memarian, B. and Doleck, T. (2023). A review of assessment for learning with artificial intelligence. *Computers in Human Behavior: Artificial Humans*, page 100040.

Messer, M., Brown, N. C., Kölling, M., and Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education*, 24(1):1–43.

Pinheiro, A. and Vilhena, D. (2022). Teste de reconhecimento de palavras e pseudopalavras: validades de conteúdo e externa. *Signo*, 47:147–164.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., and Perdigão, F. (2017). Automatic evaluation of children reading aloud on sentences and pseudowords. In *INTERSPEECH*, pages 2749–2753.

RODRIGUES, A., Ribeiro, G., Silva, V., Carvalho, W., Ramírez, M., Alves, L., Finger, M., Navas, A. L., and Ribeiro, C. (2023). Ai and reading fluency for brazilian portuguese: A preliminary study. *Preprint available at SSRN 4429229*.

Silva, C. N., Ferreira, A. L. V., de Assis, E. C., and de Souza, J. F. (2022). Definição de heurística para identificação automática da fluência em leitura de crianças em fase de alfabetização. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 39–50. SBC.

Soares, E., Carchedi, L. C., Gomes Jr, J., Barrére, E., and Souza, J. (2018). Avaliaçao automática da fluência em leitura para crianças em fase de alfabetizaçao. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 11.

Xu, L. (2020). The dilemma and countermeasures of ai in educational application. In *Proceedings of the 2020 4th International Conference on Computer Science and Artificial Intelligence*, pages 289–294.

Yıldız, M., Keskin, H. K., Oyucu, S., Hartman, D. K., Temur, M., and Aydoğmuş, M. (2024). Can artificial intelligence identify reading fluency and level? comparison of human and machine performance. *Reading & Writing Quarterly*, pages 1–18.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. (2021). A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021(1):8812542.

Zhang, J., Pan, P., and Yan, Y. (2012). Automatic scoring on english passage reading quality. *Procedia Engineering*, 29:2744–2748. 2012 International Workshop on Information and Electronics Engineering.