# AVALIANDO OS AVALIADORES: QUÃO EFICIENTES SÃO OS SISTEMAS DE CORREÇÃO AUTOMÁTICA DE REDAÇÕES?

## *EVALUATING THE EVALUATORS: HOW EFFICIENT ARE AUTOMATED ESSAY SCORING SYSTEMS?*

Ronaldo Teixeira Martins https://orcid.org/0009-0001-4305-0636
Universidade do Distrito Federal
ronaldo.martins@undf.edu.br

Luan Daniel dos Santos Sousa https://orcid.org/0009-0004-3552-8507
Universidade do Distrito Federal
luan.sousa@undf.edu.br

*http://doi.org/ 10.35572/rle.v25i2.6493*

*Recebido em 07de maio de 2025*
*Aceito em 28 de julho de 2025*

**Resumo**: Este artigo avalia a eficácia dos sistemas de correção automática de redações em português do Brasil, comparando suas avaliações com as de professores experientes. Foram analisados três sistemas disponíveis no mercado brasileiro: coRedação, CRIA e Glau, utilizando um corpus de 40 redações de alunos do ensino médio. Os resultados revelaram divergências significativas entre as avaliações dos sistemas e as dos avaliadores humanos, especialmente na avaliação geral das redações. Embora os sistemas mostrem alguma convergência na avaliação isolada das competências do ENEM, o desempenho global apresentou discrepâncias consideráveis. As análises estatísticas, incluindo correlação de Pearson, erro médio absoluto e diferença média quadrática, confirmaram a variabilidade no desempenho dos sistemas. A pesquisa destaca a necessidade de aprimoramentos nos sistemas de correção automática para que possam ser utilizados de forma mais eficaz como ferramentas de apoio ao ensino.

**Palavras-chave**: Correção automática de redações. Avaliação automática de redações.

**Abstract**: This article evaluates the effectiveness of automated essay scoring systems in Brazilian Portuguese, comparing their assessments with those of experienced teachers. Three systems available in the Brazilian market were analyzed: coRedação, CRIA, and Glau, using a corpus of 40 essays from high school students. The results revealed significant discrepancies between the systems' evaluations and those of human evaluators, especially in the overall assessment of the essays. Although the systems show some convergence in the isolated evaluation of ENEM competencies, the overall performance showed considerable differences. Statistical analyses, including Pearson correlation, mean absolute error, and mean squared error, confirmed the variability in the systems' performance. The research highlights the need for improvements in automated scoring systems to be used more effectively as teaching support tools.

**Keywords**: *Automated Essay Automated Essay Evaluation*

## Introdução

O campo da "correção automática de redações" (em inglês: *Automated Essay Scoring*, ou AES) e da "avaliação automática de redações" (*Automated Essay Evaluation*, ou AEE) vem sendo explorado sistematicamente pelo menos desde a década de 1960. (Shermis, Burstein, 2013). Trata-se de subdomínio do processamento automático das línguas naturais (PLN), que explora o uso de inteligência artificial para a automatização dos processos 1) de identificação de problemas e desvios em redações escolares; 2) de atribuição de nota aos textos; e 3) de geração de feedback para os autores. (Rassi, Lopes, 2023).

A principal hipótese subjacente à área é a de que a competência linguística humana envolvida na atividade de correção e avaliação de redações pode ser emulada pela máquina. Essa hipótese já vem sendo testada na avaliação de textos em língua inglesa, como o indicam aplicações como e-rater, (Attali, Burstein, 2006), IntelliMetric[14], PEG, (Page, Ericsson, 2002), Criterion[2], SAGrader[3], Grammarly[4] e Quillbot[5]

No entanto, como as práticas de correção de textos são culturalmente orientadas e dependentes de língua, as possibilidades de transferência tecnológica do inglês para o português são limitadas, e as tentativas de desenvolvimento de sistemas para o português brasileiro são relativamente recentes. Citem-se, entre os estudos disponíveis: Amorim, Veloso, 2017; Fonseca, 2018; Haendchen Filho, 2018; Handchen Filho, 2019; Bittencourt, 2020; Da Silva Jr., 2021; Ferreira Mello, 2022; Marinho, 2022. Além de lidar com os fenômenos característicos da língua portuguesa, os sistemas devem operar também com a variedade de critérios utilizados pelas instituições brasileiras (ENEM, UnB, Fuvest, Unesp, Unicamp etc.) e com os problemas de letramento observados no processo de escolarização no Brasil. (Rassi, Lopes, 2023; Lima, 2023).

Essas especificidades não impediram, porém, o desenvolvimento de sistemas próprios para correção de redações em português. Em levantamento preliminar, reconhecem-se já pelo menos três aplicações no mercado brasileiro:

- coRedação (https://coredacao.com/ )
- CRIA (https://cria.net.br/ )
- Glau (https://www.glau.com.vc/ )

A esses três sistemas soma-se a plataforma Redação SP, lançada pela Secretaria de Educação do Estado de São Paulo em agosto de 2023, mas de uso interno e exclusivo aos professores da rede pública paulista[6].

Todas essas aplicações operam da mesma forma: aceitam, como dado de entrada, apenas textos digitados (isto é, não trabalham com reconhecimento de escrita manual); proveem, para cada texto, cinco notas, segundo os critérios de avaliação por competências utilizados pelo ENEM; localizam, no texto, principalmente problemas de natureza ortográfica e gramatical (competência 1 do ENEM); e fornecem, para as demais competências (tema, texto, coerência, intervenção), um feedback padronizado, com indicações genéricas sobre pontos que poderiam ser aprimorados.

A avaliação dos textos é feita de forma completamente automática, sem

[1] https://www.intellimetric.com/direct/
[2] https://www.ets.org/criterion.html
[3] https://www.sagrader.com/
[4] https://www.grammarly.com/
[5] https://quillbot.com/
[6] https://www1.folha.uol.com.br/educacao/2023/12/escolas-de-sp-comecam-a-corrigir-redacao-com-inteligencia-artificial.shtml

intervenção humana. No entanto, os sistemas não detalham a metodologia utilizada. Infere-se que utilizem algoritmos treinados com ,de máquina, mas não há nenhuma informação sobre os modelos de linguagem empregados. Da mesma forma, embora provejam uma avaliação imediata, nenhum dos sistemas indica o grau de confiabilidade das notas atribuídas.

O presente artigo tem por objetivo reportar um experimento de avaliação desses sistemas a partir de um corpus de referência corrigido por humanos. Os resultados revelam que, apesar de convergirem na avaliação isolada de cada uma das competências do ENEM, os sistemas divergiram substancialmente do desempenho humano na avaliação geral. O achado permite a formulação de inúmeras hipóteses, que serão avaliadas na quarta seção, dedicada à interpretação dos dados. As duas primeiras seções são dedicadas ao detalhamento dos sistemas, à descrição do experimento e à apresentação dos resultados.

## 1. Sistemas de avaliação e correção automática de redações para o português brasileiro

No experimento aqui relatado foram avaliados três sistemas de correção disponíveis no mercado brasileiro: coRedação, CRIA e Glau. Como sistemas protegidos pelo segredo industrial, nenhum deles detalha a metodologia empregada, fazendo apenas referência genérica ao uso de "inteligência artificial". No que se segue, reportam-se as parcas informações técnicas encontradas sobre cada um deles.

### 1.1 coRedação[7]

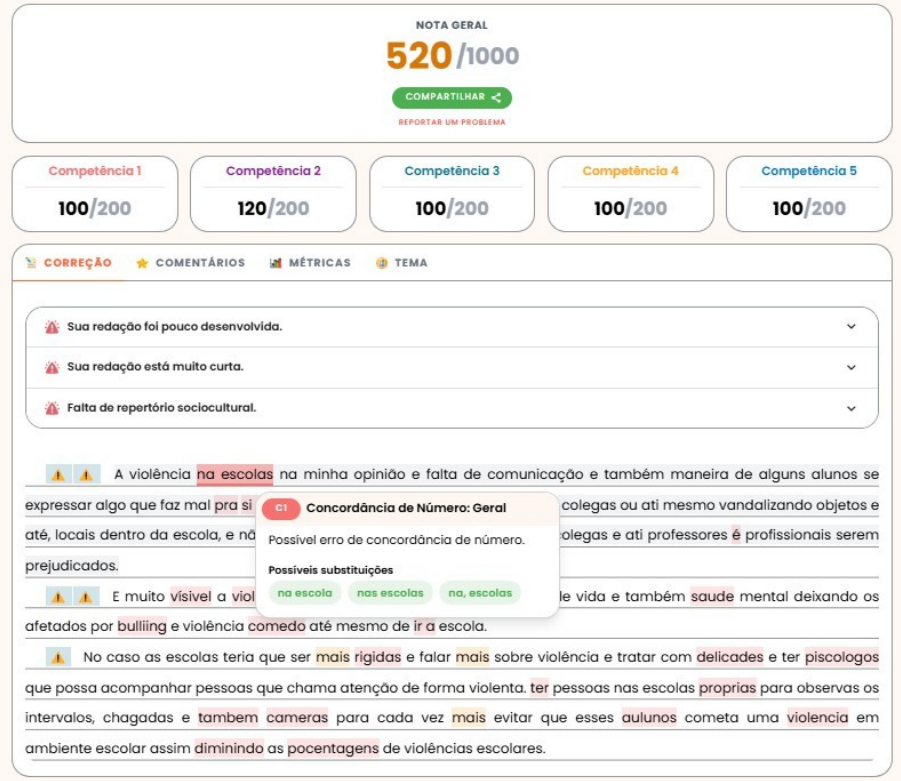
O coRedação é de propriedade da GRQTECH Sistemas de Informação Ltda. O sistema "oferece o serviço de correção de redações segundo as normas gramaticais e ortográficas da norma culta da língua portuguesa; as técnicas de retórica; e as exigências de editais de vestibulares"[8].

O sistema disponibiliza mais de 200 temas de redação a serem escolhidos pelo usuário, que também pode optar pelo "tema livre". Selecionado o tema, o usuário digita o texto e o envia para correção. O texto é avaliado nas cinco competências do ENEM, e os resultados são disponibilizados em três abas: correção, comentários e métricas.

Na aba "correção" (Fig. 1), são apresentados comentários gerais ("sua redação foi pouco desenvolvida", "sua redação está muito curta", etc.), inferidos a partir de estatísticas do texto (número de palavras, número de parágrafos, número de palavras por período etc.). Também são identificados, nessa aba, os problemas, que são assinalados em diferentes cores, de acordo com as competências avaliadas. Os problemas são detalhados quando se passa o mouse sobre os trechos destacados, e são fornecidas "possíveis substituições" para os erros.

[7] https://coredacao.com/
[8] https://coredacao.com/termos-de-uso/

**Figura 1** - Aba "correção" do CoRedação

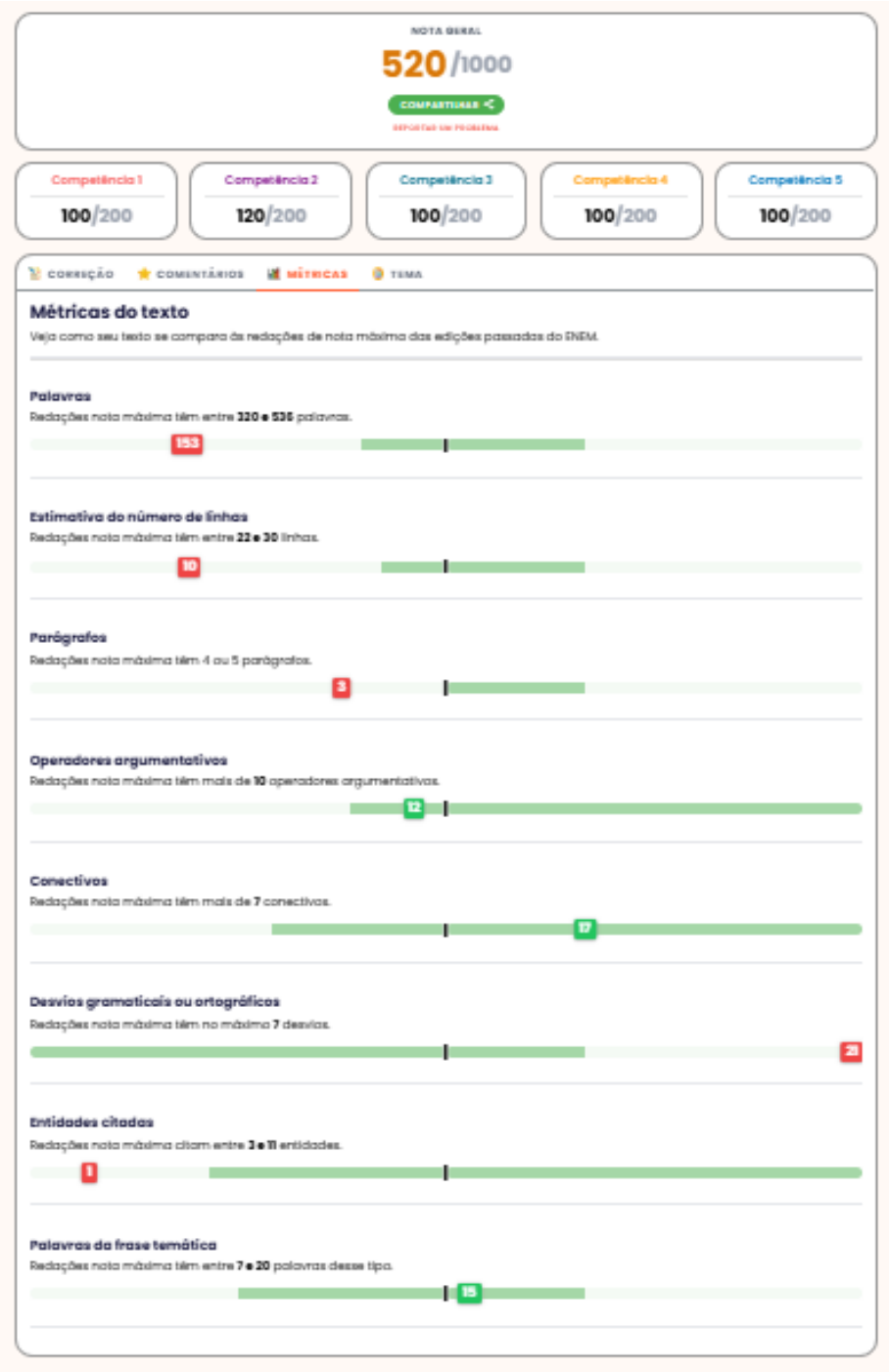



Na aba “comentários” (Fig. 2) é apresentada a avaliação por competência e, ao fim, sugestões de melhorias. Os comentários são genéricos e reproduzem textos padronizados repetidos em diferentes avaliações: “A introdução apresenta o tema, mas de forma confusa e com erros de gramática”, “A coerência do texto é prejudicada pela falta de clareza nas ideias apresentadas”, “A proposta de intervenção está presente, mas é vaga e carece de detalhamento”. Nesta aba não são incorporados dados do texto que possam explicar ou ilustrar a nota alcançada. Algumas palavras-chave (“conectivos”, “argumentos”, etc.) conduzem a material de apoio, em que são apresentadas dicas para o desenvolvimento desses elementos. As sugestões de melhoria são também genéricas: não há exemplos concretos de como aperfeiçoar o texto.

**Figura 2** - Aba "comentários" do CoRedação


A aba “métricas” (Fig. 3) traz uma avaliação das estatísticas do texto em comparação “às redações de nota máxima das edições passadas do ENEM”. Trata-se da única indicação, em todo o site, de que a avaliação toma como parâmetro redações sobre o mesmo tema previamente corrigidas. Não há, porém, nenhuma informação de como esse corpus de referência foi compilado e de quais são suas dimensões. Algumas das métricas comparativas exibidas são o número de palavras, linhas, parágrafos, operadores argumentativos, conectivos, desvios gramaticais ou ortográficos, entidades citadas e “palavras da frase temática”. Embora a metodologia de correção não seja explicitada, pode-se inferir que essas métricas determinam a nota do texto, ou seja, que a nota atribuída constitui, provavelmente, uma função do desvio médio dos indicadores em relação aos observados nos textos de referência. Se confirmada essa hipótese, a pontuação seria estritamente quantitativa, baseada exclusivamente em critérios estatísticos.

**Figura 3** - Aba "métricas" do CoRedação




Outras funcionalidades do sistema estão apenas subsidiariamente relacionadas ao campo da correção automática de redações. O CoRedação prevê a possibilidade de reescrita, fornece “ideias para redação” e disponibiliza cursos e materiais didáticos para aprimorar a escrita de textos.

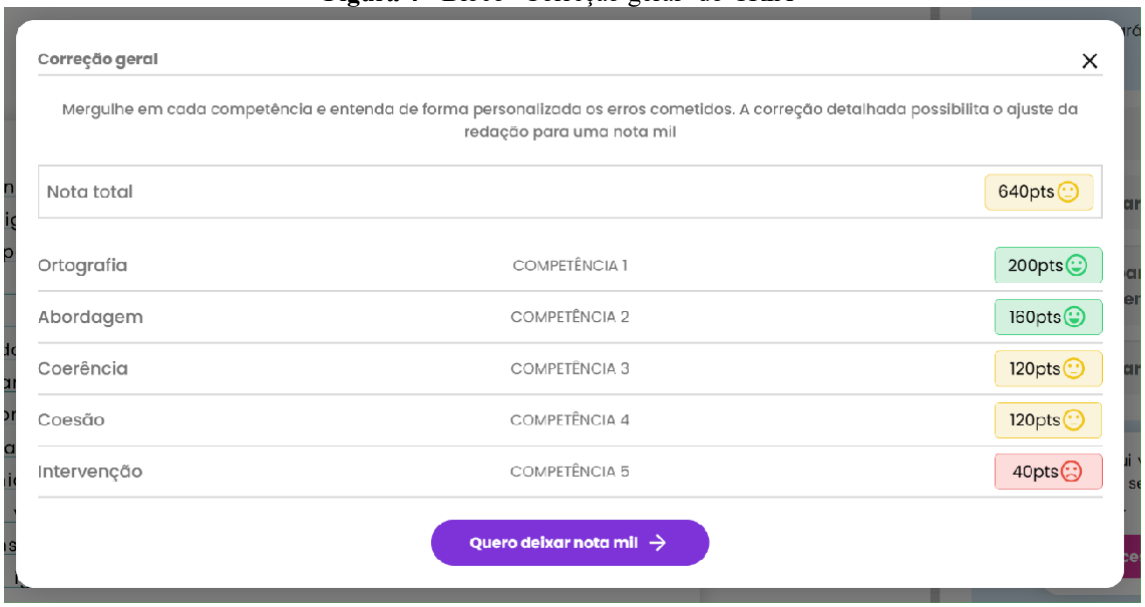
## 1.2. CRIA[9]

Assim como o coRedação, o CRIA também é uma plataforma de correção automática de redações usando inteligência artificial. O sistema foi desenvolvido pela Tecnologia Única Serviços de Informática Ltda. Segundo a descrição no website, "o CRIA é o resultado de anos de pesquisa e desenvolvimento, para entregarmos o algoritmo ideal, que analisa em até 2 minutos e com grande precisão a redação do aluno, baseando-se nas competências do ENEM"[10] .

É possível que o usuário escolha entre mais de 1.128 temas disponíveis dentro de 14 áreas diferentes para criar uma redação. Existe a possibilidade de filtrar a busca por vestibular (ENEM, UNB, Unicamp etc.), por área ou por tema. Ainda na página de escolha do tema, o usuário deve selecionar o tipo de texto, narrativo ou dissertativo, e, dentro de cada tipo de texto, o gênero textual (dissertação argumentativa, resenha crítica, editorial, carta aberta, artigo de opinião, dissertação expositiva) que pretende desenvolver.

Após o usuário escrever e enviar a redação para avaliação, o CRIA exibe a pontuação total e a pontuação por competência do texto. As competências reproduzem, em linhas gerais, os critérios do ENEM: Ortografia, Abordagem, Coerência, Coesão e Intervenção (Fig. 4). É possível obter um relatório detalhado da redação e sugestão de melhorias, mas para isso devem ser usados "CRIA *coins,*", moedas que podem ser obtidas através de compras de pacotes ou assinatura de um dos planos oferecidos.

**Figura 4** - Bloco "Correção geral" do CRIA




No modo detalhado, é possível observar erros e as sugestões presentes em cada uma das cinco competências (Fig. 5). Ao selecionar uma das opções do menu de competências à direita da tela, o usuário pode ver o tipo de erro cometido e detalhes sobre o erro. Erros ortográficos e gramaticais são apontados diretamente no texto, mas a avaliação das demais competências é feita de forma geral, sem a identificação clara de elementos textuais.

[9] https://cria.net.br/
[10] https://cria.net.br/quem-somosutm_sourcesiteutm_mediummenu/

**Figura 5** - Página do relatório detalhado da redação no CRIA


O CRIA também oferece a opção de baixar um "resumão" do relatório em PDF e a opção de corrigir a redação levando em consideração os comentários oferecidos pela plataforma.

### 1.3. Glau[11]

A Glau, assim como o coRedação e o CRIA, é uma plataforma de correção automática de redações usando inteligência artificial. Na seção de perguntas do site é possível encontrar a seguinte informação:

> A Glau é uma plataforma de estudos que usa inteligência artificial para te ajudar a se preparar para o Enem, vestibulares e concursos públicos. Na Glau, você terá acesso a correções detalhadas de redações em 3 segundos, uma base com mais de 160 mil questões do Enem e vestibulares, relatórios personalizados do seu desempenho, raio-x de provas anteriores, e muito mais! [12]

A plataforma afirma que seu desenvolvimento foi supervisionado por um time de professores experientes. Segundo os Termos de Uso, há três pilares essenciais para a correção: "a) normas gramaticais e ortográficas da norma culta da língua portuguesa; b) técnicas de retórica, segundo bibliografia selecionada; c) exigências dos editais de vestibulares ou concursos públicos".

A plataforma permite que o usuário escolha não apenas o tema mas também o critério de correção da redação (ENEM, concursos, vestibulares, ou da própria ferramenta). Os temas apresentados são limitados ao critério escolhido, ou seja, a seleção de temas depende do critério.

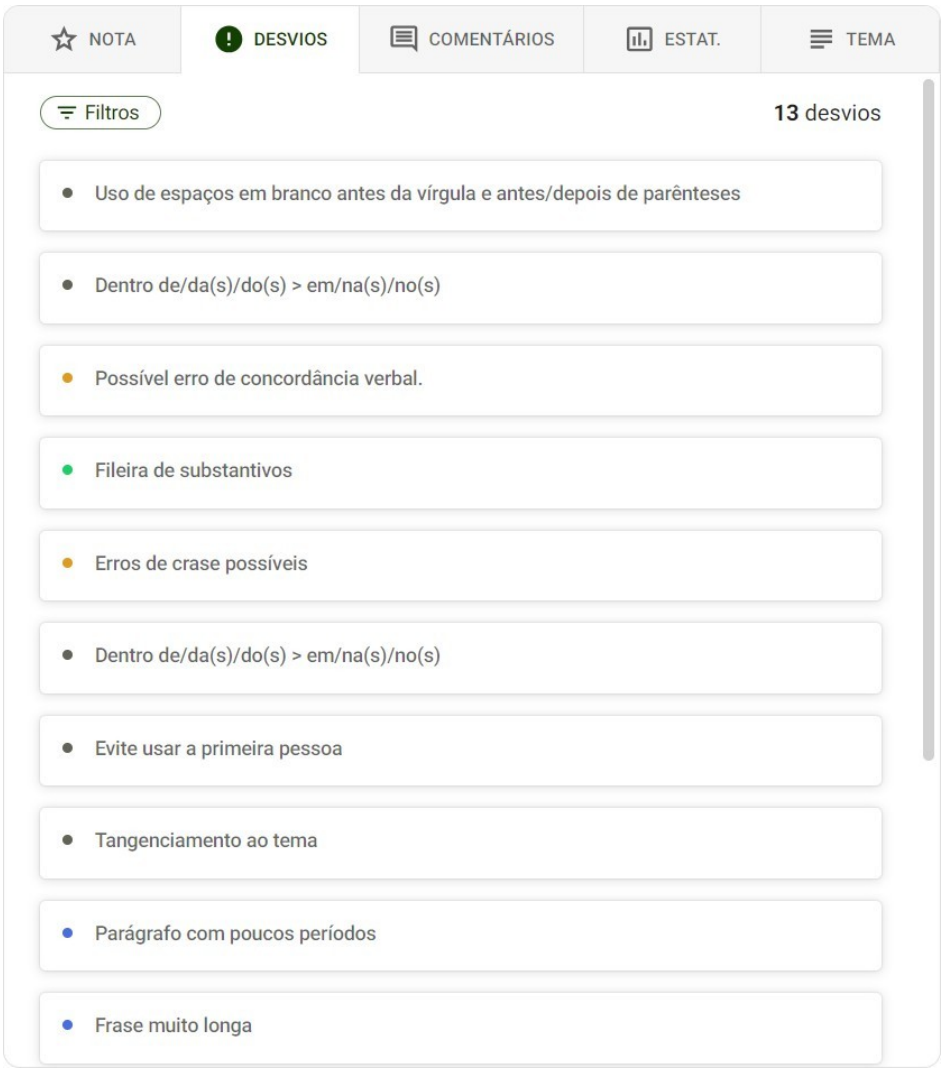
A redação é inserida em uma caixa de texto, e as estatísticas (número de caracteres, de palavras, frases etc.) são fornecidas em tempo real. A correção propriamente dita é exibida em cinco abas: NOTA, DESVIOS, COMENTÁRIOS, ESTAT(ÍSTICAS) e TEMA. Na aba NOTA é informada a nota final da redação e, caso o

[11] https://www.glau.com.vc/
[12] https://www.glau.com.vc/termos-de-uso

usuário seja assinante do plano Glau+, também são informadas as notas por competência.

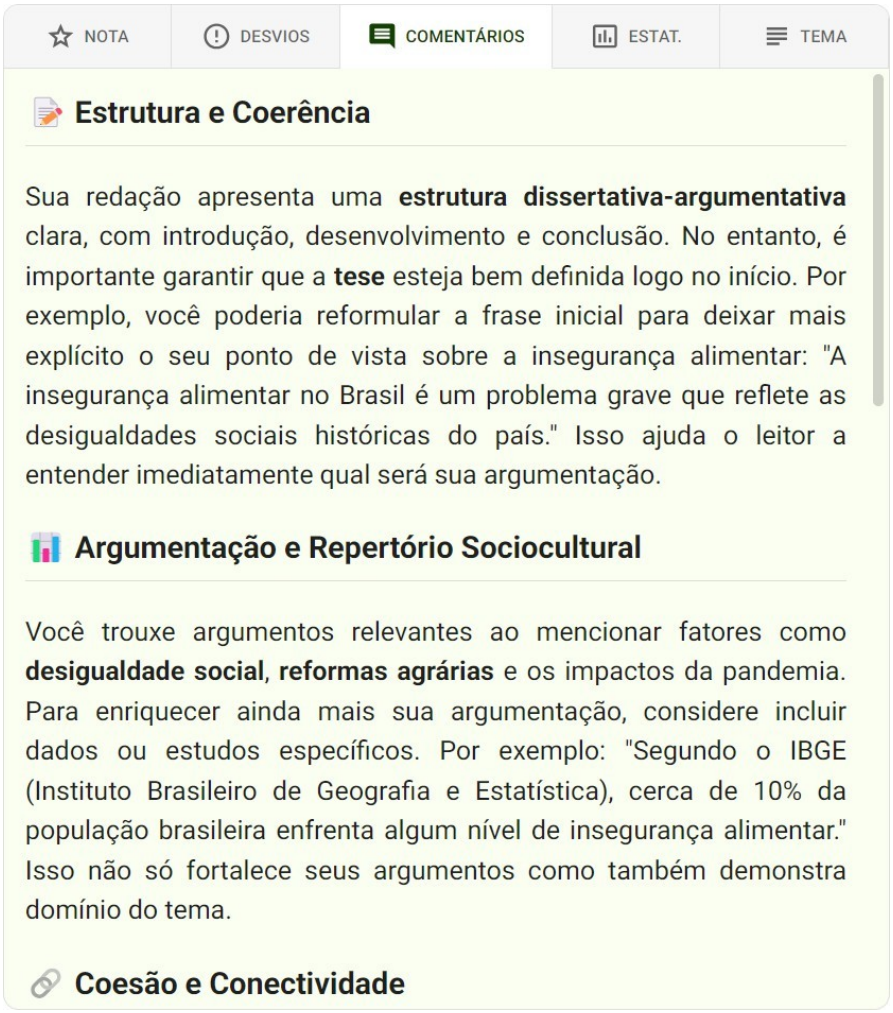
A aba DESVIOS (Fig. 6) explicita erros de ortografia e gramática e oferece sugestões de alteração.

**Figura 6** - Aba DESVIOS na página do resultado da correção




Se o usuário for assinante do plano Glau+, será possível ver comentários sobre a redação divididos em tópicos na aba COMENTÁRIOS (Fig. 7). Os comentários, como nos casos das outras ferramentas, são observações genéricas, não ancoradas em elementos do texto do usuário.

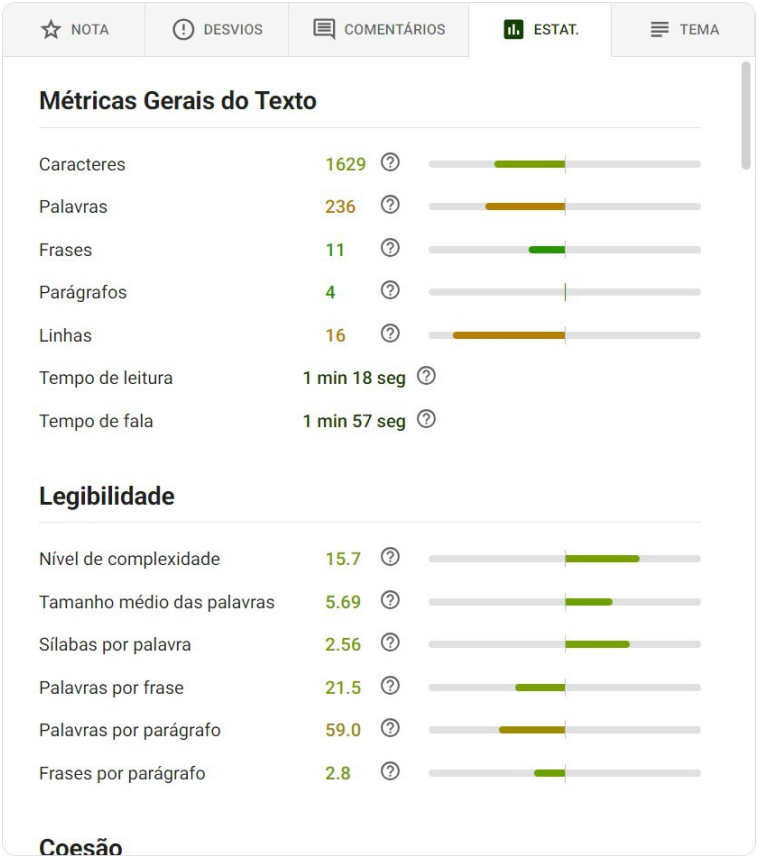
**Figura 7** - Aba COMENTÁRIOS na página do resultado da correção




A aba ESTAT. (Fig. 8) traz dados sobre o texto: métricas gerais, legibilidade, coesão, vocabulário, desvios, adequação ao tema, sinais de pontuação, verbos, substantivos, pronomes, adjetivos, advérbios, conjunções, preposições, entidades nomeadas e outras.

Por fim, a aba TEMA mostra os textos motivadores do tema escolhido pelousuário.

**Figura 8** - Aba "ESTAT." na página do resultado da correção




A Glau também oferece algumas outras funcionalidades, como busca de questões anteriores do ENEM e outros vestibulares, simulados, listas de exercícios, videoaulas, resumos e raios x de provas anteriores. As mais relevantes para o campo de correção automática de redações são, porém, o relatório de desempenho e a ferramenta de argumentos, que fornece modelos de argumentos.

## 2. Descrição do experimento

Para que pudéssemos testar os sistemas, compilamos um corpus de referência composto de 40 redações desenvolvidas por alunos do ensino médio da rede pública do Distrito Federal. As redações foram desenvolvidas em sala de aula, sob a supervisão do professor da disciplina de Língua Portuguesa, e versaram sobre quatro temas de edições passadas do ENEM:

- Medidas para o enfrentamento da recorrência de insegurança alimentar no Brasil (10 textos));
- Desafios para a (re)inserção socioeconômica da população em situação de rua no Brasil (14 textos));
- Desafios para a prevenção da violência na escola (10 textos)); e
- Desafios para o enfrentamento da invisibilidade do trabalho de cuidado realizado pela mulher no Brasil (6 textos)).

Os textos, produzidos a mão, foram digitados e anonimizados. Em seguida, foram avaliados por professores que já fizeram parte das bancas de correção do ENEM

que forneceram, para cada texto, uma nota de referência para cada uma das competências avaliadas pelo ENEM:

- Competência 1: Língua Portuguesa (0 a 200 pontos));
- Competência 2: Abordagem temática e adequação ao tipo textual (0 a 200 pontos));
- Competência 3: Progressão textual e defesa do ponto de vista (0 a 200 pontos));
- Competência 4: Coesão e articulação (0 a 200 pontos));
- Competência 5: Proposta de intervenção (0 a 200 pontos)).

Cada um dos textos foi submetido então aos três sistemas de avaliação automática descritos na segunda seção. Dos 40 textos, 8 não foram pontuados por todos os sistemas porque, segundo eles, seriam demasiadamente curtos. Os outros 32 foram avaliados em cada uma das competências do ENEM. Os resultados comparativos são apresentados nos gráficos de 1 a 6.

**Gráfico 1** - Avaliação comparativa dos resultados para a competência 1 do ENEM (Língua Portuguesa)

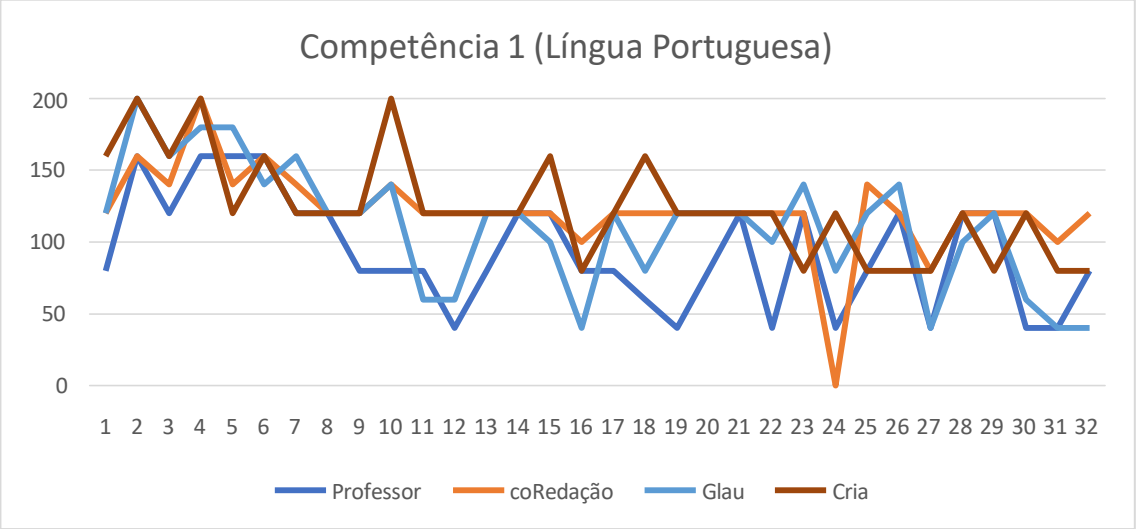



**Gráfico 2** - Avaliação comparativa dos resultados para a competência 2 do ENEM (Tema e Tipo de Texto)

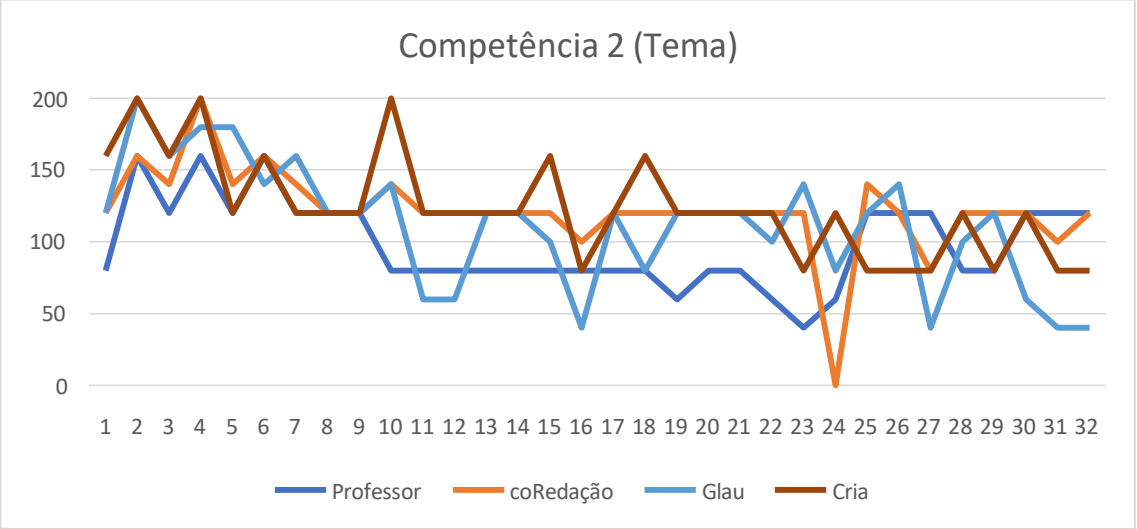


**Gráfico 23** - Avaliação comparativa dos resultados para a competência 3 do ENEM (Progressão Textual)

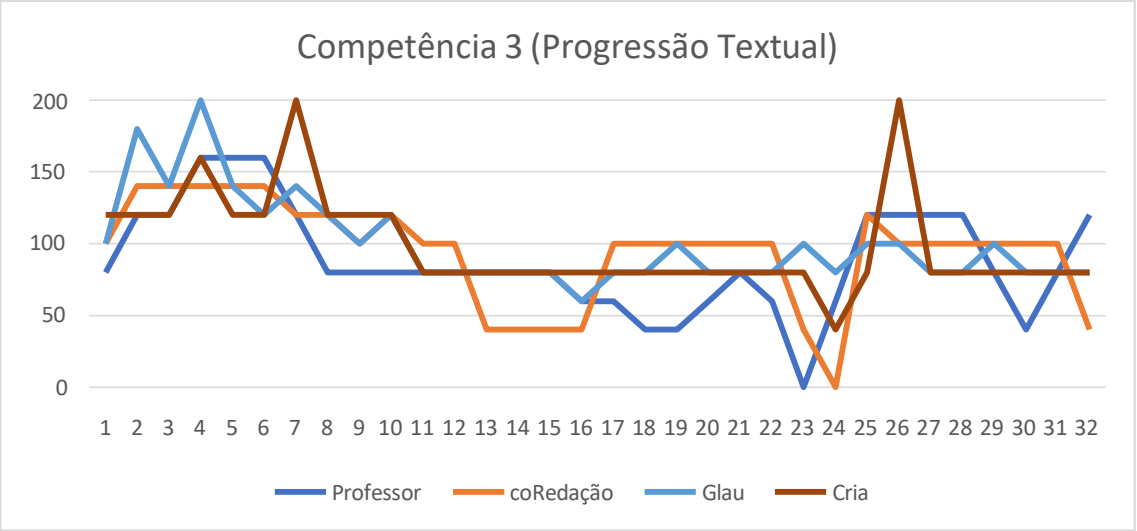



**Gráfico 34** - Avaliação comparativa dos resultados para a competência 4 do ENEM (Coesão e Coerência)

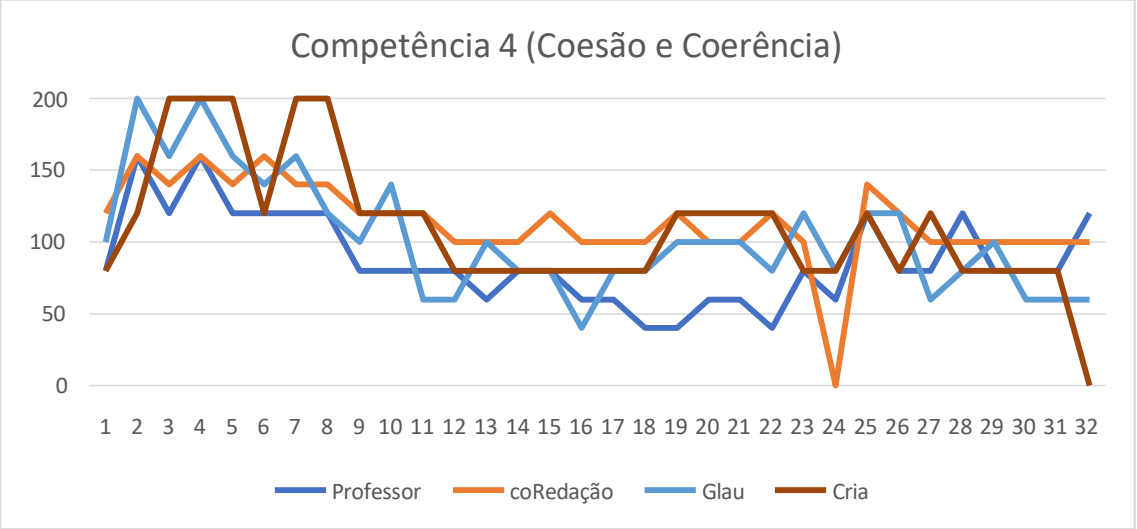



**Gráfico 5** - Avaliação comparativa dos resultados para a competência 5 do ENEM (Proposta de Intervenção)

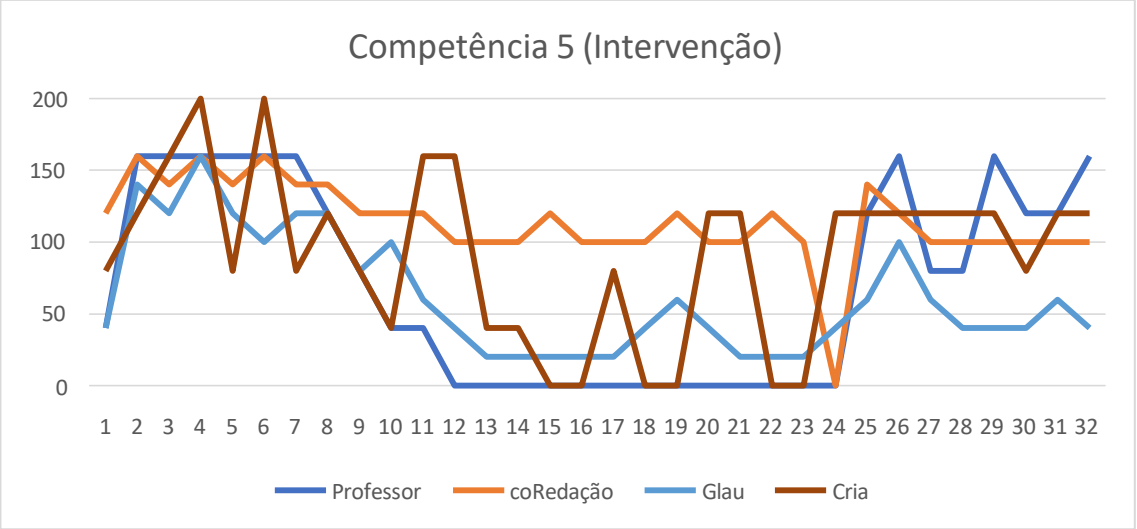
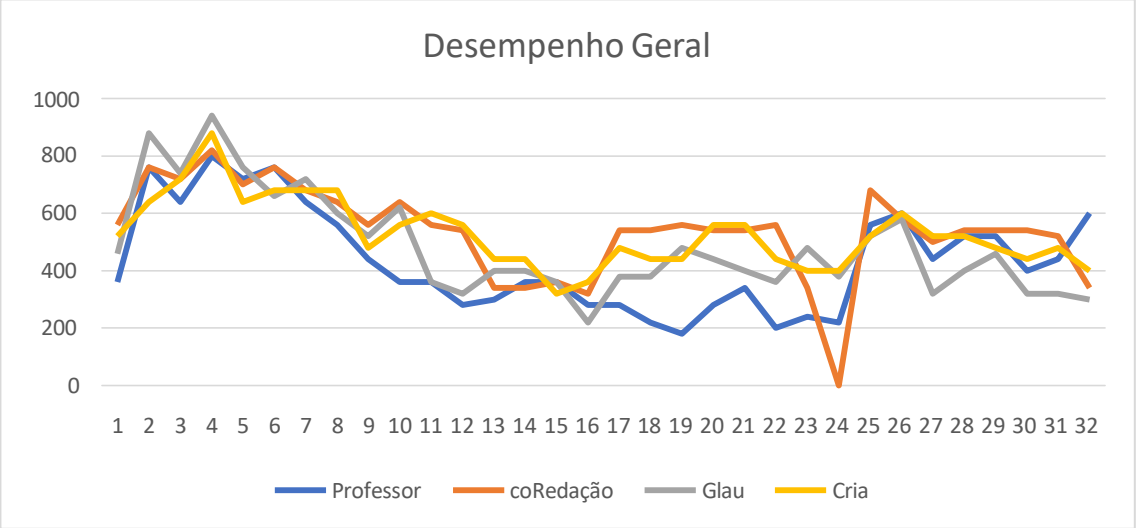


**Gráfico 6** - Avaliação comparativa da pontuação geral do ENEM




Observa-se, nos gráficos acima, que há considerável variação entre os sistemas e entre os sistemas e os avaliadores humanos. Para que se pudesse estimar o grau e a intensidade dessa divergência, investigamos as principais medidas de discrepância, seja para os dados absolutos, seja para os dados agrupados.

A primeira métrica utilizada foi o coeficiente de correlação de Pearson, que foi empregado para medir a força e a direção da relação linear entre as notas atribuídas em cada caso. Trata-se de uma medida estatística que varia entre +1 (correlação linear positiva perfeita, quando duas variáveis aumentam ou diminuem na mesma proporção), 0 (ausência de correlação linear) e -1 (correlação linear negativa perfeita, quando uma variável aumenta e outra diminui na mesma proporção)[13]. Os resultados são apresentados na Tab. 1:

**Tabela 1** - Coeficiente de correlação de Pearson para cada competência e para a nota global das redações, por referência à avaliação humana

| PEARSON (r) | CoRedação | CRIA | Glau |
|---|---|---|---|
| C1 (Língua Portuguesa) | 0,610 | 0,350 | 0,747 |
| C2 (Tema) | 0,520 | 0,266 | 0,312 |
| C3 (Progressão Textual) | 0,516 | 0,530 | 0,575 |
| C4 (Coesão e Coerência) | 0,603 | 0,439 | 0,673 |
| C5 (Intervenção) | 0,537 | 0,583 | 0,739 |
| Nota global | 0,638 | 0,722 | 0,743 |

Os dados permitem perceber que as ferramentas têm desempenho bastante variável. O Glau é o sistema que, em regra, mais se aproxima da avaliação humana, com correlação forte nas competências 1 (Língua Portuguesa) e 5 (Proposta de Intervenção), mas tem correlação desprezível na avaliação da competência 2 (Tema) e apenas moderada em relação à competência 3 (Progressão Textual).

[13] O Coeficiente de correlação de Pearson é calculado pela fórmula:em que x e y são os valores dos dados.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

O resultado global, no entanto, é positivo. O CRIA é a ferramenta de desempenho mais irregular, que varia entre uma correlação fraca para as competências 1 e 2 e uma correlação moderada nos demais casos. No entanto, e apesar da baixa correlação por competência, é o sistema que tem o segundo melhor desempenho na avaliação global, cuja correlação com as notas atribuídas pelos avaliadores humanos também é forte. O coRedação é o mais estável: há pouca variação de correlação entre as competências, que sempre se situam acima do patamar de 0,5; o resultado global, porém, é o mais distante da avaliação humana.

A correlação de Pearson é confirmada pelo cálculo do erro médio absoluto[14], que mede a diferença dos valores absolutos entre as notas previstas e as observadas, e cujos resultados são apresentados na Tab. 2. Nesse levantamento, quanto maior o erro médio, tanto maior a diferença entre as notas atribuídas pelos sistemas automáticos e os avaliadores humanos.

**Tabela 2** - Erro médio absoluto (MAE) entre as notas atribuídas a cada competência e para a nota global das redações, por referência à avaliação humana

| MAE | CoRedação | CRIA | Glau |
|---|---|---|---|
| C1 (Língua Portuguesa) | 32,50 | 40,63 | 28,13 |
| C2 (Tema) | 31,25 | 38,13 | 38,13 |
| C3 (Progressão Textual) | 31,25 | 26,88 | 25,63 |
| C4 (Coesão e Coerência) | 32,50 | 35,00 | 29,38 |
| C5 (Intervenção) | 58,75 | 42,50 | 38,13 |
| Nota global | 131,88 | 126,88 | 109,38 |

Na tabela acima verifica-se, mais uma vez, que o Glau é, em geral, a ferramenta cujas notas são mais próximas do avaliador humano, ou seja, aquela em que o erro médio é menor. O CRIA ocupa, novamente, o segundo lugar, embora tenha registrado discrepâncias maiores do que as outras ferramentas em relação a três competências. O coRedação, que tem resultados parciais melhores do que o CRIA, é prejudicado pela avaliação da competência 5, em que se destaca negativamente entre todos os sistemas, registrando o erro médio mais alto de toda a série.

Para eliminar a possibilidade de que o erro médio tenha sido afetado por uma tendência natural à atribuição de notas medianas aos textos, calculamos também a diferença média quadrática (MSE), que amplifica os desvios em relação à referência humana, especialmente os grandes, devido ao uso do quadrado[1521]. Isso faz com que o MSE seja mais sensível a valores extremos, e permite a identificação de um comportamento apenas medianamente aceitável. Os resultados são apresentados na Tab. 3:

[14] O erro médio absoluto é calculado pela fórmula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

[15] A diferença média quadrática é calculada pela fórmula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

**Tabela 3** - Diferença média quadrática (MSE) entre as notas atribuídas a cada competência e para a nota global das redações, por referência à avaliação humana

| MSE | CoRedação | CRIA | Glau |
|---|---|---|---|
| C1 (Língua Portuguesa) | 42,426 | 52,082 | 34,095 |
| C2 (Tema) | 37,749 | 47,302 | 45,415 |
| C3 (Progressão Textual) | 36,056 | 36,228 | 33,727 |
| C4 (Coesão e Coerência) | 37,417 | 48,218 | 33,727 |
| C5 (Intervenção) | 72,111 | 61,644 | 48,088 |
| Nota global | 175,178 | 152,520 | 133,838 |

Os dados da Tab. 3 confirmam a tipicidade dos erros da Tab. 2 e afastam a possibilidade de que o ranking de desempenho pudesse ter sido afetado por coincidências estatísticas derivadas do cálculo da média aritmética simples. Percebe-se que os sistemas mantiveram os resultados mesmo com a sobrevalorização dos desvios mais pronunciados. Ou seja, não houve, aparentemente, nenhum texto isolado que pudesse ter afetado o desempenho das ferramentas e prejudicado ou beneficiado sua colocação.

Por fim, para a obtenção de métricas relacionadas a problemas de classificação, em que os dados são categóricos, operamos à discretização do desempenho das ferramentas, observando as regras do ENEM, segundo o qual são aceitáveis desvios por competência de até 80 pontos e desvios por avaliação global de até 100 pontos[16]. Os resultados da discretização são apresentados na Tab. 4.

**Tabela 4** – Número de resultados aceitáveis (A) e inaceitáveis (I) em relação a cada competência e à nota global das 32 redações, por referência à avaliação humana[17]

| A/I | CoRedação | CRIA | Glau |
|---|---|---|---|
| C1 (Língua Portuguesa) | 28/4 | 24/8 | 31/1 |
| C2 (Tema) | 31/1 | 28/4 | 28/4 |
| C3 (Progressão Textual) | 31/1 | 29/3 | 31/1 |
| C4 (Coesão e Coerência) | 30/2 | 25/7 | 32/0 |
| C5 (Intervenção) | 17/15 | 24/8 | 29/3 |
| Nota global | 16/16 | 16/16 | 14/18 |

A Tab. 4 promove um interessante reordenamento dos resultados anteriores. A Glau, que até então figurava como o sistema mais próximo da avaliação humana, passa a ocupar a terceira e última posição se considerada a necessidade de terceira correção. Dos 32 textos avaliados pela ferramenta, 18 textos (56%) teriam de ser submetidos a nova avaliação, segundo as regras do ENEM. O desempenho dos outros dois sistemas (coRedação e Cria) é apenas um pouco melhor: 50% dos textos avaliados por essas ferramentas teriam de passar por terceira correção. Em quaisquer casos, é considerável o número de textos que seriam submetidos a nova avaliação e, portanto, verifica-se ainda

[16] Segundo as regras do ENEM, são submetidos a terceira correção textos em que há discrepância maior de 100 pontos na soma total das competências ou em que há discrepância maior de 80 pontos em uma ou mais competências

[17] Na Tab. 4, o primeiro número em cada célula corresponde ao número de textos que não teriam de passar por terceira correção (ou seja, cujo resultado seria aceitável); e o segundo, ao número de textos que precisariam passar por terceira correção (ou seja, cujo resultado seria inaceitável). Em cada célula, a soma corresponde ao total de textos avaliados (32).

limitado o alcance dessas ferramentas como substitutos de avaliadores humanos, pelo menos no âmbito do ENEM (em que a média de discrepância entre humanos é de cerca de 20%)(Fossey, 2018).

Os dados também revelam que os problemas não estariam, propriamente, na avaliação isolada das competências, mas no acúmulo de notas individualmente pouco discrepantes que, ao fim, contribuiriam para uma discrepância significativa na nota global. O comportamento das ferramentas, no entanto, não é uniforme: o CRIA revela o pior desempenho na avaliação isolada, mas as diferenças parecem ser compensadas na avaliação global; o oposto ocorre com o coRedação e o Glau, que têm desempenhos bons na avaliação isolada, mas desempenhos medianos na avaliação global.

A partir dos resultados discretizados, pudemos calcular o coeficiente kappa de Cohen, que avalia a concordância entre dois avaliadores para além do acaso[18]. Os dados são apresentados na Tab. 5 abaixo, onde $\kappa = 1$ indica concordância plena e $\kappa = 0$ indica ausência de concordância.

**Tabela 5** - Coeficiente kappa de Cohen entre o desempenho discretizado das ferramentas por referência à avaliação humana

| Kappa (κ) | CoRedação | CRIA | Glau |
|---|---|---|---|
| C1 (Língua Portuguesa) | 0,88 | 0,75 | 0,97 |
| C2 (Tema) | 0,97 | 0,88 | 0,88 |
| C3 (Progressão Textual) | 0,97 | 0,91 | 0,97 |
| C4 (Coesão e Coerência) | 0,94 | 0,78 | 1,00 |
| C5 (Intervenção) | 0,53 | 0,75 | 0,91 |
| Nota global | 0,50 | 0,50 | 0,44 |

Confirma-se, pela Tab. 5, que a concordância é forte entre os sistemas automáticos e a avaliação humana quando consideradas as competências isoladamente. Destaca-se, neste caso, especialmente a competência 3 (Progressão Textual), em que as discrepâncias são mínimas para as três ferramentas (ou seja, pouquíssimos seriam os textos encaminhados para terceira correção). No entanto, a concordância parcial não se reproduz na concordância global, em que todos os sistemas revelaram apenas desempenho moderado.

A última métrica utilizada para a comparação foi a Medida F (*F-Score*), que combina precisão e revocação (*recall*) em uma única métrica harmônica, amplamente empregada em problemas de classificação para avaliar o desempenho de um modelo[19]. A precisão mede a proporção de predições positivas que estão corretas, ou seja, o número de textos que receberam avaliação aceitável e não seriam encaminhados para terceira correção. A revocação mede a proporção de instâncias positivas reais que foram corretamente identificadas, ou seja, o número de textos que foram corrigidos. Para o cálculo da revocação utilizamos todos os 40 textos presentes originariamente no corpus, incluindo aqueles que, embora avaliados por humanos, não receberam nota dos sistemas porque seriam exageradamente curtos. Nessa medida, destaca-se o CRIA, que deixou de

[18] O coeficiente kappa é calculado pela fórmula

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

[19] A Medida F é calcula pela fórmula:

$$F_1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

corrigir apenas 3 textos, contra 8 textos nos dois outros casos. As medidas são apresentadas na Tab. 6

**Tabela 6** – Medida F1 para o desempenho discretizado das ferramentas por referência à avaliação humana

| F1 | CoRedação | CRIA | Glau |
|---|---|---|---|
| C1 (Língua Portuguesa) | 0,778 | 0,737 | 0,861 |
| C2 (Tema) | 0,861 | 0,849 | 0,778 |
| C3 (Progressão Textual) | 0,861 | 0,877 | 0,861 |
| C4 (Coesão e Coerência) | 0,833 | 0,765 | 0,889 |
| C5 (Intervenção) | 0,472 | 0,737 | 0,806 |
| Nota global | 0,444 | 0,512 | 0,389 |

A medida-F indica que, em relação às competências isoladas, há, em geral, bom equilíbrio entre precisão e revocação, à exceção da avaliação da competência 5 pelo coRedação. No entanto, esses indicadores isolados não contribuem para um bom desempenho geral dos modelos, que se revelam bastante aquém das expectativas. Nesse contexto, torna-se particularmente emblemática a situação da ferramenta CRIA, que figura aqui como a de melhor desempenho, embora seja a que revela, em geral, os piores indicadores quando avaliadas as competências isoladamente. O dado parece indicar que não há relação direta e óbvia entre as avaliações parciais e a avaliação global, o que contradiz, em larga medida, a expectativa de que a avaliação global pudesse ser o corolário direto das avaliações parciais. A explicação mais evidente para essa discrepância é a de que, pelo menos nos casos avaliados, a concordância com a nota global atribuída pelo avaliador humano foi produto da distribuição compensatória entre avaliações parciais, ou seja, muitos dos textos que receberam a mesma nota o fizeram por força do acaso, mais do que por mérito dos sistemas, como já o indicavam, aliás, os valores baixos para o coeficiente de kappa apresentados na Tab 5.

## Conclusões

A correção de redações é uma atividade escolar extremamente laboriosa e cara. Segundo Bittencourt (2020, p. 19), um professor leva, em média, 12 minutos para atribuir nota a uma redação do ENEM. Se acrescentarmos também o tempo necessário para a identificação dos problemas e o fornecimento de feedback individualizado para o estudante, percebe-se que a correção de redações é uma prática não escalável e pouco adaptada à realidade dos sistemas de ensino, que contam com salas de aulas com muitos alunos e professores que atuam em diferentes turmas. Por essa razão, os alunos têm tido poucas oportunidades de produção de textos na educação básica, principalmente na rede pública, situação que termina por dificultar o desenvolvimento das habilidades de escrita esperadas da formação escolar.

O emprego de sistemas automáticos de correção de redações como ferramentas auxiliares de ensino – asseguradas sua eficiência, precisão e credibilidade – poderia contribuir positivamente para a transformação desse cenário, estimulando a produção mais intensiva e mais frequente de textos sem sobrecarga docente e com resultados mais rápidos. Mas os resultados aqui expostos permitem perceber que há ainda um caminho considerável a ser percorrido antes que os sistemas hoje disponíveis possam desempenhar esse papel.

Este texto se debruçou sobre uma avaliação estritamente estatística e quantitativa do desempenho dos sistemas de correção automática disponíveis para o português do Brasil. Uma segunda etapa da avaliação deverá se deter sobre as diferenças qualitativas no processo de correção, de forma que possamos identificar o número de falsos negativos, ou seja, de desvios do texto que não foram identificados; e o número dos falsos positivos, ou seja, o grau de sobrecorreção do texto, com a identificação de desvios inexistentes. Só assim poderemos efetivamente avaliar o alcance das ferramentas ora utilizadas para corrigir e avaliar textos.

Os resultados alcançados, porém, permitem já uma primeira aproximação sobre a precisão e a abrangência dos sistemas disponíveis, e a constituição de um primeiro *benchmarking* das ferramentas de correção e avaliação automática de redações, ainda indisponível para o português brasileiro. No entanto, para que possamos levar essa proposta a cabo seria necessário confirmar os atuais valores para *corpora* ampliados, que incluíssem um número mais expressivo de temas e de redações escolares. Por isso, consideramos temerária a generalização dos resultados aqui alcançados e não defenderemos que um ou outro sistema seja, visivelmente, melhor do que os demais. Pelo contrário: os resultados parecem indicar que a qualidade dos sistemas varia muito em função das competências e do tipo de análise (regressiva ou classificatória) que se proceda a partir dos dados obtidos.

De qualquer forma, os resultados deste estudo indicam que os sistemas avaliados ainda não atingiram um nível de precisão e confiabilidade que permita sua utilização como substitutos dos avaliadores humanos, principalmente no contexto do ENEM. Há uma necessidade de mais pesquisas e desenvolvimento para aprimorar a capacidade desses sistemas de avaliar a qualidade geral dos textos, além de identificar e corrigir desvios linguísticos de forma eficaz.

**Referências**

AMORIM, E.; VELOSO, A. A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Anais...Valencia, Spain: Association for Computational Linguistics, abr. 2017.

BITTENCOURT JR., J. A. S. *Avaliação automática de redação em língua portuguesa empregando redes neurais profundas*. Universidade Federal de Goiás, 2020.

DA SILVA JR., J. A. *Um avaliador automático de redações*. Universidade Federal do Espírito Santo, 2021.

FERREIRA MELLO, R. et al. Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in Portuguese. (A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds.) LAK22 Conference Proceedings. Anais...United States of America: Association for Computing Machinery (ACM), 2022.

FONSECA, E. R. et al. Automatically Grading Brazilian Student Essays. (A. Villavicencio et al., Eds.) *Computational Processing of the Portuguese Language*. Anais...Springer International Publishing, 2018.

FOSSEY, Marcela Franco. Avaliação de redações de vestibular: da teoria à prática. *Trabalhos em Linguística Aplicada*, 57 (2), mai-ago 2018. Disponível em: https://doi.org/10.1590/01031813865218137742l. Acesso em 29 Mai 2024.

HAENDCHEN FILHO, A. et al. An approach to evaluate adherence to the theme and the argumentative structure of essays. International Conference on KnowledgeBased Intelligent Information & Engineering Systems. Anais...2018.

HAENDCHEN FILHO, A. et al. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. Procedia Computer Science, v. 159, p. 764–773, jan. 2019.

LIMA, T. B. DE et al. *Avaliação Automática de Redação*: Uma revisão sistemática. Revista Brasileira de Informática na Educação, v. 31, p. 205--221, maio 2023.

MARINHO, J. et al. Automated Essay Scoring: An approach based on ENEM competencies. Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. Anais...SBC, 2022.

MARINHO, J.; ANCHIÊTA, R.; MOURA, R. *Essay-BR*: a Brazilian Corpus to Automatic Essay Scoring Task. Journal of Information and Data Management, v. 13, n. 1, p. 65–76, 2022.

MARTINS, A. B.; SOUZA, R. C. (Eds.). *Avaliação de redações*: abordagens teóricas e práticas. Editora XYZ, 2015.

PAGE, Erica B.; ERICSSON, Patricia Freitag. "The development of the Project Essay Grade (PEG) system." *Assessing Writing* 8.1 (2002): 39-58.

RASSI, Amanda Pontes; LOPES, Priscilla de Abreu. Correção automática de redação. In CASELI, Helena; NUNES, Maria da Graça Volpe. (ed). *Processamento de linguagem natural*: conceitos, técnicas e aplicações em português. BPLN, 2024. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-aes/cap-aes.html. Acesso em: 29 Abr 2024.

SHERMIS, M. D.; BURSTEIN, J. *Handbook of Automated Essay Evaluation*: Current Applications and New Directions. [s.l.] Routledge/Taylor & Francis Group, 2013.

SILVA, C. M. *Correção de redações*: teoria e prática. Editora ABC, 2018.