Desafios de grandes modelos de linguagem generativa na reprodução de complexidade textual: um estudo com editoriais jornalísticos

Challenges for large generative language models in reproducing textual complexity: a study on newspaper editorials

André Luis Antonelli *[1]

[1] *Universidade Estadual de Maringá, Departamento de Língua Portuguesa, Maringá, PR, Brasil.*

Resumo

Este artigo avalia a *performance* do modelo de linguagem generativa *Sabiá-3* na tarefa de reproduzir aspectos de complexidade textual do gênero discursivo editorial, usando como ponto de referência editoriais produzidos por humanos. Para essa tarefa, utilizamos métricas da ferramenta computacional NILC-Metrix. Nossos resultados revelaram diferenças em quatro das cinco métricas analisadas. Os textos humanos demonstraram maior complexidade nas medidas "proporção de types em relação à quantidade de tokens" e "entropia cruzada". Argumentamos que esse resultado pode estar vinculado, por exemplo, à capacidade humana de selecionar palavras ou realizar combinações lexicais sem a limitação de parâmetros probabilísticos. Já os textos gerados pelo modelo *Sabiá-3* apresentaram maior complexidade nas métricas "sílabas por palavra" e "orações subordinadas", possivelmente devido ao fato, entre outros aspectos, de que ferramentas do tipo não sofrem restrições de processamento cognitivo. A única métrica sem diferença estatística significativa foi "conjunções difíceis". Atribuímos esse resultado à natureza fechada dessa classe gramatical, que limitaria variações. O estudo reforça a importância de se considerar múltiplas dimensões da complexidade textual ao avaliar a produção de grandes modelos de linguagem generativa, especialmente quando se trata de gêneros que exigem domínio linguístico refinado, tais como o editorial.

Palavras-chave: Modelos de linguagem generativa. Complexidade textual. Gêneros discursivos. Inteligência artificial e linguagem. Análise comparativa humano-IA.

Abstract

This article evaluates the performance of the generative language model *Sabiá-3* in the task of reproducing aspects of textual complexity characteristic of the editorial discourse genre, using human-produced editorials as a reference point. For this task, we employed metrics from the computational tool NILC-Metrix. Our results revealed differences in four out of the five analyzed metrics. Human texts showed greater complexity in the measures of "type-token ratio" and "cross-entropy". We argue that this outcome may be linked, for instance, to the human ability to select words or form lexical combinations without the constraints of probabilistic parameters. In contrast, the texts generated by the *Sabiá-3* model displayed higher complexity in the "syllables per word" and "subordinate clauses" metrics, possibly due, among other factors, to the fact that such tools do not face cognitive processing limitations. The only metric without a statistically significant difference was "hard conjunctions". We attribute this result to the closed-class nature of this grammatical category, which tends to limit variation. This study reinforces the importance of considering multiple dimensions of textual complexity when evaluating the output of large generative language models, especially in genres that require refined linguistic control, such as the editorial.

Keywords: Generative language models. Textual complexity. Discourse genres. Artificial intelligence and language. Human-AI comparative analysis.

Textoline
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2025.58530

Seção:
Artigos

Autor Correspondente:
André Luis Antonelli

Editor de seção:
Danierveli Pereira
Editor de layout:
Leonado Araújo

Recebido em:
8 de abril de 2025
Aceito em:
12 de maio de 2025
Publicado em:
31 de maio de 2025

1 Introdução

Grandes modelos de linguagem generativa (LLMs, na sigla em inglês) têm desempenhado um papel crescente na produção automatizada de textos (Reisenbichler *et al.*, 2022; Franceschelli; Musolesi, 2024; Naik; Naik; Naik, 2024), sendo empregados em uma ampla variedade de aplicações, desde a

*Email: alantonelli@uem.br

redação de conteúdos jornalísticos até a assistência na escrita acadêmica e técnica. Deixando de lado a questão ética, que envolve discussões, entre outros aspectos, sobre autoria, autenticidade e impactos no mercado de trabalho, o fato é que essas ferramentas, treinadas a partir de enormes *corpora*, têm demonstrado notável habilidade na geração de textos em diversos domínios. No entanto, apesar de tais avanços, muitos pontos ainda precisam ser melhor explorados. Um deles diz respeito à capacidade de tais modelos de mimetizar com precisão aspectos relacionados à complexidade textual de gêneros discursivos específicos.

Como já bem estabelecido na literatura especializada, cada gênero do discurso apresenta convenções próprias, que moldam a organização da informação, a escolha lexical, a estrutura sintática e os recursos argumentativos empregados na construção dos enunciados linguísticos (Bakhtin, 1997). A combinação dessas propriedades impacta naquilo que podemos chamar de nível de complexidade de um texto. De acordo com Frantz, Starr e Bailey (2015), complexidade textual refere-se ao grau de sofisticação e desafio que um texto coloca ao leitor, podendo ser analisada com base em fatores linguísticos, cognitivos e discursivos. Ela é frequentemente avaliada por meio da diversidade lexical, do grau de estruturação sintática, da coesão e de medidas de leiturabilidade. Essa natureza multifacetada da complexidade textual é um elemento importante na diferenciação dos gêneros discursivos entre si. Por exemplo, em termos de tipologia textual, uma charge pode ser tão opinativa quanto um editorial de jornal, ainda que sejam gêneros do discurso distintos. No entanto, não há dúvida de que estamos diante de manifestações linguísticas que mobilizam diferentes níveis de complexidade, seja na forma de estrutura organizacional, seja na forma de disposição dos argumentos ou mesmo no grau de diversidade lexical, para nomear apenas algumas distinções. Isso mostra que a relação entre gêneros discursivos e complexidade textual é intrínseca, pois cada gênero, ao se adaptar às convenções que lhe são impostas, apresenta níveis de complexidade específicos. Assim, quando pensamos em produção automatizada de textos por LLMs, a questão que se coloca é saber não apenas o nível de proficiência dessas ferramentas em produzir conteúdo que seja coerente e adequado às normas linguísticas, mas também a capacidade de se adaptar ao nível de complexidade textual exigida pelo gênero do discurso a ser gerado.

Como contribuição a esse debate, neste artigo avaliaremos a *performance* de um modelo de linguagem generativa na tarefa de reproduzir aspectos de complexidade textual do gênero discursivo editorial, usando como ponto de referência editoriais produzidos por humanos. Para essa tarefa, empregaremos uma série de métricas quantitativas de complexidade linguística, o que possibilitará uma avaliação multidimensional do fenômeno. Esperamos, com isso, fornecer uma análise objetiva das capacidades e limitações de LLMs na reprodução de propriedades de complexidade textual intrínsecas à configuração discursiva de qualquer produção verbal.

O artigo está organizado da seguinte maneira. A Seção 2 apresenta a fundamentação teórica a ser explorada no trabalho. Na Seção 3, detalhamos os procedimentos metodológicos adotados. A Seção 4 é dedicada à apresentação dos resultados, que serão analisados na Seção 5. Por fim, na Seção 6, apresentamos as considerações finais.

2 Fundamentação teórica

2.1 Gêneros textuais e o editorial jornalístico

Os gêneros textuais são categorias que agrupam textos com características linguísticas, discursivas e funcionais semelhantes, moldadas por contextos sociocomunicativos específicos (Bakhtin, 1997). Eles surgem a partir das necessidades de comunicação em diferentes esferas da atividade humana. Cada gênero é caracterizado por uma estrutura e uma função social próprias, que orientam tanto a produção quanto a interpretação dos textos.

Dentro da esfera de gêneros jornalísticos que circulam socialmente, o editorial é um exemplo emblemático de texto opinativo, cuja função principal é expressar a posição institucional de um veículo de comunicação sobre temas de relevância pública. Os editoriais são marcados por uma estrutura clássica que inclui uma introdução, na qual o tema é contextualizado; um desenvolvimento, que apresenta argumentos fundamentando a tese a ser defendida; e uma conclusão, que reforça a posição assumida no texto (Vieira, 2009). Linguisticamente, esse gênero se destaca pelo emprego de

um estilo formal, que se caracteriza pelo uso de vocabulário e estruturas gramaticais típicas da norma urbana culta. Além disso, os editoriais se caracterizam pelo uso de estratégias discursivas que buscam persuadir o leitor, como o emprego de evidências estatísticas, referências a autoridades e apelos emocionais. Essas características fazem desse gênero textual uma estrutura sofisticada, que exige do autor não apenas domínio linguístico, mas também profundidade argumentativa e sensibilidade ao contexto sociopolítico.

A complexidade dos editoriais também se reflete em sua função social. Como textos opinativos, desempenham um papel crucial na formação da opinião pública, procurando influenciar a maneira como os leitores percebem e interpretam eventos e questões relevantes do cotidiano. Essa função exige que os editoriais sejam ao mesmo tempo claros e persuasivos, equilibrando a necessidade de transmitir informações de maneira acessível com a de apresentar argumentos convincentes e bem fundamentados. Esse conjunto de propriedades de natureza gramatical, argumentativa e social tornam o editorial um gênero particularmente interessante para análises comparativas entre humanos e inteligência artificial (IA), uma vez que estamos diante de uma manifestação linguística com diferentes camadas, todas elas contribuindo para o grau de complexidade textual do material produzido.

2.2 Modelos de linguagem generativa

Grandes modelos de linguagem generativa, como, por exemplo, os da família GPT (*Generative Pre-trained Transformer*) e Llama, têm reconfigurado o campo de estudos sobre processamento de linguagem natural. Esses modelos são baseados na arquitetura *Transformer*, proposta por Vaswani *et al.* (2018), que utiliza mecanismos de atenção (*attention mechanisms*) para processar e gerar textos.

Historicamente, os modelos de linguagem evoluíram de abordagens estatísticas simples, como n-gramas, para sistemas capazes de lidar com tarefas complexas, como tradução automática, sumarização e geração de texto. Os LLMs modernos são treinados em *corpora* de grande escala, que incluem materiais de diversos gêneros e domínios, o que lhes permite gerar textos em uma ampla variedade de estilos e contextos.

Uma característica central dessas ferramentas de IA é justamente a capacidade de produzir textos de forma contextualizada, baseada na análise de padrões sintáticos, semânticos e discursivos aprendidos durante o treinamento. De acordo com Radford *et al.* (2019), essa capacidade é possibilitada pelo mecanismo de atenção, que permite ao modelo atribuir pesos distintos a diferentes partes do prompt de entrada, resultando em respostas mais coesas e pertinentes. Tal mecanismo confere aos modelos uma flexibilidade que lhes permite adaptar-se, por exemplo, a diferentes gêneros textuais. Além disso, estudos como o de Touvron *et al.* (2023) demonstram que esses grandes modelos de linguagem generativa conseguem alcançar um nível avançado de compreensão textual, respondendo adequadamente às tarefas solicitadas.

2.3 Métricas de complexidade textual

A medição da complexidade textual é uma tarefa multifacetada por excelência, abrangendo dimensões variadas. A depender de como o conceito de complexidade é desenvolvido, ele pode incluir desde elementos estritamente linguísticos, como o grau de sofisticação do vocabulário empregado, até fatores predominantemente cognitivos, como a necessidade de o leitor integrar as novas informações apresentadas com seu repertório de conhecimentos prévios (Frantz; Starr; Bailey, 2015).

Diversas métricas e ferramentas já foram desenvolvidas para mensurar a complexidade textual. Entre as abordagens mais utilizadas estão métricas de diversidade lexical, como a razão type-token (TTR), medidas de complexidade sintática, como o comprimento médio das sentenças, e indicadores de coesão e coerência textual. Uma das vantagens dessas métricas é o fato de permitirem uma avaliação quantitativa da complexidade, conferindo um nível adicional de objetividade às análises.

Entretanto, tais métricas apresentam limitações se aplicadas de maneira isolada. Um argumento interessante nesse sentido é dado por McNamara, Louwerse e Graesser (2002). Comparando os exemplos (1) e (2), os autores mostram que o primeiro deles apresenta um nível de coesão mais baixo, pois não há nenhuma indicação linguística que aponte a relação de causa entre as duas sentenças. O resultado desse menor nível de coesão é que, diante de um exemplo como (1), o leitor teria um

desafio maior para estabelecer a relação de causa, já que isto poderia ser inferido unicamente a partir da condição de adjacência entre as duas sentenças do enunciado Em outras palavras, unicamente por essa métrica, o exemplo (1) seria considerado mais complexo, pois impõe um nível maior de interpretação. Porém, um índice de complexidade que levasse em consideração apenas o parâmetro sintático de tamanho da sentença apontaria que o exemplo (1) é mais simples do que o exemplo (2), pois se trata de um enunciado menor. O que esses dados mostram, na análise dos autores, é que a complexidade textual deve ser medida de maneira mais ampla, por meio da aplicação de diferentes métricas.

1. One part of the cloud develops a downdraft. Rain begins to fall[1].
2. One part of the cloud develops a downdraft, which causes rain to fall[2].

No contexto do português brasileiro, a ferramenta NILC-Metrix[3] destaca-se como, talvez, o mais importante recurso para análise de complexidade textual. Desenvolvido pelo Núcleo Interinstitucional de Linguística Computacional (NILC)[4], o NILC-Metrix integra 200 métricas linguísticas, cobrindo áreas como complexidade sintática, diversidade lexical, uso de conectivos e coesão textual, entre outros aspectos (Leal *et al.*, 2024). Esse conjunto de métricas oferece uma visão multidimensional da estrutura textual, permitindo a análise não apenas de características superficiais, como comprimento de palavras e sentenças, mas também de aspectos mais profundos, como a organização discursiva e a progressão temática. E é justamente por essa capacidade de analisar múltiplas dimensões da complexidade textual de forma integrada que o NILC-Metrix é especialmente útil para avaliar a performance de LLMs na reprodução de um gênero do discurso tão complexo como o editorial.

3 Metodologia

3.1 Coleta de dados

Para a investigação que nos propusemos a fazer aqui, coletamos cinquenta editoriais do jornal *Folha de São Paulo* (FSP)[5], publicados entre os dias 27 de setembro e 21 de outubro de 2024. Esses textos foram selecionados para representarem a amostra de editoriais produzidos por humanos No caso específico dos editoriais da FSP, são textos escritos por uma equipe de editorialistas e representam o ponto de vista da empresa sobre os assuntos discutidos[6].

Paralelamente, foram produzidos cinquenta textos utilizando o modelo de inteligência artificial generativa *Sabiá-3*, desenvolvido pela empresa brasileira Maritaca AI[7]. Trata-se de um modelo treinado com um corpus de dados disponíveis até meados de 2023. O *Sabiá-3* é baseado na arquitetura *Transformers*, no entanto a configuração exata, incluindo o número de parâmetros, é uma informação não disponibilizada pela empresa. Apesar dessa não publicização, o que pode ser visto como uma propriedade desfavorável, uma vantagem da ferramenta nacional em relação a outros modelos é o fato de ter sido treinada exclusivamente com dados em língua portuguesa, o que a torna particularmente adequada para a geração de textos nessa língua. Um relatório técnico sobre esse LLM é apresentado em Abonizio *et al.* (2025).

Os textos gerados pela IA foram produzidos com base no seguinte prompt: "Escreva um editorial de jornal, com aproximadamente 425 palavras[8], sobre o seguinte assunto:", seguido do título e subtítulo de um editorial real da FSP. Para cada editorial produzido pelo LLM, utilizou-se o título e o subtítulo de um dos editoriais da amostra humana. Como o *Sabiá-3* foi treinado com material disponível até meados de 2023, essa abordagem garantiu que o modelo não tivesse acesso aos editoriais humanos que utilizamos, evitando desse modo eventuais plágios. Além disso, como os editoriais da FSP e aqueles gerados pela IA abordam a mesma temática, isso minimiza eventuais discrepâncias que pudessem

[1] "Uma parte da nuvem desenvolve uma corrente de ar descendente. A chuva começa a cair" (tradução nossa).
[2] "Uma parte da nuvem desenvolve uma corrente de ar descendente, o que faz com que a chuva comece a cair" (tradução nossa).
[3] http://fw.nilc.icmc.usp.br:23380/nilcmetrix.
[4] https://sites.google.com/view/nilc-usp/.
[5] https://www.folha.uol.com.br/.
[6] https://www1.folha.uol.com.br/tv/2022/07/folha-explica-a-diferenca-entre-os-tipos-de-texto-que-voce-le-no-jornal.shtml.
[7] https://www.maritaca.ai/.
[8] Essa é quantidade média de palavras dos editoriais da FSP.

surgir entre as amostras por conta de diferenças temáticas.

3.2 Ferramenta de análise e métricas selecionadas

Para avaliar a complexidade textual dos editoriais, utilizamos o NILC-Metrix, já apresentado na Seção 2.3. Tendo em vista as restrições de espaço para se analisar em um único artigo os resultados de todas as métricas, escolhemos cinco delas, apresentadas a seguir.

I. **Sílabas por palavra.** Essa métrica calcula o número médio de sílabas por palavras de conteúdo no texto. São consideradas palavras de conteúdo os itens vocabulares de quatro classes gramaticais: substantivos, verbos, adjetivos e advérbios. Assume-se que um número maior de sílabas por palavras corresponde a um grau de complexidade maior.

II. **Proporção de types em relação à quantidade de tokens.** Trata-se de uma métrica que computa a proporção de palavras sem repetições (types) em relação ao total de palavras com repetições (tokens). Cada flexão é tratada como um type diferente. Na documentação oficial do NILC-Metrix, considera-se que a complexidade de um texto é proporcional ao valor dessa métrica, isto é, valores maiores correspondem a textos mais complexos.

III. **Orações subordinadas.** Essa métrica calcula a proporção de orações subordinadas pela quantidade de orações do texto. Assume-se que orações dependentes são estruturas mais complexas, demandando um maior esforço de processamento. Assim, quanto maior o resultado dessa métrica, maior a complexidade textual.

IV. **Conjunções difíceis.** Trata-se de uma métrica que computa a proporção de conjunções difíceis em relação a todas as palavras do texto. A interpretação da métrica é que quanto maior o resultado obtido, maior a complexidade textual. Os itens lexicais que fazem parte dessa categoria são: todavia, eis, a fim de, ao passo que, para que, conforme, tais, ou seja, contudo, bem como, logo, à medida que, entretanto, desde que, mesmo que, ainda que, de acordo com, uma vez que, por sua vez, sobretudo, até, ainda, caso, no entanto, nem, quanto, já, como, já que, outrossim, mas também, como também, não só, mas ainda, tampouco, senão também, bem assim, ademais, antes, não obstante, sem embargo, ao passo que, de outra forma, em todo caso, aliás, de outro modo, por conseguinte, em consequência de, por consequência, consequentemente, conseguintemente, isso posto, pelo que, de modo que, de maneira que, de forma que, em vista disso, por onde, porquanto, posto que, isto é, ademais, senão, dado que, visto como, vez que, de vez que, pois que, agora, na medida em que, sendo que, como que, como quer que, eis que, sendo assim, tal qual, ao invés de, conquanto, por muito que, visto que, uma vez que, quanto mais, quanto menos, se bem que, apesar de que, suposto que, ainda quando, quando mesmo, a despeito de, conquanto que, sem embargo de que, por outro lado, em contrapartida, sem embargo, muito embora, inclusive se, por mais que, por menos que, por pouco que, contanto que, salvo se, com tal que, caso que, consoante, tal que, de forma que, à proporção que, ao passo que, mal, tão logo, entretanto, sob esse aspecto, sob esse prisma, sob esse ponto de vista, sob esse enfoque, embora, portanto, além disso[9].

V. **Entropia cruzada.** Essa métrica avalia o grau de "surpresa" de um modelo de linguagem ao processar uma sentença, calculando a média do valor de entropia cruzada das sentenças do texto. Para calcular essa medida, o NILC-Metrix utiliza um modelo estatístico de trigramas com suavização Kneser-Ney modificada, gerado pela ferramenta KenLM (https://github.com/kpu/kenlm). Os valores obtidos podem variar de 0 a 1. Um índice maior de entropia cruzada corresponde a uma maior complexidade das palavras em relação ao modelo de língua estatístico treinado, isto é, valores maiores representam um nível maior de combinações não usuais de palavras.

[9] Essa lista foi definida por um linguista da antiga plataforma Guten Educação, agora denominada Árvore Atualidades (https://www.arvore.com.br/). Na documentação oficial do NILC-Metrix, nenhuma informação adicional de autoria e critérios de classificação é apresentada.

4 Resultados

Na Tabela 1, apresentamos a média geral para cada uma das cinco métricas escolhidas no estudo[10].

Tabela 1. Média dos valores das métricas escolhidas para as amostras de editoriais produzidos por humanos e editoriais gerados pelo modelo *Sabiá-3*.

Métrica	Humanos	IA
Sílabas por palavra	2.931442	3.247690
Proporção de types em relação à quantidade de tokens	0.7845508	0.7373788
Orações subordinadas	0.5021336	0.5897614
Conjunções difíceis	0.0234116	0.0211284
Entropia cruzada	0.5843956	0.4559052

Fonte: Elaboração própria.

Para verificar se as diferenças de média entre a amostra de editoriais humanos e a amostra de editoriais gerados por IA são estatisticamente significativas, aplicamos o teste t para duas amostras independentes. Esse teste tem o objetivo de comparar o efeito de uma mesma variável em duas populações independentes e distribuídas identicamente. A estatística do teste t é dada por

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{1}$$

onde:

- $\bar{X}_1, \bar{X}_2$ são as médias amostrais dos dois grupos;
- s_1^2, s_2^2 são as variâncias amostrais;
- n_1, n_2 são os tamanhos das amostras;
- s_p^2 é a variância combinada.

O teste t para duas amostras independentes testa as seguintes hipóteses:

- Hipótese nula (H_0): as médias das duas populações são iguais, ou seja, eventuais diferenças não são estatisticamente significativas;
- Hipótese alternativa (H_1): as médias das populações são diferentes, ou seja, diferenças atestadas são estatisticamente significativas.

No nosso caso, H_0 assume que as diferenças de média entre as duas amostras para cada métrica de complexidade textual não são estatisticamente significativas se o nível de significância (valor de p) associado ao índice de t for maior que 0.05, o que equivale dizer que as diferenças de média não são afetadas pela variável tipo de produtor (humanos *versus* IA). Já H_1 postula que as diferenças de média entre as duas amostras são estatisticamente significativas (com o valor de $p \leq 0.05$), ou seja, a variável tipo de produtor afeta as diferenças de média. Os resultados de aplicação do teste t são apresentados na Tabela 2.

Tabela 2. Valores do teste t para as métricas escolhidas e respectivos valores de p.

Métrica	valor de t	valor de p
Sílabas por palavra	-12.134	<2.2e-16
Proporção de types em relação à quantidade de tokens	12.554	<2.2e-16
Orações subordinadas	-6.5845	2.276e-09
Conjunções difíceis	1.3349	0.185
Entropia cruzada	20.223	<2.2e-16

Fonte: Elaboração própria.

[10] Os resultados obtidos foram submetidos a uma análise estatística com o software R 4.2.2 (Team, 2021).

Como mostram os resultados da Tabela 2, "conjunções difíceis" é a única métrica que não apresenta uma diferença de média estatisticamente significativa. Todas as outras apresentam valores de t associados a níveis de significância abaixo de 0.05, o que nos leva a assumir que, no caso dessas quatro métricas, as diferenças de média são estatisticamente significativas, implicando que há efeito da variável tipo de produtor sobre o resultado da média.

Esses resultados são reforçados quando olhamos para os diagramas de caixa na Figura 1. Ali vemos que, apenas na métrica "conjunções difíceis", o intervalo interquartil das duas amostras apresenta uma considerável sobreposição, o que aponta para uma dispersão dos dados centrais mais próxima entre os dois conjuntos. Isso vai ao encontro do resultado de que a diferença de média entre textos humanos e textos gerados por IA na métrica "conjunções difíceis" não é estatisticamente significativa. Nas demais métricas, o que se vê é uma clara falta de alinhamento da dispersão dos dados centrais, o que corrobora o resultado já apresentado de que, para tais métricas, as diferenças de média entre as duas amostras são estatisticamente significativas.

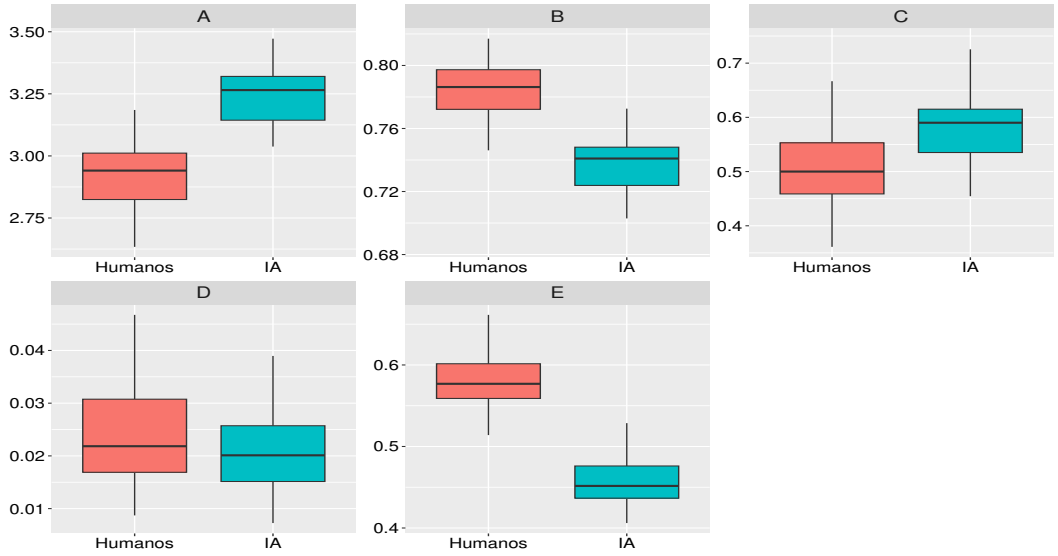

Figura 1. Medidas de dispersão e tendência central para as métricas "sílabas por palavra" (A), "proporção de types em relação à quantidade de tokens" (B), "orações subordinadas" (C), "conjunções difíceis" (D) e "entropia cruzada" (E).

Fonte: Elaboração própria.

5 Discussão

5.1 Métricas com maior complexidade na amostra humana

Iniciamos a discussão dos resultados comentando as duas métricas em que a amostra de editoriais produzidos por humanos apresentou índices de média, com diferenças estatisticamente significativas, superiores aos valores obtidos pela amostra de textos gerados pelo modelo *Sabiá-3*. Essas duas métricas, como apontado na Seção 4, são "proporção de types em relação à quantidade de tokens" e "entropia cruzada". Em ambas, valores maiores correspondem a índices mais elevados de complexidade textual, o que nos permite dizer que, nesses quesitos, os textos humanos se mostram mais complexos.

Um primeiro ponto a ser destacado para explicar essa maior complexidade dos editoriais humanos diz respeito à métrica "proporção de types em relação à quantidade de tokens". A média maior nessa medida sugere um uso mais diversificado do léxico. No caso dos editoriais da FSP, essa diversidade mais elevada pode estar vinculada à criatividade típica da escrita humana. Enquanto LLMs operam com distribuições estatísticas aprendidas a partir de *corpora*, escritores humanos selecionam palavras não apenas com base em probabilidades, mas também motivados por nuances semânticas, estilísticas e pragmáticas (Piantadosi, 2014). Essa flexibilidade permite maior inovação lexical, incluindo o uso de palavras menos frequentes ou combinações inusitadas, elevando a proporção de types. Já as ferramentas de IA priorizam tokens mais comuns para garantir coerência, o que acaba por reduzir a

diversidade lexical (Holtzman *et al.*, 2020).

Outro aspecto da relação type/token é que LLMs fazem uso de padrões repetitivos como estratégia para minimizar erros ou incoerências. Esse comportamento, conhecido como "degeneração" (*degeneracy*), é documentado, por exemplo, em Welleck *et al.* (2019), que observam a tendência de modelos generativos repetirem estruturas ou palavras para manter a fluência superficial. Em contraste, particularmente em contextos formais, escritores humanos monitoram ativamente a redundância, ajustando o texto para evitar repetições desnecessárias — um processo vinculado à "teoria da monitoração" (Levelt, 1989). Assim, a menor proporção de types/tokens nos textos gerados por IA pode refletir essa busca conservadora dos grandes modelos de linguagem generativa por segurança estatística.

Outro ponto a ser detacado em relação a essa métrica tem a ver com o fato de que, como mencionado anteriormente, o NILC-Metrix considera flexões como types distintos, o que pode apontar para uma certa incapacidade dos LLMs em utilizar variações morfológicas sutis. Por exemplo, em textos escritos formais, ainda é usual em produções humanas a presença de formas verbais mais raras na fala, como flexões do modo subjuntivo (Almeida; Callou, 2010). Já em relação a grandes modelos de linguagem generativa, Gulordava *et al.* (2018) demonstram que tais ferramentas apresentam uma maior dificuldade em generalizar padrões morfológicos complexos, especialmente em línguas com alta carga flexional, como é o caso do português. Essa limitação poderia reduzir a diversidade de types nos textos gerados.

Ainda em relação à proporção types/tokens, cabe dizer que textos humanos são moldados por contextos situacionais e objetivos comunicativos específicos. Editoriais, em particular, exigem adaptação a audiências, tons e finalidades — o que incentiva a variação lexical (Biber; Conrad, 2009). Modelos de linguagem generativa, por outro lado, carecem de intencionalidade genuína. Sua produção é orientada por prompts, o que pode levar a um repertório lexical mais estático.

No que diz respeito à métrica "entropia cruzada", o maior nível de complexidade dos textos humanos nesse quesito também se explica nas mesmas linhas da questão da diversidade lexical explorada para a métrica anterior. Podemos retomar, por exemplo, o ponto de que LLMs priorizam combinações de palavras mais prováveis, alinhadas aos padrões estatísticos dominantes em seus dados de treinamento, o que reduz a "surpresa" medida pela entropia cruzada. Já textos humanos, como refletem escolhas individuais do autor, influências estilísticas e adpatações contextuais, podem apresentar mais combinações idiossincráticas, consequentemente aumentando o grau de entropia cruzada.

5.2 Métricas com maior complexidade na amostra automatizada

Passemos agora às duas métricas em que os textos gerados pelo modelo *Sabiá-3* apresentam uma média maior em relação ao conjunto de editoriais da FSP, a saber, "sílabas por palavra" e "orações subordinadas". Nessas métricas, como nas duas anteriores discutidas, vimos que as diferenças são estatisticamente significativas. Aqui também valores maiores indicam um maior nível de complexidade. Um primeiro aspecto explicando essa diferença pode estar relacionado a questões de processamento linguístico. Gibson (2000) apresenta evidências de que a produção humana de linguagem é limitada por restrições de memória operacional. Por exemplo, esse autor mostra que o aumento de orações relativas tende a aumentar a dificuldade de compreensão, como ilustrado nos exemplos de (3) a (5).

3. The reporter disliked the editor.
4. The reporter [$_S$ who the senator attacked] disliked the editor.
5. *The reporter [$_S$ who the senator [$_S$ who John met] attacked] disliked the editor.

Em (3), temos uma oração simples com nenhum material interveniente entre o sujeito *the reporter* e o verbo com o qual ele está associado, *disliked*. Em (2), é encaixada uma oração relativa entre o sintagma nominal *the reporter* e o verbo *disliked*. Em (3), é colocada uma oração relativa entre o sujeito e o verbo da primeira oração encaixada (*the senator* e *attacked*). De acordo com o autor, essa última sentença é tão complexa que, para a maioria dos falantes nativos do inglês, ela apresenta uma alta dificuldade de processamento[11]. Aplicando esse fato à nossa discussão, podemos dizer que

[11] Em português, as respectivas traduções para os enunciados de (3) a (5) recebem juízos de gramaticalidade equivalentes, sugerindo não se tratar de uma restrição específica do inglês, mas sim de algo mais geral em termos de processamento cognitivo.

i. O repórter não gostava do editor.

o uso excessivo de palavras mais longas ou de orações subordinadas pode sobrecarregar o leitor, o que levaria escritores mais experientes a moderar a sua frequência. Já os modelos de IA, desprovidos de tais limitações cognitivas, podem gerar estruturas gramaticais densas (sejam elas de natureza fonológica, morfoloógica ou sintática) sem avaliar seu impacto na compreensão. Isso explicaria por quê o *Sabiá-3* apresenta uma média maior de sílabas por palavras de conteúdo e também de orações subordinadas, quando comparado com a amostra de editoriais humanos.

Especificamente em relação à métrica "orações subordinadas", é importante destacar que estruturas encaixadas seguem padrões sintáticos relativamente previsíveis (isto é, são construções sintáticas introduzidas por um conjunto fechado de conjunções, tais como "que", "quando", "embora"), o que, por hipótese, facilita sua geração automática. LLMs podem recorrer a essas estruturas como "atalhos" para estender o texto de forma coerente, mesmo quando a subordinação não acrescenta informação relevante. Consequentemente, essa eventual estratégia de LLMs pode resultar em textos com mais sentenças encaixadas, quando comparados com produções similares realizadas por humanos.

5.3 Similaridade de complexidade nas duas amostras

Por fim, discutimos a métrica "conjunções difíceis", para a qual não encontramos diferença de média estatisticamente significativa entre as duas amostras. Em termos quantitativos, isso nos permite dizer que, nessa métrica, as duas amostras apresentam comportamento similar no que diz respeito à complexidade textual. Ao contrário do que fizemos anteriormente, o ponto agora é saber por que ocorre essa semelhança. Um aspecto que nos parece importante é a natureza gramatical dos itens lexicais avaliados nessa métrica. A classe das conjunções é, em termos morfossintáticos, de natureza fechada, já que abriga essencialmente elementos funcionais. Isso pode ter um impacto no grau de diversidade de uso desses elementos, pois, diferentemente de classes abertas, como substantivos e verbos, as conjunções não permitem uma expansão ilimitada de seu repertório lexical, o que restringe a margem para variação estilística ou criativa, reduzindo o espectro de possíveis diferenças entre produção humana e automatizada.

Uma forma de testar essa hipótese é comparar o comportamento das duas amostras em uma métrica que avalie itens lexicais de alguma classe aberta. Quando discutimos a métrica "proporção de types em relação à quantidade de tokens", um dos pontos que levantamos é que o LLM em análise se mostra menos complexo nessa métrica justamente porque o seu processo de escolha lexical é essencialmente baseado em escolhas probabilísticas. Essa característica impacta no grau de diversidade lexical atestado pela relação type/token. No caso de produções humanas, as escolhas lexicais não são motivadas unicamente por fatores de ordem probabilística, mas também por uma série de outras variáveis, tais como questões de ordem estilística, semântica ou pragmática. Essa maior flexibilidade permitiria um nível maior de diversidade lexical. Seguindo esse raciocínio, a ideia que assumimos aqui é que classes abertas, justamente por não imporem um limite quantitativo em seu universo de escolhas, sejam o espaço por excelência em que textos humanos se mostrem mais complexos do que produções textuais geradas por IA. Para verificar a validade desse argumento, apresentamos duas métricas que avaliam itens lexicais de classes abertas. Uma delas é a métrica "diversidade de substantivos", que calcula a proporção de types de substantivos em relação à quantidade de tokens de substantivos no texto. A outra métrica é "diversidade de verbos", que computa a proporção de types de verbos em relação à quantidade de tokens de verbos no texto. Nos dois casos, índices quantitativos maiores são interpretados como indicadores de maior complexidade textual. Na Tabela 3, apresentamos a média dessas duas métricas para os dois grupos comparados.

Como se vê, em ambas as métricas, a amostra de editoriais da FSP mostrou um índice de média superior ao apresentado pelos textos gerados por IA. Essa diferença é estatisticamente significativa nas duas medidas, como mostram os resultados na Tabela 4 dos valores de t e p resultantes da aplicação do teste t para duas amostras independentes.

Essa maior complexidade dos textos humanos nas métricas "diversidade de substantivos" e "diversidade de verbos" pode ser visualizada também nos diagramas de caixa na Figura 2, que apresentam

ii. O repórter [$_S$ que o senador atacou] não gostava do editor.
iii. *O repórter [$_S$ que o senador [$_S$ que o João encontrou] atacou] não gostava do editor.

Tabela 3. Média dos valores das métricas "diversidade de substantivos" e "diversidade de verbos" para as amostras de editoriais produzidos por humanos e editoriais gerados pelo modelo *Sabiá-3*.

Métrica	Humanos	IA
Diversidade de substantivos	0.8486522	0.7412310
Diversidade de verbos	0.9030712	0.7890242

Fonte: Elaboração própria.

Tabela 4. Valores do teste t para as métricas "diversidade de substantivos" e "diversidade de verbos" e respectivos valores de p.

Métrica	valor de t	valor de p
Diversidade de substantivos	9.6068	8.641e-16
Diversidade de verbos	11.077	<2.2e-16

Fonte: Elaboração própria.

medidas de dispersão e tendência central.

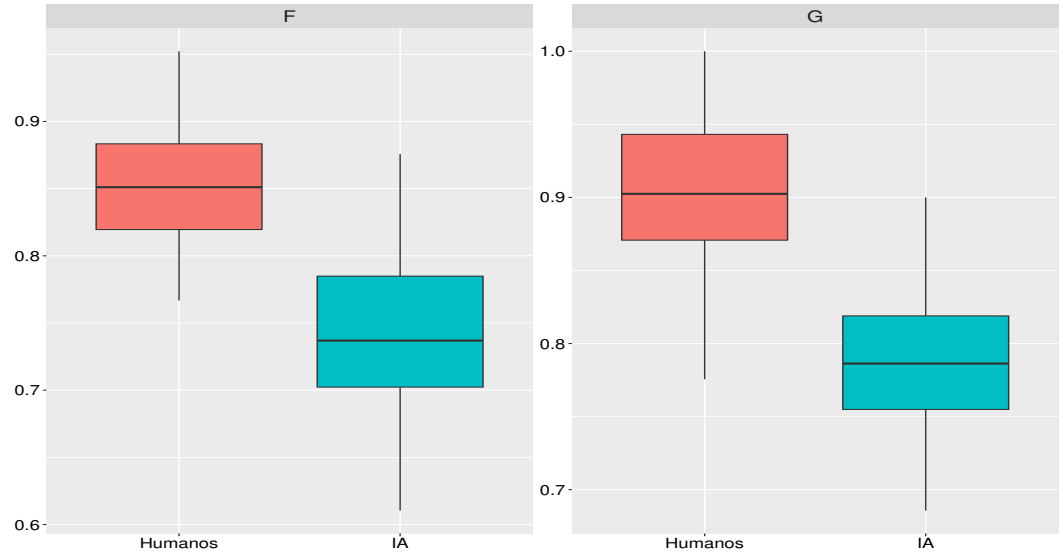

Figura 2. Medidas de dispersão e tendência central para as métricas "diversidade de substantivos" (F) e "diversidade de verbos" (G).

Fonte: Elaboração própria.

Esses resultados de métricas de classes abertas corroboram a nossa hipótese de que a questão da diversidade lexical, quando comparados textos humanos e textos gerados por IA, é mais saliente no primeiro grupo, que não se baseia unicamente em parâmetros probabilísticos para a escolha de itens vocabulares. Em relação a classes fechadas, como é o caso das conjunções, o espaço para diversificação lexical é limitado, pela própria natureza da classe gramatical, o que explicaria o nível de complexidade semelhante entre os editoriais humanos e os editoriais do modelo *Sabiá-3* na métrica "conjunções difíceis".

6 Conclusão

Neste artigo, avaliamos a *performance* do modelo de linguagem generativa *Sabiá-3* na tarefa de reproduzir aspectos de complexidade textual do gênero discursivo editorial, usando como ponto de referência editoriais produzidos por humanos. Para essa tarefa, utilizamos métricas da ferramenta

computacional NILC-Metrix. Em nossa análise comparativa, os resultados revelaram diferenças estatisticamente significativas em quatro das cinco métricas analisadas, destacando padrões distintos entre os dois grupos.

Os textos humanos demonstraram maior complexidade nas métricas "proporção de types em relação à quantidade de tokens" e "entropia cruzada". Argumentamos que esse resultado pode estar vinculado, por exemplo, à capacidade humana de selecionar palavras ou realizar combinações lexicais não limitada por parâmetros probabilísticos, ao contrário de LLMs. Essa flexibilidade inerente de falantes de línguas naturais permitiria maior inovação lexical, capacidade esta que se concretizaria no uso mais saliente de palavras menos frequentes ou combinações inusitadas. Por outro lado, os textos gerados pela IA apresentaram maior complexidade nas medidas "sílabas por palavra" e "orações subordinadas", possivelmente devido ao fato, entre outros aspectos, de que essas ferramentas de geração automatizada de textos não sofrem restrições de processamento cognitivo, diferentemente do que ocorre com seres humanos. A única métrica sem diferença estatística significativa foi "conjunções difíceis". Atribuímos esse resultado à natureza fechada dessa classe gramatical, que limitaria variações.

Diante dessas considerações, concluímos que, embora o modelo *Sabiá-3* apresente indícios de sofisticação linguística em certos aspectos da linguagem escrita, ainda há lacunas em replicar a riqueza lexical e a adaptabilidade estilística da escrita humana. Esses resultados reforçam a importância de se considerar múltiplas dimensões da complexidade textual ao avaliar a produção de LLMs, especialmente quando se trata de gêneros que exigem domínio estilístico e argumentativo refinado, como é o caso do editorial.

7 Agradecimentos

Este artigo se beneficiou da leitura atenta de Juliano Desiderato Antonio e de dois pareceristas anônimos da revista. Eventuais problemas ainda presentes são de inteira responsabilidade do autor do trabalho.

Referências

ABONIZIO, Hugo; ALMEIDA, Thales Sales; LAITZ, Thiago; MALAQUIAS JUNIOR, Roseval; BONÁS, Giovana Kerche; NOGUEIRA, Rodrigo; PIRES, Ramon. *Sabiá-3 technical report*. [*S. l.: s. n.*], 2025. arXiv: 2410.12049. Disponível em: https://arxiv.org/abs/2410.12049.

ALMEIDA, Erica; CALLOU, Dinah. Sobre o uso variável do subjuntivo em português: um estudo de tendência. *In:* BRITO, Ana Maria; SILVA, Maria de Fátima Henriques da; VELOSO, João; FIÉIS, Alexandra (ed.). *XXV Encontro Nacional da Associação Portuguesa de Linguística: Textos seleccionados*. Porto: Apl, 2010. p. 143–152.

BAKHTIN, Mikhail. *Estética da criação verbal*. São Paulo: Martins Fontes, 1997.

BIBER, Douglas; CONRAD, Susan. *Register, genre, and style*. Cambridge: Cambridge University Press, 2009.

FRANCESCHELLI, Giorgio; MUSOLESI, Mirco. On the creativity of large language models. *Ai & Society*, p. 1–11, 2024.

FRANTZ, Roger; STARR, Laura; BAILEY, Alison. Syntactic complexity as an aspect of text complexity. *Educational Researcher*, v. 44, n. 7, p. 387–393, 2015.

GIBSON, Edward. The dependency locality theory: a distance-based theory of linguistic complexity. *In:* MARANTZ, Alec; MIYASHITA, Yasushi; O'NEIL, Wayne (ed.). *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*. Cambridge, MA: The MIT Press, 2000. p. 95–126.

GULORDAVA, Kristina; BOJANOWSKI, Piotr; GRAVE, Edouard; LINZEN, Tal; BARONI, Marco. *Colorless green recurrent networks dream hierarchically*. [*S. l.: s. n.*], 2018. arXiv: 1803.11138. Disponível em: https://arxiv.org/abs/1803.11138.

HOLTZMAN, Ari; BUYS, Jan; DU, Li; FORBES, Maxwell; CHOI, Yejin. *The curious case of neural text degeneration*. [*S. l.: s. n.*], 2020. arXiv: 1904.09751. Disponível em: https://arxiv.org/abs/1904.09751.

LEAL, Sidney Evaldo; DURAN, Magali Sanches; SCARTON, Carolina Evaristo; HARTMANN, Nathan Siegle; ALUÍSIO, Sandra Maria. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, v. 58, n. 1, p. 73–110, 2024.

LEVELT, Willem. *Speaking: from intention to articulation*. Cambridge, MA: The MIT Press, 1989.

MCNAMARA, Danielle; LOUWERSE, Max; GRAESSER, Arthur. *Coh-Metrix: automated cohesion and coherence scores to predict text readability and facilitate comprehension*. [*S. l.: s. n.*], 2002. Technical report, Institute for Intelligent Systems, University of Memphis, TN.

NAIK, Dishita; NAIK, Ishita; NAIK, Nitin. Applications of AI chatbots based on generative AI, large language models and large multimodal models. *In:* NAIK, Nitin; JENKINS, Paul; PRAJAPAT, Shaligram; GRACE, Paul (ed.). *Contributions Presented at The International Conference on Computing, Communication, Cybersecurity and AI, July 3–4, 2024, London, UK*. Cham: Springer Nature Switzerland, 2024. p. 668–690.

PIANTADOSI, Steven T. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*, v. 21, p. 1112–1130, 2014.

RADFORD, Alec; WU, Jeffrey; CHILD, Rewon; LUAN, David; AMODEI, Dario; SUTSKEVER, Ilya. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 1–9, 2019.

REISENBICHLER, Martin; REUTTERER, Thomas; SCHWEIDEL, David A.; DAN, Daniel. Frontiers: supporting content marketing with natural language generation. *Marketing Science*, v. 41, n. 3, p. 441–452, 2022.

TEAM, R Core. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: https://www.R-project.org/.

TOUVRON, Hugo; LAVRIL, Thibaut; IZACARD, Gautier; MARTINET, Xavier; LACHAUX, Marie-Anne; LACROIX, Timothée; ROZIÈRE, Baptiste; GOYAL, Naman; HAMBRO, Eric; AZHAR, Faisal; RODRIGUEZ, Aurelien; JOULIN, Armand; GRAVE, Edouard; LAMPLE, Guillaume. *LLaMA: open and efficient foundation language models*. [*S. l.: s. n.*], 2023. arXiv: 2302.13971. Disponível em: https://arxiv.org/abs/2302.13971.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan; KAISER, Lukasz; POLOSUKHIN, Illia. Attention is all you need. *In:* LUXBURG, Ulrike von; GUYON, Isabelle; BENGIO, Samy; WALLACH, Hanna; FERGUS, Rob; VISHWANATHAN, S.; GARNETT, Roman (ed.). *Advances in Neural Information Processing Systems 30*. Red Hook: Curran Associates, Inc., 2018. p. 5999–6009.

VIEIRA, Rosaura Maria Marques. O editorial de jornal. *In:* DELL'ISOLA, Regina Lúcia Péret (ed.). *Nos domínios dos gêneros textuais*. Belo Horizonte: Fale/ufmg, 2009. v. 2. p. 15–20.

WELLECK, Sean; KULIKOV, Ilia; ROLLER, Stephen; DINAN, Emily; CHO, Kyunghyun; WESTON, Jason. *Neural text generation with unlikelihood training*. [*S. l.: s. n.*], 2019. arXiv: 1908.04319. Disponível em: https://arxiv.org/abs/1908.04319.