Avaliação de eficiência na leitura: uma abordagem baseada em PLN

Túlio Sousa de Gois[1,2], **Raquel Meister Ko. Freitag**[1,3]

[1]Laboratório Multiusuário de Informática e Documentação Linguística
Universidade Federal de Sergipe – Brazil (UFS)
Didática II – 49.107-230 – São Cristóvão – SE – Brazil

[2]Departamento de Computação – Universidade Federal de Sergipe – UFS

[3]Departamento de Letras Vernáculas – Universidade Federal de Sergipe – UFS

{tuliosg, rkofreitag}@academico.ufs.br

Abstract. *The cloze test, widely used due to its low cost and flexibility, makes it possible to assess reading comprehension by filling in gaps in texts, requiring the mobilization of diverse linguistic repertoires. However, traditional correction methods, based only on exact answers, limit the identification of nuances in student performance. This study proposes an automated evaluation model for the cloze test in Brazilian Portuguese, integrating orthographic (edit distance), grammatical (POS tagging) and semantic (similarity between embeddings) analyses. The integrated method demonstrated its effectiveness, achieving a high correlation with human evaluation ($\rho = 0.832$). The results indicate that the automated approach is robust, sensitive to variations in linguistic repertoire and suitable for educational contexts that require scalability.*

Resumo. *O teste cloze, amplamente difundido por seu baixo custo e flexibilidade, permite avaliar a compreensão leitora por meio do preenchimento de lacunas em textos, exigindo mobilização de repertórios linguísticos diversos. No entanto, os métodos tradicionais de correção, baseados apenas em respostas exatas, limitam a identificação de nuances no desempenho dos estudantes. Este estudo propõe um modelo automatizado de avaliação para o teste cloze em português brasileiro, integrando análises ortográfica (distância de edição), gramatical (via POS tagging) e semântica (similaridade entre embeddings). O método integrado mostrou-se efetivo, atingindo uma alta correlação com a correção humana ($\rho = 0{,}832$). Os resultados indicam que a abordagem automatizada é robusta, sensível às variações do repertório linguístico e adequada para contextos educacionais que exigem escalabilidade.*

1. Introdução

Os problemas de leitura no Brasil são historicamente graves e foram intensificados pela pandemia de COVID-19, que expôs e acirrou ainda mais as desigualdades e educacionais preexistentes. Dados do PISA 2022 revelam que os estudantes brasileiros continuam apresentando desempenho insatisfatório em leitura, com escore de 410 pontos, inferior à de países vizinhos como Chile e Uruguai. Esses resultados reiteram que a leitura, enquanto

ferramenta básica para a construção do conhecimento e da cidadania, não tem sido desenvolvida de maneira efetiva na escola, comprometendo a formação crítica dos estudantes e sua participação plena na sociedade.

As avaliações diagnósticas, como Prova Brasil, PISA e PIRS, apresentam o panorama do sistema, mas não permitem que sejam implementadas medidas específicas em tempo hábil para atenuar as dificuldades dos estudantes com a leitura, de modo que as intervenções pedagógicas adequadas possam ser realizadas ainda no mesmo ano letivo. Assim, no desenvolvimento de instrumentos e matrizes de avaliação para o acompanhamento próximo e sistemático da leitura, o teste cloze tem se mostrado promissor por sua ampla difusão e baixo custo operacional [Freitas et al. 2025].

O teste cloze é um instrumento que avalia a compreensão leitora a partir da omissão de palavras em um texto, que devem ser completadas pelo estudante com base em seus conhecimentos linguísticos, textuais e contextuais. Por exigir a mobilização de diferentes habilidades cognitivas e linguísticas, o teste cloze permite diagnosticar não apenas o vocabulário e a gramática dominados pelo estudante, mas também seu grau de proficiência em leitura. Além de ser autoaplicável, esse tipo de teste pode ser adaptado a diferentes níveis de ensino e utilizado de forma contínua para acompanhar o desenvolvimento da leitura, tornando-se uma estratégia eficaz para planejar intervenções e promover avanços concretos no processo de aprendizagem.

O processo de correção do teste, no entanto, ainda baseia-se em um modelo de respostas certas/erradas, o que restringe a potencialidade de identificar especificidades da deficiência na compreensão leitora, em especial a diferenciação entre os níveis do repertório do estudante e o seu conhecimento gramatical [Cardoso et al. 2024]. O uso de ferramentas de processamento de linguagem natural tem permitido avançar neste campo, com a adoção de procedimentos para avaliação automática que saiam do limite do certo e errado, utilizando, por exemplo, similaridade semântica para acessar as lacunas preenchidas [de Gois et al. 2024].

Na continuidade do desenvolvimento e validação de um método automatizado para a avaliação de testes cloze em português brasileiro, expandindo os critérios de correção para além da tradicional verificação de respostas exatas, neste trabalho visamos integrar a análise ortográfica (via distância de edição), gramatical (via *POS tagging*) e semântica (via similaridade semântica de *embeddings*). Ampliando a exploração de aspectos do repertório, o trabalho também compara modelos de linguagem e a arquitetura de *embedding* (contextual vs. não contextual) mais eficazes para a tarefa de avaliação semântica. O desempenho do sistema final de avaliação automática através de seus resultados, é, então, comparado com as anotações de uma juíza especialista.

2. Antecedentes

Tradicionalmente, a correção do teste cloze prioriza a resposta exata como critério de avaliação [Freitas et al. 2025], considerando acertos apenas as palavras idênticas às do texto original. No entanto, há métodos alternativos de correção e/ou aplicação de técnicas computacionais ao processo.

[Kleijn et al. 2019] apresentam uma alternativa ao TC tradicional, o *Hybrid Text Comprehension cloze (HyTeC-cloze)*. Dentre as mudanças propostas, está a correção dos

testes utilizando uma pontuação semântica (também chamada de correção por resposta aceitável), que considera tanto as respostas exatas quanto as com semântica correta.

Em uma via distinta, [Mirault et al. 2021] propôs um algoritmo para correção do cloze baseado em regras. O método verifica se a resposta é uma palavra válida (usando bases lexicais), aplica remoção de afixos e calcula distâncias ortográficas (via `stringdist` [Van der Loo 2014]) para tolerar erros de grafia. Apesar do método automatizado, a abordagem não suporta a correção por resposta aceitável.

Uma estratégia baseada em categorização foi proposta por [Cardoso et al. 2024], que classificou as respostas por classe gramatical e campo semântico, considerando corretas aquelas alinhadas à palavra esperada em ambos os critérios. A eficiência em leitura foi medida combinando taxa de acerto e tempo de resolução, classificando os participantes quanto ao perfil na escala de [Bormuth 1968]. No entanto, a metodologia exige anotação manual das categorias, inviabilizando a aplicação em larga escala.

Buscando automatizar a análise semântica, o estudo de [de Gois et al. 2024] utilizou *word embeddings* (WEs) e similaridade de cosseno para avaliar respostas semanticamente próximas às esperadas. O ranking das palavras, gerado pela avaliação de similaridade, foi validado contra avaliações humanas. Embora promissor, o trabalho não integra critérios gramaticais ou tolerância a erros ortográficos, limitando sua aplicação prática.

Considerando estes antecedentes, neste trabalho propomos uma abordagem baseada em PLN que integra três níveis de análise: (1) verificação ortográfica via distância de Damerau-Levenshtein, (2) validação gramatical através da comparação de *POS tags*, e (3) avaliação semântica das respostas. Diferentemente dos apresentados anteriormente, o método proposto implementa um sistema de pontuação que preserva a robustez contra erros ortográficos, automatiza a avaliação semântica, considera acertos de classe gramatical e contabiliza as respostas por tipo.

3. Método

3.1. Dados

Os dados utilizados neste trabalho são provenientes da aplicação de testes cloze que ocorreu no Colégio de Aplicação da Universidade Federal de Sergipe (CODAP/UFS). Na coleta, foram aplicados quatro testes diferentes, em turmas do 6º ao 9º ano do ensino fundamental ($n = 210$) [Santos 2025].

As respostas foram tabuladas em planilhas, onde cada linha representa uma resposta ao teste, contendo as palavras inseridas nas lacunas e as informações do aluno respondente. Após essa etapa, os testes foram corrigidos por uma juíza especialista seguindo o padrão descrito em [Cardoso et al. 2024]: a resposta da lacuna era classificada quanto à classe gramatical e ao campo semântico. Para viabilizar uma análise quantitativa, foi realizado o mapeamento das categorias atribuídas pela juíza para um sistema de classificação padronizado.

3.2. Distância de edição Damerau-Levenshtein

A conferência de erros ortográficos é baseada no trabalho de [Mirault et al. 2021], que utilizou *Optimal String Alignment* (OSA) como parâmetro para aceitação da resposta. O OSA é uma variação da distância de Levenshtein, quantificando a diferença entre

duas palavras pelo número mínimo de edições para transformar uma na outra, mas considerando a transposição de letras vizinhas como custo 1. Em [Mirault et al. 2021], uma distância menor que 3 do gabarito era aceita. Contudo, esse valor arbitrário é problemático, pois o tamanho das palavras-alvo em um teste cloze varia.

Nossa abordagem utiliza a função `edit_distance` da biblioteca `NLTK` [Bird and Loper 2004], com parâmetro para considerar transposições, que implementa a distância de Damerau-Levenshtein [Damerau 1964, Levenshtein et al. 1966]. O critério de aceitação é dinâmico: a distância deve ser menor que $\frac{1}{3}$ do tamanho da resposta esperada (sendo 1 o valor mínimo), considerando assim a quantidade de caracteres da palavra-alvo na comparação, que ocorre exclusivamente com o gabarito.

3.3. *POS tags*

A verificação da classe gramatical de uma resposta é um indicador importante da compreensão sobre a estrutura sintática da frase [Cardoso et al. 2024], então também foi acrescida como ponto de avaliação. Para extrair essa informação, foi utilizada a biblioteca `spacy` [Honnibal et al. 2020] e a pipeline *pt_core_news_lg*[1]. A escolha do modelo ocorreu através devido ao desempenho na extração de *POS tags* para o Português Brasileiro (acurácia de $0{,}97$)[2] e também pelo tempo de execução.

Para alinhar o nível de detalhe do modelo (que distingue `VERB` de `AUX`) com a anotação humana (que trata ambos como "verbo"), implementou-se uma regra de mapeamento que agrupa a predição `AUX` na categoria `VERB` para fins de avaliação, garantindo uma comparação justa.

3.4. Similaridade semântica

A implementação da avaliação semântica das respostas é uma abordagem que se mostrou promissora [de Gois et al. 2024]. Neste trabalho, utilizamos modelos baseados em *transformers*, diferente do de [de Gois et al. 2024], que se baseou em modelos estáticos. Com base nos desempenhos para o português na tarefa de similaridade semântica proposta no ASSIN [Real et al. 2020], foram selecionados os modelos `BERTimbau Base` [Souza et al. 2020] e `Albertina 100M PTBR` [Santos et al. 2024].

Para a seleção do modelo mais eficaz na abordagem proposta, foi conduzido um experimento para determinar se embeddings contextuais (palavra-alvo inserida na frase) superavam as não contextuais (palavra-alvo isolada). O desempenho foi medido comparando os *scores* de similaridade de cosseno de cada abordagem com a classificação ordinal de uma juíza especialista. Para esta etapa, as anotações da juíza foram mapeadas para uma escala ordinal: respostas exatas valiam 3; aceitáveis, 2; de classe gramatical correta, 1; e incorretas ou em branco, 0. O experimento utilizou a biblioteca `transformers` [Wolf et al. 2019] para o uso dos modelos e `pytorch` [Paszke et al. 2019] para as operações de tensores.

3.5. Fluxo de avaliação

Partindo da integração das técnicas e da validação do modelo de linguagem, o método proposto para avaliação de cada lacuna do cloze segue o fluxo hierárquico presente na Figura 1. O sistema classifica cada resposta em uma das seguintes categorias:

[1] `https://spacy.io/models/pt#pt_core_news_lg`

[2] `https://spacy.io/models/pt#pt_core_news_lg-accuracy`

- **Exata**: idêntica ao gabarito (pontuação $1{,}0$);
- **Grafia incorreta**: resposta com pequenos erros ortográficos em relação ao gabarito (pontuação $1{,}0$);
- **Aceitável**: resposta semanticamente equivalente ao gabarito, com mesma classe gramatical (pontuação $1{,}0$);
- **Classe correta**: possui a mesma classe gramatical do gabarito, mas não é semanticamente equivalente (pontuação $0{,}5$);
- **Em branco**: a não-resposta (pontuação 0);
- **Incorreta**: resposta que não se enquadra nas demais categorias (pontuação 0);

A taxa de compreensão final do respondente é calculada como a média da pontuação obtida em todas as lacunas, expressa em percentual.

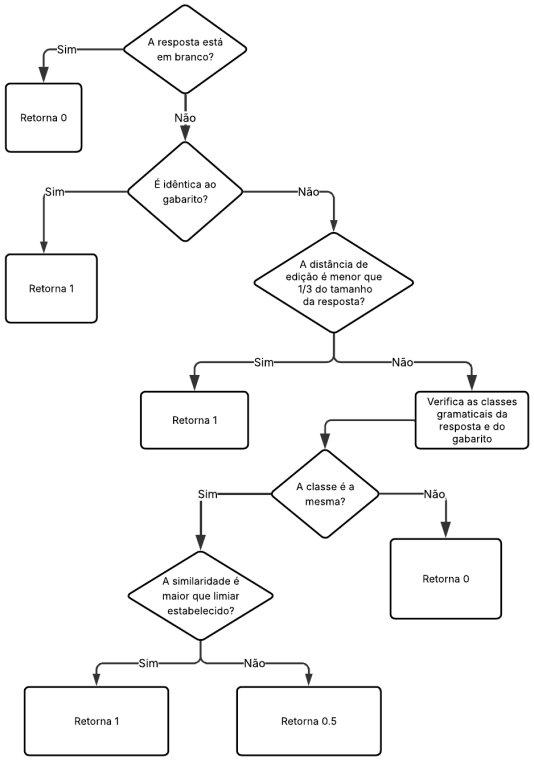

Figure 1. Fluxo para correção da lacuna

3.6. Implementação

A metodologia de avaliação descrita foi encapsulada em uma classe Python modular e reutilizável, denominada `NLPcloze`. Esta classe centraliza todas as funcionalidades necessárias, desde o carregamento dos modelos de linguagem e o gerenciamento de cache de *embeddings*, até a aplicação do fluxo de avaliação hierárquico em um conjunto de dados de respostas de alunos.

A arquitetura da classe foi projetada para ser extensível, permitindo a fácil substituição de modelos de linguagem ou o ajuste de parâmetros, como o limiar de aceitação semântica. O fluxo de decisão principal, implementado no método `avaliar_lacuna`, segue a lógica apresentada na Figura 1.

4. Resultados

4.1. Modelo de Linguagem e Limiar de Aceitabilidade

Os desempenhos dos modelos de linguagem foram avaliados através da correlação de Spearman com os dados ordinais da especialista, bem como por métricas de classificação (AUC e *F1-score*), onde as respostas foram categorizadas entre "aceitáveis" (*scores* 2 e 3) e "não-aceitáveis" (*scores* 0 e 1).

Table 1. Comparação das Métricas dos Modelos de Linguagem

Modelo	AUC	*F1-score* máximo	Limiar ótimo	Spearman
BERTimbau (Contexto)	0,880	0,727	0,652	0,682
Albertina (Contexto)	0,847	0,706	0,770	0,622
BERTimbau (Palavra)	0,815	0,637	0,877	0,532
Albertina (Palavra)	0,741	0,605	0,993	0,439

Conforme exposto na Tabela 1, o modelo BERTimbau, utilizando contexto, apresentou o melhor desempenho. Ele alcançou a maior correlação de Spearman com a avaliação humana ($\rho = 0{,}682$), a maior AUC ($0{,}880$) e o maior *F1-score* Máximo ($0{,}727$).

Partindo do experimento, consideramos o ponto onde o F1 é maximizado como o limiar ótimo. Para o BERTimbau (com contexto), o limiar foi de $0{,}652$, assim, estabelecemos o ponto de "aceitabilidade" para a nossa avaliação semântica.

4.2. Validação da Avaliação Automática

Após integrar o modelo e o limiar selecionados no fluxo de avaliação, a implementação final da abordagem foi validada contra o gabarito humano em duas frentes: análise de *scores* e concordância de categorias.

Primeiramente, foi comparada a pontuação final gerada pelo sistema com o *score* ordinal mapeado da avaliação da especialista (ver Seção 3.4). A abordagem proposta obteve uma alta correlação de Spearman, $\rho = 0{,}832$.

Em seguida, foi avaliada a concordância entre os rótulos categóricos. Para uma comparação justa, os rótulos do sistema (incluindo "grafia incorreta") e da juíza foram padronizados em um conjunto comum de classes (exata, aceitável, classe correta, incorreta). O coeficiente Kappa de Cohen (κ) foi de $0{,}727$, indicando uma concordância substancial, de acordo com [Landis and Koch 1977].

5. Conclusões

Neste estudo, expandimos os critérios tradicionais de correção em testes cloze para além da verificação de respostas exatas, ao integrar análises ortográfica (por meio da distância de edição), gramatical (via *POS tagging*) e semântica (com base na similaridade de cosseno entre *embeddings*). Também comparamos diferentes modelos de linguagem e arquiteturas de *embedding*, contextual e não contextual, para identificar as soluções mais eficazes na avaliação automatizada de aceitabilidade semântica. O desempenho do sistema final foi confrontado com as anotações de uma juíza especialista, a fim de validar sua aderência aos critérios humanos de julgamento.

Entre os modelos avaliados, o BERTimbau com contexto se destacou como o mais eficaz. Esse modelo apresentou os melhores resultados em todas as métricas analisadas: maior correlação de Spearman com os julgamentos humanos ($\rho = 0{,}682$), maior AUC ($0{,}880$) e *F1-score* máximo ($0{,}727$). Esses resultados demonstram sua capacidade de captar nuances contextuais relevantes para o julgamento semântico, o que é essencial em tarefas que demandam sensibilidade linguística próxima à humana.

Na fase de validação do método, os resultados apontaram para a alta correlação de Spearman ($\rho = 0{,}832$) entre os *scores* gerados pela correção automática e os atribuídos pela especialista, sugerindo forte alinhamento entre as avaliações. Além disso, a concordância categórica aferida pelo coeficiente Kappa de Cohen ($\kappa = 0{,}727$) foi considerada substancial, [Landis and Koch 1977], evidenciando que, mesmo após a categorização dos *scores*, o sistema mantém correspondência com as decisões humanas.

Estes resultados sugerem a confiabilidade da abordagem proposta, apontando seu potencial para uso em contextos educacionais e em pesquisa que demandam escalabilidade, precisão e sensibilidade semântica na avaliação de teste cloze em larga escala. Em suma, este estudo reforça a importância de aplicações em processamento de linguagem natural para o desenvolvimento de instrumentos e matrizes de avaliação para o acompanhamento próximo e sistemático da leitura, contribuindo para a educação de qualidade.

6. Limitações e Trabalhos futuros

Os testes cloze que resultaram nos dados utilizados no presente trabalho tinham apenas lacunas de verbos, o que pode enviesar os resultados aqui apresentados.

A necessidade de mapeamentos e a diferença entre as classificações atribuídas pela juíza especialista e pela abordagem proposta podem afetar a avaliação, a superestimando ou subestimando. Em trabalhos futuros, é indicada a avaliação por mais de um especialista, testes com lacunas de outras classes gramaticais e protocolos de correção bem definidos para facilitar a validação posterior.

Disponibilidade de Dados e Códigos

O conjunto de dados anonimizado utilizado para a avaliação, bem como o código-fonte completo da classe `NLPcloze` e os notebooks utilizados para as análises apresentadas neste trabalho, estão publicamente disponíveis em um repositório no GitHub para fins de reprodutibilidade e reuso: `https://github.com/tuliosg/nlp-cloze`.

Agradecimentos

Este trabalho está vinculado ao projeto *Impactos da pandemia de COVID-19 na linguagem da criança e do adulto: foco no desenvolvimento e na aprendizagem da leitura*, financiado pelo edital Capes 12/2021 - PDPG Impactos da Pandemia. Agradecemos ao suporte de infraestrutura e equipe do Laboratório Multiusuário de Informática e Documentação Linguística (LAMID) da Universidade Federal de Sergipe. Agradecemos também à Rede Brasileira de Reprodutibilidade pelo suporte financeiro, permitindo a participação no STIL.

References

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of educational measurement*, 5(3):189–196.

Cardoso, P. B., Menezes, K. V., Freitas, F. O., and Freitag, R. M. K. (2024). Eficiência na leitura: medidas de precisão e velocidade entre alunos do colégio de aplicação da universidade federal de sergipe. *Revista Científica Sigma*, 5(5):120–143.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

de Gois, T. S., Freitas, F. O., Tejada, J., and Freitag, R. M. K. (2024). Nlp and education: Using semantic similarity to evaluate filled gaps in a large-scale cloze test in the classroom. *The Mental Lexicon*, 19(1):90–99.

Freitas, F. O., dos Santos, G. E., and Freitag, R. M. K. (2025). The use of the cloze test in reading comprehension assessment in brazil: post-pandemic challenges. *Cadernos de Linguística*, 6(2):e787–e787.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Kleijn, S., Pander Maat, H., and Sanders, T. (2019). Cloze testing for comprehension assessment: The hytec-cloze. *Language Testing*, 36(4):553–572.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mirault, J., Massol, S., and Grainger, J. (2021). An algorithm for analyzing cloze test results. *Methods in Psychology*, 5:100064.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Real, L., Fonseca, E., and Gonçalo Oliveira, H. (2020). The assin 2 shared task: a quick overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 406–412. Springer.

Santos, G. E. (2025). O preenchimento de lacunas de aspecto verbal em teste cloze: pistas de compreensão em leitura.

Santos, R., Rodrigues, J., Gomes, L., Silva, J., Branco, A., Cardoso, H. L., Osório, T. F., and Leite, B. (2024). Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1):111–122.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.