# Densidade lexical em textos gerados pelo ChatGPT: implicações da inteligência artificial para a escrita em línguas adicionais

## Lexical density in texts generated by ChatGPT: implications of artificial intelligence for writing in additional languages

**Antonio Marcio Da Silva** [*1] **e Lucia Rottava** [†2]

[1] *University of Essex, Colchester, Inglaterra.*
[2] *Universidade Federal do Rio Grande do Sul, Instituto de Letras, Porto Alegre, RS, Brasil.*

## Resumo

O avanço tecnológico tem tido um grande impacto na produção escrita, especialmente em Línguas Adicionais (LAs). Embora a tecnologia tenha trazido novas oportunidades para o ensino de LAs, ela também apresenta desafios, incluindo preocupações sobre a complexidade da escrita e a autenticidade dos trabalhos dos alunos. Uma dessas ferramentas é o ChatGPT, plataforma de inteligência artificial (IA) que tem sido objeto de debates desde sua popularização em 2022. Este estudo analisa um corpus composto por seis tarefas produzidas pelo ChatGPT em cinco idiomas (alemão, espanhol, francês, italiano e português), considerando os níveis de proficiência propostos pelo Quadro Comum Europeu de Referência para Línguas (CEFR), que totalizou 2991 textos e 706,401 palavras. Os dados foram gerados por alunos em um laboratório de informática em uma universidade britânica a partir de 100 diferentes perfis na plataforma do ChatGPT, seguindo instruções dos pesquisadores. A análise dos dados utiliza a linguística sistêmico-funcional (LSF) e o conceito de densidade lexical (Halliday, 1985, 1987, 1993; Halliday; Matthiessen, 2014) para investigar a complexidade dos textos produzidos, dado que a complexidade lexical está relacionada à proficiência na escrita, na qual textos mais avançados usam proporcionalmente mais "palavras de conteúdo" (nomes, verbos, adjetivos e alguns advérbios de modo). Os resultados revelam que o ChatGPT não segue as instruções das tarefas quanto ao número de palavras solicitadas, impactando, assim, no cálculo da densidade lexical, nem produz textos que mostram diferenças significativas da densidade lexical entre as línguas adicionais e níveis de proficiência.

*Palavras-chave*: Línguas adicionais. ChatGPT. Inteligência Artificial. Linguística Sistêmico-funcional. Densidade lexical.

## Abstract

Technological advancement has had a significant impact on written production, especially in Additional Languages (ALs). Although technology has brought new opportunities for AL teaching, it also poses challenges, including concerns about the complexity of writing and the authenticity of students' work. One such tool is ChatGPT, an artificial intelligence (AI) platform that has been the subject of debate since its popularization in 2022. This study analyses a corpus consisting of six tasks produced by ChatGPT in five languages (German, Spanish, French, Italian, and Portuguese), considering the proficiency levels proposed by the Common European Framework of Reference for Languages (CEFR), totalling 2991 texts and 706,401 words. The data were generated by students in a computer lab at a British university from 100 different profiles on the ChatGPT platform, following the researchers' instructions. Data analysis employs Systemic Functional Linguistics (SFL) and the concept of lexical density (Halliday, 1985, 1987, 1993; Halliday; Matthiessen, 2014) to investigate the complexity of the produced texts, as lexical complexity is related to proficiency in writing, where more advanced texts proportionally use more "content words" (nouns, verbs, adjectives, and some adverbs of manner). The results reveal that ChatGPT does not adhere to task instructions regarding the requested word count, thereby impacting the calculation of lexical density, nor does it produce texts that show significant differences in lexical density among additional languages and proficiency levels.

*Keywords*: Additional languages. ChatGPT. Artificial Intelligence. Systemic Functional Linguistics. Lexical density.

Textoline
Linguagem e Tecnologia

**DOI:** 10.1590/1983-3652.2024.47836

**Seção:**
Dossiê

**Autor Correspondente:**
Antonio Marcio da Silva

**Editor de seção:**
Danniervelin Pereira
**Editor de layout:**
Leonado Araújo

**Recebido em:**
24 de agosto de 2023
**Aceito em:**
5 de outubro de 2023
**Publicado em:**
29 de novembro de 2023



*Email: antonio.dasilva@essex.ac.uk
†Email: luciarottava@yahoo.com.br

## 1 Introdução

Pensar a relação entre a inteligência artificial (IA) e a produção escrita em línguas adicionais (LAs) é um desafio pelo conjunto de variáveis: a tecnologia envolvida, o propósito motivador e orientador da escrita e o nível de conhecimento de uma língua que o texto gerado por IA apresenta. Essas variáveis impactam o texto escrito em termos de sua densidade lexical por ser um indicador da complexidade da escrita (Halliday, 1985, 1987, 1989, 1993; Colombi, 2000; Martins, 2017), das diferenças entre oralidade e escrita (Halliday, 1985, 2005; Johansson, 2008; González Fernández, 2018), da linguagem científica (Halliday, 1993; Moohebat *et al.*, 2015; Riffo; Osuna; Lagos, 2019; Nasseri; Thompson, 2021), da proficiência linguística em uma língua adicional (Kondal, 2015; Nalesso, 2018; Schnur; Rubio, 2021) e da facilidade na leitura que um texto pode apresentar (Kembaren; Aswani, 2022). A densidade lexical refere-se à porcentagem de itens lexicais em um texto inferido pelo cálculo de "palavras de conteúdo" (substantivos, adjetivos e verbos e alguns advérbios de modo) em relação à totalidade de palavras presentes no texto (Ure, 1971).

Estudos relacionados à densidade lexical datam de longa data (Ure, 1971; Halliday, 1985; Johansson, 2008; Read, 2010) com utilização de corpora para fazer análises quantitativas por meio de programas computacionais (*Text Analyzer, Textalyser, RANGE* (Nation, 2013), *Lexicool, Microsoft Word, Online Utility, TAALED, Python, IBM SPSS, t-SNE*, por exemplo). Porém, não há estudos que investiguem um corpus resultante de atividades produzidas pelo ChatGPT (Transformador Generativo Pré-Treinado) com dados em diferentes línguas adicionais.

A característica principal do ChatGPT é a geração de grandes quantidades de dados ou textos, visto ter a possibilidade de gerá-los a partir de perguntas ou tarefas que lhe são solicitadas, fazendo ajustes, correções, inclusão de palavras e usando determinado gênero textual ou formato requerido de acordo com as instruções inseridas por humanos (Kumar, 2023; Kasneci *et al.*, 2023). Trata-se de um recurso de IA que possibilita a interação com o usuário para que, a partir de instruções, produza texto escrito semelhante ao que seria produzido por humanos (Perkins, 2023; King; ChatGPT, 2023; Rospigliosi, 2023; Anderson *et al.*, 2023). Um dos aspectos positivos, de acordo com Mitrović, Andreoletti e Ayoub (2023) e Ramos (2023), é que a ferramenta tende a produzir textos com registro escrito mais formal dado o uso de vocabulário menos coloquial.

Entretanto, apesar da aparente facilidade oferecida pelo ChatGPT com relação à escrita, as respostas podem apresentar textos sem coerência semântica (Perkins, 2023; Dale, 2021), baixa diversidade lexical (Gehrmann; Strobelt; Rush, 2019), vocabulário repetido (Dehouche, 2021; Fröhling; Zubiaga, 2021) e informações pouco confiáveis, com falta de exemplos reais (Kumar, 2023).

Não há, no entanto, estudos que investiguem o funcionamento do ChatPGT em termos de respostas ou textos gerados quando são inseridas as mesmas tarefas em diferentes línguas adicionais, usuários distintos, instruções específicas referentes aos níveis de proficiência do Quadro Comum Europeu de Referência para Línguas (CEFR)[1]. Considerando essa lacuna, este artigo busca responder às seguintes perguntas de pesquisa:

a. o ChatGPT produz dados seguindo as instruções no que diz respeito ao número de palavras solicitadas para as diferentes línguas adicionais e níveis de proficiência?
b. o ChatGPT produz dados que revelam densidade lexical distinta entre as línguas adicionais e os níveis de proficiência?
c. o ChatGPT produz dados que revelam densidade lexical distinta em uma mesma língua, considerando os níveis de proficiência?
d. o ChatGPT produz dados que demonstram correlação entre a densidade lexical e a extensão textual em cada nível de proficiência?

Para tanto, este estudo tem como objetivo compreender a densidade lexical em textos produzidos pelo ChatGPT e analisar a correlação dessa densidade nas línguas adicionais e nos níveis de proficiência previstos nas tarefas. As contribuições deste estudo trazem indicações sobre especificidades na natureza de textos que o ChatGPT produz nas línguas e respectivos níveis de proficiência, além de informações importantes sobre o uso dessa ferramenta de IA na produção de textos e na proposição

[1] Fonte: https://www.britishcouncil.org.br/quadro-comum-europeu-de-referencia-para-linguas-cefr. Acesso em: 03 jul. 2023.

de tarefas de escrita no ensino de línguas adicionais.

Este artigo está organizado em cinco seções, incluída esta introdução, além das referências e um anexo com as tarefas aplicadas para a geração de dados. A segunda seção aborda o conceito de densidade lexical; a terceira descreve o desenho do estudo e detalha a natureza dos dados; a quarta sistematiza os dados, analisando-os e destacando correlações e coeficientes e, finalmente, a quinta reporta os resultados gerais e contribuições.

## 2 Densidade lexical

A densidade lexical diz respeito à porcentagem de itens lexicais em um texto, inferida pelo cálculo de "palavras de conteúdo" (substantivos, adjetivos, verbos e alguns advérbios de modo) em relação à totalidade de palavras presentes no texto (Ure, 1971; Ure; Ellis, 1977; Halliday, 1985, 1993). Esse conceito é um indicativo de desenvolvimento e complexidade da escrita, diferenças entre oralidade e escrita, linguagem científica (Halliday, 1993) e proficiência linguística (Halliday, 1987, 1993, 1989; Colombi, 2000).

A investigação da densidade lexical, de acordo com Johansson (2008), tem sido orientada por diferentes perspectivas (Ure, 1971; Ure; Ellis, 1977; Halliday, 1985, 1993). A perspectiva de Ure (1971) é pela distinção entre palavras que têm propriedades lexicais (termos gramaticais ou que possuem uma função sintática-gramatical) daquelas que não possuem tal propriedade. Assim, o número total de palavras com propriedades lexicais, dividido pelo número total de palavras ortográficas, define o conceito de densidade lexical para esse autor. Halliday (1985, p. 65) sugere que as medidas se relacionam aos "padrões de distribuições de palavras em diferentes tipos de textos falados e escritos". A média que caracteriza a densidade lexical alta ou baixa está também ligada ao fato de o texto ser falado ou escrito. Ainda de acordo com Halliday (1985, p. 80), a "média típica do inglês falado está entre 1,5 e 2, enquanto o valor para o escrito fica entre 3 e 6, dependendo do quão formal é a escrita" e destaca que pode haver variações de acordo com os critérios adotados. Sob essa perspectiva, Johansson (2008) usa os seguintes porcentuais: para textos falados, densidade lexical inferior a 40% e, para textos escritos, densidade lexical de 40% ou mais.

Ure e Ellis (1977) afirmam que, tradicionalmente, substantivos, verbos e adjetivos são as três classes de palavras com propriedades lexicais. A esse respeito, Johansson (2008) destaca que esses itens são chamados de "palavras de conteúdo" por apresentarem a possibilidade de inclusão de novos itens, diferente das classes fechadas ou partes mais gramaticais.

Halliday (1985) retoma as discussões para esclarecer a diferença de função de um item lexical e de um item gramatical, visto que um item pode se constituir por mais de uma palavra, a exemplo de verbos auxiliares e modais. Johansson (2008) explicita essas diferenças, ao destacar:

> Um item lexical é definido por Halliday como um item que "funciona em conjuntos lexicais e não em sistemas gramaticais: isto é, eles entram em contrastes abertos e não fechados" (Halliday 1985: 63). O item lexical faz parte de um conjunto aberto, que pode ser contrastado com vários itens do mundo. Um item gramatical, por outro lado, entra em um sistema fechado, de acordo com Halliday (Johansson, 2008, p. 66).

O fundamento da visão de Halliday (1985) advém de sua compreensão do desenvolvimento da linguagem. No início do desenvolvimento linguístico, as construções oracionais das crianças se constituem em um *continuum*, preponderando substantivos, verbos e adjetivos, estes identificados como itens lexicais.

Em 1993[2], Halliday observa que a densidade lexical é uma medida ligada à densidade de informação de qualquer texto em termos de itens lexicais (palavras de conteúdo em relação às palavras gramaticais) presentes na estrutura gramatical e é medida pelo número de palavras lexicais em cada oração. A densidade de informação representa a noção de empacotamento, visto que um texto com alta proporção de palavras de conteúdo contém mais informação do que um texto com alta proporção

[2] Neste estudo, o foco de Halliday é o texto científico e suas características em termos de: definições entrelaçadas; taxonomias/classificações técnicas, expressões/termos específicos, densidade lexical, ambiguidade sintática, metáfora gramática e descontinuidade semântica (Halliday, 1993, p. 78).

de palavras funcionais (preposições, interjeições, pronomes, conjunções e palavras contáveis). Ainda para Halliday (1993), essa quantidade pode variar bastante de oração para oração, mas salienta que é possível verificar uma tendência de um *continuum* entre fala e escrita. Na escrita, a linguagem tende a ser mais planejada e formal, a densidade é alta e com tendência de ser infinitamente maior na escrita acadêmica. Em casos de densidade lexical muito alta, de acordo com Halliday (1993), a leitura pode ser mais difícil, tornando o texto pouco inteligível.

A densidade lexical tem relação com a diversidade e complexidade lexical/de vocabulário (Martins, 2017; González Fernández, 2018; Riffo; Osuna; Lagos, 2019), e a diversidade lexical é frequentemente usada como um equivalente à riqueza lexical (Johansson, 2008). No entanto, neste artigo, o foco é na densidade lexical para entender a complexidade da escrita produzida pelo ChatGPT. Assim, a seguir apresenta-se o desenho deste estudo que contempla dados quantitativos a respeito da densidade lexical e sua correlação entre línguas adicionais e níveis de proficiência.

## 3 Desenho do estudo

Neste estudo, a geração dos dados, os procedimentos técnicos e a abordagem analítica dessas informações caracterizam-se por uma pesquisa quantitativa (Dörnyei, 2007), visto priorizar uma amostra grande de dados, permitindo fazer algumas generalizações da densidade lexical dos textos produzidos nas diferentes línguas e associados aos níveis de proficiência previstos pelo CEFR. Este artigo também se qualifica como exploratório e descritivo; exploratório, porque o volume de dados das diferentes línguas e dos respectivos níveis de proficiência previstos em cada uma das atividades proporciona maior familiaridade com as ocorrências de densidade lexical; descritivo, porque permite descrever suas características e estabelecer possíveis correlações entre as variáveis (Gil, 2002).

Os participantes deste estudo são estudantes de uma universidade britânica que se dispuseram a participar da pesquisa a partir de um anúncio veiculado nas plataformas virtuais da própria universidade e *e-mails* pessoais. No referido anúncio, constava o link de acesso ao formulário (Microsoft Forms) para que os interessados pudessem obter informações a respeito da pesquisa, contendo o detalhamento dos objetivos e a natureza da participação (o que lhes era exigido fazer, línguas adicionais que a pesquisa contemplava, conhecimento inicial do ChatGPT; quando, onde e pró-labore a receber por sua participação de duas horas no laboratório de informática). Os interessados, então, preenchiam o formulário, davam ciência da data em que a tarefa deveria ser realizada e assinavam o termo de compromisso de ética na pesquisa[3].

Considerando que a pesquisa incluía tarefas em cinco línguas adicionais (alemão, espanhol, francês, italiano e português), planejadas de acordo com os níveis de proficiência previstos pelo CEFR, aos participantes selecionados foi solicitado simplesmente que copiassem as instruções de cada uma das seis tarefas previamente elaboradas pelos pesquisadores, em um dos cinco idiomas de cada vez, e as inserissem na plataforma IA (O Anexo A mostra as tarefas em português). Em seguida, copiavam a resposta produzida pelo ChatGPT para cada tarefa e a inseriam em um formulário (Microsoft Forms) específico ao idioma. Ao final, completaram cinco formulários com seis textos cada um, respectivamente. As tarefas foram produzidas a partir de 100 contas diferentes no ChatGPT e acompanhadas por, pelo menos, um dos pesquisadores em função das especificidades das atividades e em virtude de possíveis dúvidas que pudessem surgir e das dificuldades com a ferramenta de IA; o tempo máximo para a realização das atividades era de duas horas. Em geral, a plataforma funcionou sem problemas, e a exceção foi em nove casos nos quais não foi possível completar a tarefa individualmente, sendo estes textos excluídos. Entretanto, a margem de problemas foi bastante baixa, e, dos 3000 textos esperados, 2991 (99.7%) foram concluídos com sucesso e formam o *corpus* deste estudo, conforme a Tabela 1.

Portanto, na Tabela 1 verificou-se o número de textos após os seguintes ajustes: para a Língua Alemã, dois textos foram excluídos, um devido ao fato de ter copiado a atividade não indicada e outro por ter adicionado apenas as instruções; para a Língua Espanhola, em três textos, somente foram copiadas as instruções para a atividade; para a Língua Italiana, um texto continha apenas as instruções, semelhante ao que ocorreu com dois textos em Língua Portuguesa.

[3] Processo enviado ao comitê de ética na pesquisa da Universidade de Essex, número ETH2223-0987.

**Tabela 1.** Número de textos por línguas, distribuídos por níveis do CEFR.

| | Nível A1 | Nível A2 | Nível B1 | Nível B2 | Nível C1 | Nível C2 | Total |
|---|---|---|---|---|---|---|---|
| Alemão | 100 | 100 | 100 | 100 | 99 | 99 | 598 |
| Espanhol | 98 | 100 | 100 | 99 | 99 | 100 | 596 |
| Francês | 100 | 100 | 100 | 100 | 100 | 100 | 600 |
| Italiano | 100 | 100 | 100 | 100 | 100 | 99 | 599 |
| Português | 100 | 100 | 100 | 98 | 100 | 100 | 598 |

*Fonte:* Os autores.

Para a análise dos dados, os critérios foram organizados de acordo com as perguntas de pesquisa elencadas na introdução deste artigo e um conjunto de ferramentas foi utilizado: os textos dos formulários (Microsoft Forms) completados pelos participantes do estudo foram inicialmente baixados no formato de planilhas de Excel e os textos foram preparados ("limpos") manualmente para a análise, usando o formato CSV no Excel. Em seguida, cada arquivo, organizado por língua e por nível, foi submetido a uma análise lexical usando o programa Python por meio dos ambientes de execução Google Colaboratory (referido normalmente como Colab) e Jupyter notebook, e o ChatGPT. Por fim, os dados já organizados e classificados foram transpostos manualmente para planilhas no Excel e submetidos ao tratamento analítico com uso de fórmulas para análise de dados no próprio Excel e no software SPSS para análise estatística.

## 4 Apresentação dos dados e análise dos resultados

Os textos gerados pelo ChatGPT resultaram em número significativo de palavras, considerando as variáveis: língua adicional e nível de proficiência de acordo com o CEFR. Para fins de análise, o primeiro passo foi fazer uma verificação detalhada para realizar uma limpeza dos dados (Schnur; Rubio, 2021), sendo excluídos trechos de instrução ou informações repetidas.

Visto que o número de palavras é um dos fatores principais usados no cálculo da densidade, iniciou-se a análise pela pergunta de pesquisa: *o ChatGPT produz textos seguindo as instruções no que diz respeito ao número de palavras solicitadas e há diferenças entre as línguas e os níveis de proficiência?* Em todas as tarefas, as instruções solicitavam a produção do texto pela ferramenta, usando um número mínimo e máximo de palavras. As Figuras 1 a 6 trazem essas ocorrências em cada uma das línguas e níveis de proficiência, respectivamente.

Os resultados mostram que a ferramenta de IA não segue as orientações das tarefas e sugerem uma tendência inversa ao que se esperaria de textos requeridos em níveis mais avançados de proficiência. Para os níveis iniciais, a ferramenta excede o número de palavras solicitadas e, inversamente, para os níveis finais, produz textos com número de palavras aquém do previsto.

A diferença de proporcionalidade no produto final do texto pode ser mais bem visualizada nas Tabelas 2 e 3 subsequentes. A Tabela 2 mostra que o percentual de textos com número acima do solicitado se concentra nos níveis iniciais.

**Tabela 2.** Percentual de textos com palavras acima do esperado.

| | Nível A1 | Nível A2 | Nível B1 | Nível B2 | Nível C1 | Nível C2 |
|---|---|---|---|---|---|---|
| Alemão | 99% | 84% | 83% | 78% | 33% | 0% |
| Espanhol | 95% | 88% | 97% | 88% | 45% | 3% |
| Francês | 100% | 98% | 99% | 99% | 63% | 1% |
| Italiano | 94% | 73% | 89% | 83% | 19% | 0% |
| Português | 89% | 76% | 96% | 86% | 21% | 0% |

*Fonte:* Os autores.

A tendência da ferramenta de IA foi de produzir os textos solicitados sem contemplar todos os

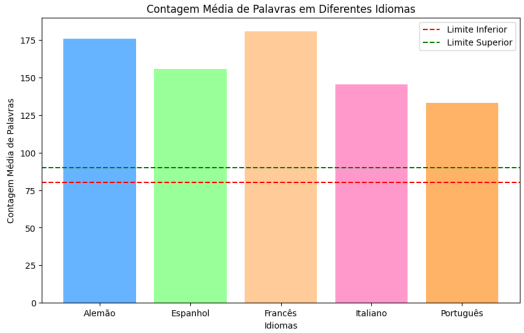


**Figura 1.** Média de palavras nas línguas e níveis de proficiência: Nível A1.

*Fonte:* Os autores.

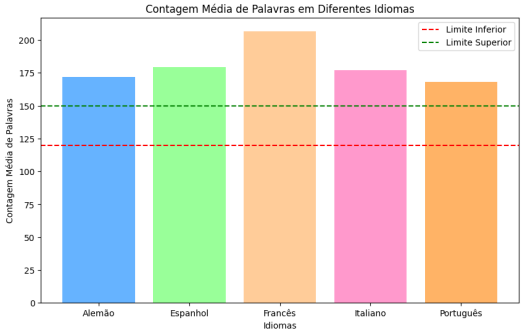


**Figura 2.** Média de palavras nas línguas e níveis de proficiência: Nível A2.

*Fonte:* Os autores.

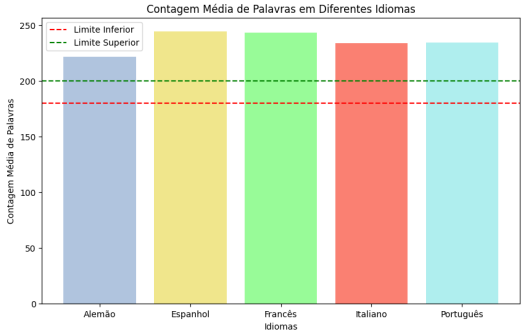


**Figura 3.** Média de palavras nas línguas e níveis de proficiência: Nível B1.

*Fonte:* Os autores.

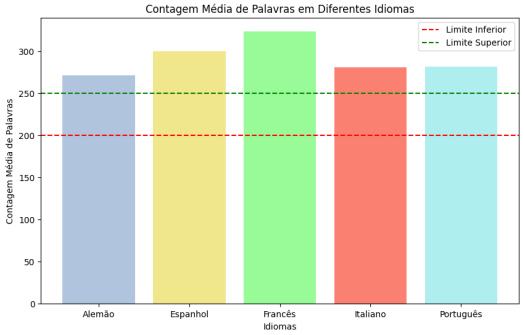


**Figura 4.** Média de palavras nas línguas e níveis de proficiência: Nível B2.

*Fonte:* Os autores.

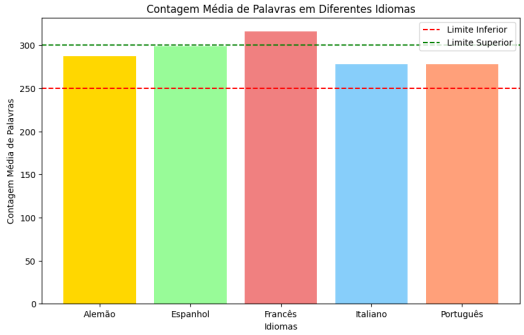


**Figura 5.** Média de palavras nas línguas e níveis de proficiência: Nível C1.

*Fonte:* Os autores.

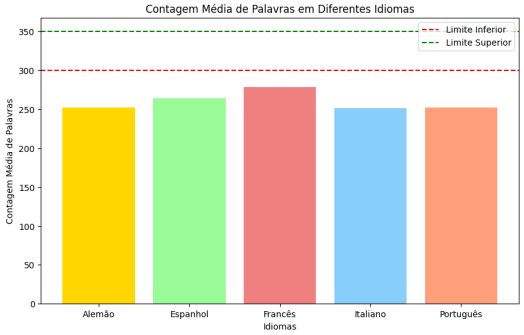


**Figura 6.** Média de palavras nas línguas e níveis de proficiência: Nível C2.

*Fonte:* Os autores.

aspectos das instruções inseridas no espaço previsto em sua plataforma. Esse resultado é reiterado nos dados apresentados na Tabela 3, observando-os de outra perspectiva, ou seja, o número de palavras aquém do solicitado se concentra nos níveis finais. Confirma-se, assim, a tendência de não seguir todas as instruções solicitadas.

**Tabela 3.** Percentual de textos com palavras abaixo do esperado.

| | Nível A1 | Nível A2 | Nível B1 | Nível B2 | Nível C1 | Nível C2 |
|---|---|---|---|---|---|---|
| Alemão | 0% | 0% | 1% | 0% | 7% | 96% |
| Espanhol | 2% | 0% | 0% | 1% | 3% | 82% |
| Francês | 0% | 0% | 0% | 0% | 0% | 72% |
| Italiano | 1% | 3% | 1% | 0% | 16% | 90% |
| Português | 0% | 1% | 0% | 0% | 19% | 93% |

*Fonte:* Os autores.

Olhando os dois movimentos apresentados pelo ChatGPT conjuntamente, desde o ponto de vista do percentual de textos com número de palavras aquém do esperado e acima do solicitado, respectivamente, a Figura 7 ilustra claramente esse produto textual e revela um movimento inverso, se comparados os níveis iniciais e finais.

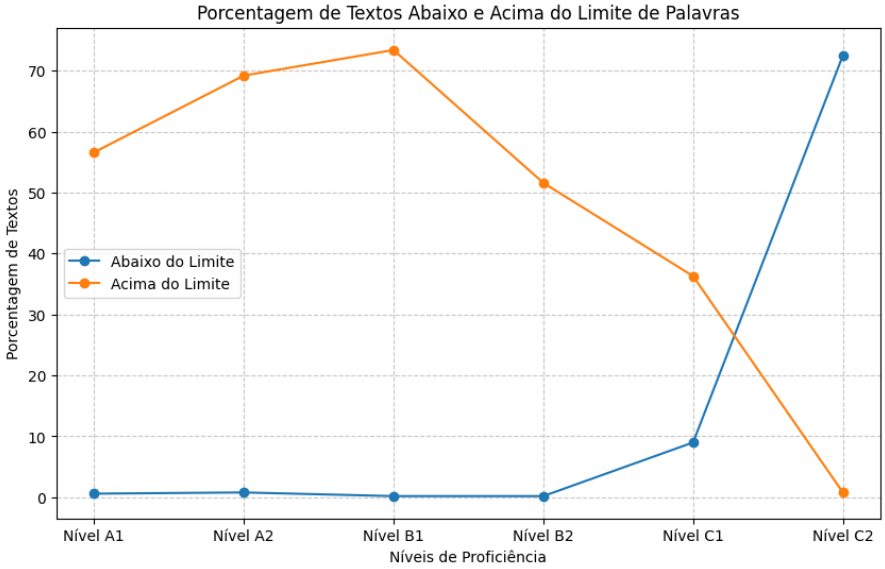


**Figura 7.** Percentual de textos com número de palavras distinto do esperado
*Fonte:* Os autores.

Considerando-se as línguas adicionais contempladas neste estudo, selecionou-se uma segunda pergunta orientadora na análise dos dados, qual seja: *O ChatGPT produz dados que revelam densidade lexical distinta entre as línguas adicionais e níveis de proficiência?*. Sistematizaram-se os resultados na Tabela 4.

Para melhor visualizar os resultados em termos percentuais, a Figura 8 ilustra esse comportamento revelado pela IA na produção escrita.

Os resultados na Figura 8 sugerem uma tendência semelhante no movimento da densidade lexical em todas as línguas e níveis. Porém, os textos do nível A2 apresentaram densidade lexical mais alta do que se espera para este nível, contrariando a literatura da área a qual indica que esse percentual de densidade representaria a escrita de textos em níveis mais avançados de proficiência (Gregori-Signes; Clavel-Arroitia, 2015; Clavel-Arroitia; Pennock-Speck, 2021). A exceção foi para os dados no alemão para o nível B2, que apresentou percentual um pouco mais alto. Um segundo aspecto é a queda brusca no nível B1 do alemão, indicando uma diferença entre as línguas, e essa tendência se verifica também em espanhol e português. Portanto, os resultados permitem que se afirme haver diferenças

**Tabela 4.** Densidade lexical por línguas e níveis.

| | Nível A1 | Nível A2 | Nível B1 | Nível B2 | Nível C1 | Nível C2 |
|---|---|---|---|---|---|---|
| Alemão | 41.57 | 44.16 | 37.74 | 45.14 | 40.64 | 43.89 |
| Espanhol | 48.00 | 50.67 | 44.03 | 46.91 | 44.45 | 47.45 |
| Francês | 42.67 | 48.18 | 47.28 | 48.38 | 43.28 | 48.09 |
| Italiano | 46.51 | 49.51 | 49.47 | 49.53 | 45.05 | 48.78 |
| Português | 50.83 | 54.31 | 46.00 | 50.96 | 47.71 | 50.63 |

*Fonte:* Os autores.

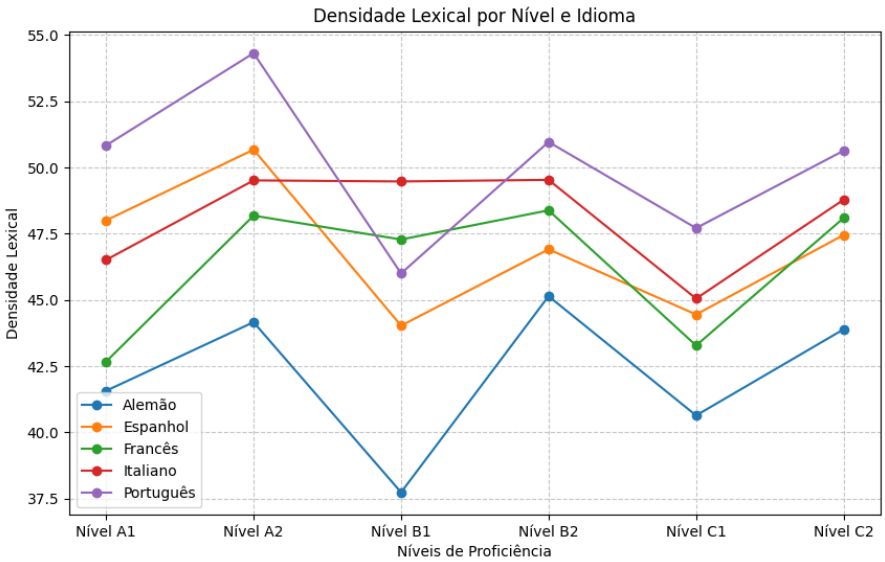


**Figura 8.** Densidade lexical por nível e idioma.
*Fonte:* Os autores.

nesse nível entre as línguas, conforme indicado pelo texto gerado para o nível B1, embora para duas línguas haja um movimento mais estável entre os níveis A2 e B2 (italiano e francês). Por outro lado, há uma semelhança entre as línguas com relação ao percentual, sugerindo que a ferramenta parece optar por certa formalidade do texto escrito. Esse resultado é corroborado no estudo de González Fernández (2018), que analisou um corpus usando mensagens escritas no Twitter e destacou que a densidade lexical varia em um *continuum* entre oralidade e escrita, sendo mais baixa quando o texto possui marcas de oralidade.

Ao observar os resultados em uma mesma língua, ou seja, se *o ChatGPT produz dados que revelam densidade lexical distinta em uma mesma língua considerando os níveis de proficiência*, os resultados ilustrados na Figura 8 sugerem haver uma distinção em todas as línguas. Se comparadas as tarefas relativas ao nível A1 e C2, que se relacionam aos níveis de proficiência, a expectativa era de que haveria um percentual mais alto no C2; no entanto, não foi o que se concretizou nos dados, e os resultados confirmam que a ferramenta não consegue perceber esses detalhes do ponto de vista linguístico.

Por fim, para entender melhor os resultados, procedeu-se à análise orientada pela quarta pergunta de pesquisa: *o ChatGPT produz dados que demonstram correlação entre a densidade lexical e a extensão textual em cada nível de proficiência?* Para responder a esta pergunta, recorreu-se a uma análise estatística usando o SPSS. Devido à quantidade de textos, 2991, e à natureza da análise, estes foram agrupados de acordo com os níveis do CEFR (498 textos no nível A1, 500 textos no nível A2, 500 textos no nível B1, 497 textos no nível B2, 498 textos no nível C1, e 498 textos no nível C2, respectivamente), visando a um tratamento estatístico mais significativo. A hipótese inicial é que a densidade aumentaria proporcionalmente aos níveis de proficiência e ao número de palavras.

Para tanto, foram utilizados dados para contemplar duas variáveis que respondem a esta pergunta de pesquisa: qual é a densidade lexical e o número total de palavras de cada texto? Ambas as variáveis foram calculadas por meio do uso do Python nos ambientes Google Colab e Jupyter notebook e em planilhas de Excel antes de serem submetidas a uma análise detalhada em SPSS.

No SPSS, adotaram-se duas funções para analisar a correlação entre a extensão textual e a densidade lexical: a *correlação bivariada* e a *regressão linear*, a qual se subdivide em quatro itens de análise: variáveis inseridas/removidas, resumo do modelo, ANOVA e coeficientes. Dentre os resultados derivados dessa análise detalhada, três coeficientes são significativos para este estudo e serão discutidos a seguir: coeficiente de correlação, coeficiente valor de t (*t-value*)/Constante e o coeficiente valor de t (*t-value*)/*TotaldePalavras*. A Tabela 5 elenca os resultados dos três coeficientes de acordo com cada nível do CEFR.

**Tabela 5.** Resumo dos Coeficientes por nível.

| | Coeficiente de correlação | Coeficiente valor de t (t-value) - Constante | Coeficiente valor de t (t-value) - TotaldePalavras |
|---|---|---|---|
| Nível A1 | -0.544 | $< 0.001$ | $< 0.001$ |
| Nível A2 | -0.142 | $< 0.001$ | 0.001 |
| Nível B1 | 0.124 | $< 0.001$ | 0.006 |
| Nível B2 | 0.011 | $< 0.001$ | 0.808 |
| Nível C1 | -0.179 | $< 0.001$ | $< 0.001$ |
| Nível C2 | -0.039 | $< 0.001$ | 0.389 |

*Fonte:* Os autores.

Os valores são também visualizados na Figura 9.

Os resultados dos coeficientes em cada nível permitem estabelecer se há uma correlação entre o número de palavras e a densidade lexical. Para cada nível, são mostrados os resultados sob o formato de Tabelas seriadas no SPSS. Primeiro, o foco de análise é nas tabelas de *Correlações* para entender o coeficiente de correlação nos seis níveis do CEFR e, em seguida, as tabelas de Coeficientes para explicar os coeficientes valor de t *Constante* e *TotaldePalavras*.

O coeficiente de correlação advém do primeiro passo da análise em SPSS, ilustrada na Figura 10 do Nível A1 e detalhada a seguir. Procedimento similar foi aplicado aos demais níveis.

A verificação da existência de uma correlação entre o número total de palavras e a densidade lexical e seus elementos oferece diferentes informações: a correlação de Pearson (*Pearson's r*) estabelece a correlação entre as duas variáveis *DensidadeLexical* e *TotaldePalavras*, a qual é de –0.544 nesse caso, e quantifica a força e a direção linear entre as duas variáveis. Essa correlação negativa indica que, à medida que o número de palavras aumenta, a densidade lexical diminui, e vice-versa. No que concerne ao Nível de Significância (*Sig.*), o valor de p (*p-value*), nesse caso $<$ 0,001, indica a probabilidade de a correlação entre as duas variáveis ter ocorrido por acaso ou não. No presente caso, é improvável ter sido por acaso e, portanto, é considerada significativa. Por último, o elemento N indica o tamanho da amostra (498 textos).

No nível A1, os resultados sugerem uma correlação negativa forte entre o número total de palavras em um texto e sua densidade lexical, e essa relação é estatisticamente significativa. À medida que os textos se tornam mais longos (com mais palavras), eles tendem a ter uma densidade lexical menor e, inversamente, textos mais curtos tendem a ter uma densidade lexical maior. Isso poderia implicar que textos mais longos podem incluir palavras redundantes ou menos informativas, levando a uma densidade lexical menor. Entretanto, isso demandaria uma análise qualitativa que não está no escopo deste estudo.

No nível A2, os resultados (cf. Figura 11) mostram que o coeficiente da correlação de Pearson entre as duas variáveis é aproximadamente –0.142, o que sugere uma correlação negativa fraca entre as duas variáveis. Como ocorreu com os textos no nível A1, à medida que o número de palavras no texto aumenta, a densidade lexical tende a diminuir, e vice-versa. A correlação é significativa no nível 0,01 (2 extremidades) e indica que a correlação é significante estatisticamente em um alto nível de

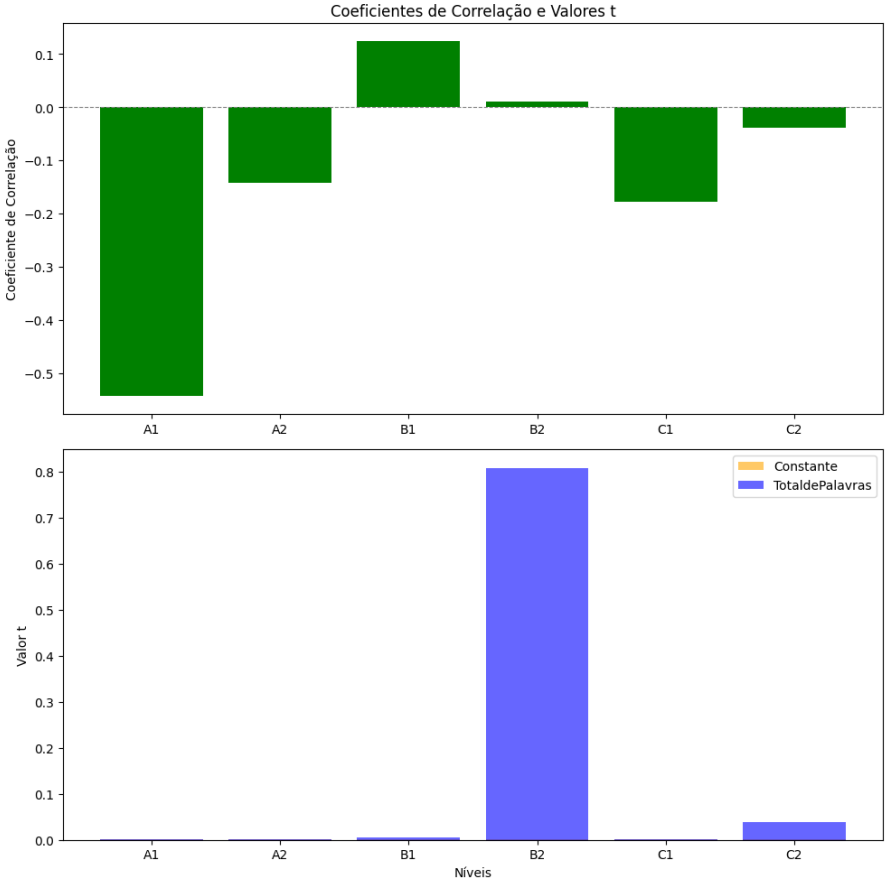


**Figura 9.** Coeficientes de Correlação e Valores de t (t-value).
*Fonte:* Os autores.

**Correlações**

| | | TotaldePalavras | DensidadeLexical |
|---|---|---|---|
| TotaldePalavras | Correlação de Pearson | 1 | -.544** |
| | Sig. (2 extremidades) | | <.001 |
| | N | 498 | 498 |
| DensidadeLexical | Correlação de Pearson | -.544** | 1 |
| | Sig. (2 extremidades) | <.001 | |
| | N | 498 | 498 |

**. A correlação é significativa no nível 0,01 (2 extremidades).

**Figura 10.** Correlações Nível A1.
*Fonte:* Os autores.

confiança e não ocorreu por acaso. No entanto, é importante frisar que, apesar de ser estatisticamente significativa, a força da correlação entre as duas variáveis é relativamente fraca.

No nível B1, evidenciados pela Figura 12, o coeficiente de correlação de Pearson é aproximadamente 0,124. Como ocorreu nos níveis A1 e A2, a correlação é considerada estatisticamente significativa no nível de significância de 0,01 (bilateral), ou seja, é improvável de ter ocorrido por acaso. O coeficiente de correlação 0,124 sugere uma correlação positiva fraca entre as variáveis *TotaldePalavras* e *DensidadeLexical* e indica que, conforme o número total de palavras em um texto aumenta, a densidade lexical tende a aumentar ligeiramente, e vice-versa, diferente do que ocorreu nos níveis A1 e A2, respectivamente. No entanto, é importante salientar que, embora estatisticamente significativa, a força da correlação é relativamente fraca e sugere que outros fatores também podem influenciar a relação entre essas variáveis, como enfatizado anteriormente sobre o possível uso de palavras redundantes ou menos informativas.

**Correlações**

| | | TotaldePalavras | DensidadeLexical |
|---|---|---|---|
| TotaldePalavras | Correlação de Pearson | 1 | -.142** |
| | Sig. (2 extremidades) | | .001 |
| | N | 500 | 500 |
| DensidadeLexical | Correlação de Pearson | -.142** | 1 |
| | Sig. (2 extremidades) | .001 | |
| | N | 500 | 500 |

**. A correlação é significativa no nível 0,01 (2 extremidades).

**Figura 11.** Correlações Nível A2.
*Fonte:* Os autores.

**Correlações**

| | | TotaldePalavras | DensidadeLexical |
|---|---|---|---|
| TotaldePalavras | Correlação de Pearson | 1 | .124** |
| | Sig. (2 extremidades) | | .006 |
| | N | 500 | 500 |
| DensidadeLexical | Correlação de Pearson | .124** | 1 |
| | Sig. (2 extremidades) | .006 | |
| | N | 500 | 500 |

**. A correlação é significativa no nível 0,01 (2 extremidades).

**Figura 12.** Correlações Nível B1.
*Fonte:* Os autores.

Os resultados evidenciados na Figura 13, do nível B2, indicam que o coeficiente de correlação entre as duas variáveis é de aproximadamente 0,011, o que sugere uma correlação positiva muito fraca. Isso quer dizer que praticamente não há relação linear entre essas variáveis. Ademais, o valor de significância (*Sig.*) elevado, 0,808, indica que a correlação observada não é estatisticamente significativa de acordo com os níveis de significância convencionais (por exemplo, $\alpha = 0{,}05$). Ou seja, diferente dos níveis anteriores, a correlação entre as duas variáveis provavelmente se deve ao acaso e não possui significado prático ou relevante estatisticamente para este estudo.

**Correlações**

| | | TotaldePalavras | DensidadeLexical |
|---|---|---|---|
| TotaldePalavras | Correlação de Pearson | 1 | .011 |
| | Sig. (2 extremidades) | | .808 |
| | N | 497 | 497 |
| DensidadeLexical | Correlação de Pearson | .011 | 1 |
| | Sig. (2 extremidades) | .808 | |
| | N | 497 | 497 |

**Figura 13.** Correlações Nível B2.
*Fonte:* Os autores.

No nível C1, a correlação entre as variáveis *TotaldePalavras* e *DensidadeLexical* é estatisticamente significativa no nível de 0,01 (duas extremidades), sugerindo que há um alto grau de confiança no resultado (Figura 14). O coeficiente de correlação é de aproximadamente -0,179 e indica haver uma correlação negativa moderada entre as duas variáveis. Isso quer dizer que, à medida que o total de palavras em um texto aumenta, a densidade lexical tende a diminuir e o sinal negativo do coeficiente transparece uma relação inversa entre as variáveis. O baixo valor de p (*Sig.*), menor que 0,001, prediz que é bastante improvável que a correlação observada tenha ocorrido por acaso e é, portanto, relevante para este estudo.

**Correlações**

| | | TotaldePalavras | DensidadeLexical |
|---|---|---|---|
| TotaldePalavras | Correlação de Pearson | 1 | -.179** |
| | Sig. (2 extremidades) | | <.001 |
| | N | 498 | 498 |
| DensidadeLexical | Correlação de Pearson | -.179** | 1 |
| | Sig. (2 extremidades) | <.001 | |
| | N | 498 | 498 |

**. A correlação é significativa no nível 0,01 (2 extremidades).

**Figura 14.** Correlações Nível C1.
*Fonte:* Os autores.

Por último, no nível C2 (Figura 15), os resultados indicam que o coeficiente de correlação entre as variáveis *TotaldePalavras* e *DensidadeLexical* é -0.039, sugerindo haver uma correlação negativa muito fraca, ou seja, há pouca ou nenhuma relação linear entre essas duas variáveis. O valor de p (0,389) é considerado relativamente alto, externando que a correlação poderia ser explicada razoavelmente pelo acaso, mas que a mesma não é estatisticamente significativa. Portanto, no nível C2, assim como ocorreu em B2, parece não haver uma correlação significativa ou relevante entre as duas variáveis.

**Correlações**

| | | TotaldePalavras | DensidadeLexical |
|---|---|---|---|
| TotaldePalavras | Correlação de Pearson | 1 | -.039 |
| | Sig. (2 extremidades) | | .389 |
| | N | 498 | 498 |
| DensidadeLexical | Correlação de Pearson | -.039 | 1 |
| | Sig. (2 extremidades) | .389 | |
| | N | 498 | 498 |

**Figura 15.** Correlações Nível C2.
*Fonte:* Os autores.

Para confirmar a análise inicialmente feita usando a função correlação, o segundo passo foi usar a função regressão linear. Dentre as subdivisões desta função, é trazida para esta discussão a chamada Coeficientes, a qual se subdivide em dois coeficientes distintos: valor de t (*t-value*)/Constante e coeficiente valor de t (*t-value*)/*TotaldePalavras*. A mesma análise foi conduzida em todos os níveis e é ilustrada na Figura 16, utilizando o resultado do nível A1.

**Coeficientes[a]**

| Modelo | | Coeficientes não padronizados B | Coeficientes não padronizados Erro Erro | Coeficientes padronizados Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constante) | 56.170 | .730 | | 76.983 | <.001 |
| | TotaldePalavras | -.065 | .004 | -.544 | -14.427 | <.001 |

a. Variável Dependente: DensidadeLexical

**Figura 16.** Coeficientes Nível A1.
*Fonte:* Os autores.

A Tabela *Coeficientes* contém informações sobre os coeficientes do modelo de regressão, os quais ajudam, neste estudo, a compreender as relações entre as chamadas *variável preditora* (*TotaldePalavras*) e a *variável dependente* (*DensidadeLexical*). O coeficiente padronizado (Beta) destaca a importância relativa de cada variável preditora na explicação da variabilidade na variável dependente. O valor de t indica quantos erros padrão o coeficiente está afastado de zero, e o coeficiente padronizado oferece uma ideia do tamanho do efeito em termos de desvio padrão. Ambos o *valor de t* e o

*valor de significância* (*Sig.*) assinalam a significância estatística de cada coeficiente.

No que diz respeito à regressão no Nível A1, portanto, o coeficiente negativo para a variável preditora *TotaldePalavras* sugere que, à medida que o número total de palavras em um texto aumenta, a densidade lexical tende a diminuir. Ambos os coeficientes são estatisticamente significativos (<0,001), mostrando que as relações dificilmente ocorreram por acaso. Assim, esses resultados confirmam aqueles obtidos na análise do coeficiente de correlação na primeira parte da análise. O mesmo ocorre no Nível A2 (Figura 17), em que o coeficiente negativo para a variável preditora *TotaldePalavras* também confirma que, à medida que o número total de palavras de um texto aumenta, a densidade lexical tende a diminuir ligeiramente. Ambos os coeficientes são estatisticamente significativos (p < 0,001 para *Constante* e p = 0,001 para *TotaldePalavras*) e sinalizam que as relações dificilmente ocorreram por acaso.

**Coeficientes**[a]

| Modelo | | Coeficientes não padronizados | | Coeficientes padronizados | | |
|---|---|---|---|---|---|---|
| | | B | Erro Erro | Beta | t | Sig. |
| 1 | (Constante) | 52.976 | 1.139 | | 46.526 | <.001 |
| | TotaldePalavras | -.020 | .006 | -.142 | -3.211 | .001 |

a. Variável Dependente: DensidadeLexical

**Figura 17.** Coeficientes Nível A2.
*Fonte:* Os autores.

Os resultados obtidos no Nível B1 (Figura 18) também corroboram os obtidos pelo cálculo do coeficiente de correlação na primeira parte da análise. O coeficiente positivo para a variável preditora *TotaldePalavras* indica que, à medida que o número total de palavras em um texto aumenta, a densidade lexical tende a aumentar ligeiramente. Ambos os coeficientes são estatisticamente significativos (p < 0,001 para *Constante* e p = 0,006 para *TotaldePalavras*), preconizando que as relações dificilmente ocorreram por acaso.

**Coeficientes**[a]

| Modelo | | Coeficientes não padronizados | | Coeficientes padronizados | | |
|---|---|---|---|---|---|---|
| | | B | Erro Erro | Beta | t | Sig. |
| 1 | (Constante) | 39.149 | 2.077 | | 18.851 | <.001 |
| | TotaldePalavras | .024 | .009 | .124 | 2.785 | .006 |

a. Variável Dependente: DensidadeLexical

**Figura 18.** Coeficientes Nível B1.
*Fonte:* Os autores.

Já no Nível B2 (Figura 19), o coeficiente positivo para a variável preditora *TotaldePalavras* denota haver uma associação positiva entre o número total de palavras e a densidade lexical, mas o efeito é extremamente pequeno. Já o valor de p não significativo (0,808) para a variável preditora *TotaldePalavras* revela que a relação entre esse preditor e a variável dependente não é estatisticamente significativa, confirmando o que foi sugerido pela análise do coeficiente de correlação na primeira parte da análise.

No nível C1 (Figura 20), o coeficiente negativo para a variável preditora *TotaldePalavras* sinaliza haver uma associação negativa entre o número total de palavras e a densidade lexical. Ou seja, à medida que o número de palavras de um texto aumenta, a densidade lexical tende a diminuir, corroborando o que foi verificado na análise do coeficiente de correlação na primeira parte da análise. O alto valor absoluto de t (*t-value*) (-4,045) e o valor de p muito baixo (menor que 0,001) para a variável preditora *TotaldePalavras* indicam que a relação entre esta e a variável dependente é estatisticamente significativa.

Por fim, no nível C2 (Figura 21), o valor de t (t-value) e o valor associado de p oferecem informação sobre a significância estatística da variável preditora *TotaldePalavras*. Nesse caso, o valor de p para

**Coeficientes**[a]

| Modelo | | Coeficientes não padronizados B | Coeficientes não padronizados Erro Erro | Coeficientes padronizados Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constante) | 47.934 | .995 | | 48.155 | <.001 |
| | TotaldePalavras | .001 | .003 | .011 | .243 | .808 |

a. Variável Dependente: DensidadeLexical

**Figura 19.** Coeficientes Nível B2.
*Fonte:* Os autores.

**Coeficientes**[a]

| Modelo | | Coeficientes não padronizados B | Coeficientes não padronizados Erro Erro | Coeficientes padronizados Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constante) | 48.855 | 1.150 | | 42.467 | <.001 |
| | TotaldePalavras | -.016 | .004 | -.179 | -4.045 | <.001 |

a. Variável Dependente: DensidadeLexical

**Figura 20.** Coeficientes Nível C1.
*Fonte:* Os autores.

*TotaldePalavras*, 0,389, é considerado relativamente alto e sugere que a relação entre as duas variáveis pode não ser estatisticamente significativa e, portanto, confirma o resultado da análise do coeficiente de correlação. Ademais, o baixo coeficiente padronizado (Beta) (-0.039) e o alto valor de p sugerem que a variável preditora não parece ter um impacto substancial na variável dependente.

**Coeficientes**[a]

| Modelo | | Coeficientes não padronizados B | Coeficientes não padronizados Erro Erro | Coeficientes padronizados Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constante) | 48.732 | 1.119 | | 43.558 | <.001 |
| | TotaldePalavras | -.004 | .004 | -.039 | -.862 | .389 |

a. Variável Dependente: DensidadeLexical

**Figura 21.** Coeficientes Nível C2.
*Fonte:* Os autores.

Apresentados os dados e indicados os resultados, orientados pelas perguntas de pesquisa, na próxima seção são discutidos esses resultados e indicadas as contribuições deste estudo.

## 5 Discussão e considerações finais

Este estudo discutiu a escrita em línguas adicionais para compreender o impacto que as ferramentas digitais exercem na produção textual. Para tanto, analisou a densidade lexical na perspectiva da LSF (Halliday, 1985, 1993) em um *corpus* resultante de seis tarefas produzidas pelo ChatGPT em cinco idiomas (alemão, espanhol, francês, italiano e português) e comparou a densidade lexical entre essas línguas e níveis de proficiência para identificar se haveria um padrão que indicasse complexidade textual.

A análise foi sistematizada em tabelas e figuras para visualizar os resultados, considerando-se as línguas adicionais e respectivos níveis de proficiência. Evidenciaram-se os valores da densidade lexical em cada língua (cf. Figuras 1 a 6 e Tabelas 2 e 3), para verificar o quão consistentes foram esses valores.

Com relação à primeira pergunta de pesquisa (cf. Figuras 1 a 6, Tabelas 2 e 3 e Figura 7), os resultados mostraram que a ferramenta de IA não é sensível às especificidades das orientações indicadas para a realização da tarefa nem à variável proficiência. O ChatGPT produziu textos com extensão diversa ao solicitado, não considerando os níveis de proficiência indicados. Portanto, é um

resultado incongruente ao esperado na produção de textos em uma língua adicional por humanos.

Os resultados relativos às segunda e terceira perguntas deste estudo mostraram que o ChatGPT produziu dados que seguem um mesmo padrão, independente da língua adicional. Esse padrão foi verificado na extensão dos textos; no entanto, entre os níveis de proficiência, observou-se uma diferença para os textos do nível B1. Esse padrão é reiterado quando se verifica o que acontece em uma mesma língua e níveis de proficiência. Em outras palavras, independente da língua adicional, o ChatGPT não produz textos que revelam diferenças marcantes entre os níveis iniciais e finais, respectivamente.

Para reforçar os resultados revelados pelos dados estatísticos, análises de correlações e coeficientes foram realizadas com o propósito de verificar se o "*ChatGPT produz dados que demonstram correlação entre a densidade lexical e a extensão textual em cada nível de proficiência*" (4a. pergunta de pesquisa). Os resultados dos coeficientes revelaram a seguinte correlação entre o número de palavras e a densidade lexical: para o nível A1, uma não correlação negativa forte; para o nível A2, negativa fraca; para o nível B1, correlação positiva fraca, diferenciando-se dos dois níveis anteriores e confirmando a diferença observada em termos de percentuais de densidade lexical. Por sua vez, no nível B2, verificou-se uma correlação positiva muito fraca; no nível C1, correlação negativa moderada; no nível C2, correlação negativa muito forte.

Para a confirmação desses resultados, empreendeu-se uma análise utilizando a função regressão linear. Os resultados mostraram coeficiente negativo para o coeficiente de valor t (*t-value*), *TotaldePalavras* no nível A1, reiterando a análise anterior; o nível A2 revelou coeficiente negativo de valor t (*t-value*)/*TotaldePalavras*; o nível B1 confirmou o resultado indicado na Figura 12, cujos coeficientes são estatisticamente positivos para o valor de t (*t-value*)/*TotaldePalavras*; o nível B2 revelou coeficiente positivo, associação positiva entre as variáveis; o nível C1 mostrou coeficiente negativo para a variável preditora *TotaldePalavras* com associação negativa entre o número total de palavras e a densidade lexical. Finalmente, o nível C2 exibiu significância relativamente alta, confirmando o que foi sugerido na análise do coeficiente de correlação.

Portanto, os resultados permitem que se conclua que, a partir da comparação entre as diferentes línguas adicionais, a ferramenta de IA não é sensível às características do sistema linguístico como, por exemplo, línguas adicionais mais desinenciais em relação às que não o são, pois as diferenças em termos percentuais são moderadas. Além disso, não se observou, com base nos padrões verificados, haver alguma língua, dentre as cinco, que revelasse um padrão muito diferente das demais. Por outro lado, levando-se em conta a variável nível de proficiência, obtiveram-se dois resultados inesperados: o nível A2, que se diferencia dos demais, e o nível C1, que apresenta um decréscimo quanto à densidade lexical em todas as línguas, resultado que difere daqueles que discutem proficiência linguística em contexto de língua adicional (Kondal, 2015; Schnur; Rubio, 2021).

Esses resultados divergem das observações feitas no estudo de Lancaster (2023), que salienta que o ChatGPT responde de maneira pré-determinada com base em sua programação (ou modelo criado), visto que, se a mesma entrada for inserida, a tendência da ferramenta é dar as mesmas respostas. A divergência diz respeito ao fato de a ferramenta produzir textos com extensão distinta entre as línguas adicionais, níveis de proficiência e densidade lexical, embora as especificidades do sistema linguístico de cada língua poderiam ser determinantes. A esse respeito, os dados requereriam um tratamento qualitativo para cada um dos textos do *corpus*.

Os resultados também dão indicações de que a ferramenta precisa ser mais bem refinada para levar em conta as características linguísticas de cada língua adicional e produzir textos escritos que tenham dados confiáveis que possam revelar a proficiência linguística em uso e real se comparada à produção escrita sem auxílio da IA. Nesse sentido, o uso da ferramenta em contexto de ensino precisa acontecer com parcimônia.

A contribuição deste estudo, a exemplo do que já sinalizou Lancaster (2023), indica que o ChatGPT é parte do desenvolvimento tecnológico que precisa se coadunar com o desenvolvimento educacional para que estudantes ou usuários da ferramenta compreendam: o uso tem implicações éticas e limitações, não sendo uma solução mágica para produzir textos escritos; os vários campos do conhecimento possuem exigências diferentes, como é o caso do ensino de línguas adicionais e a proficiência escrita nessas línguas para não gerar dados imprecisos; a possibilidade de questionar informações. Por fim,

quanto à densidade lexical, este artigo é pioneiro em abordar o tema nesse contexto, e a principal contribuição é compreender a complexidade dos textos que são produzidos por essa ferramenta.

Futuras pesquisas são necessárias com textos produzidos pelo ChatGPT para compreender melhor a qualidade e a complexidade da escrita em termos da organização dos complexos oracionais, da complexidade sintática e modalidade e avaliatividade sob a perspectiva da LSF. Além disso, pesquisas que analisem os textos qualitativamente e relacionem a qualidade da escrita em termos de níveis de proficiência de acordo com o CEFR serão bem-vindas. Por fim, pesquisas que explicitem de que forma os textos produzidos pelo ChatGPT podem levantar questões pedagógicas e suas possíveis contribuições para o desenvolvimento do currículo no ensino de línguas adicionais são desejáveis.

## 6 Financiamento

A coleta de dados para o presente estudo foi financiada por meio do Executive Dean Fund da University of Essex, Reino Unido, e faz parte do projeto "The Effect of ChatGPT on Student Writing in Multiple Languages: A Systemic Functional Linguistics Analysis".

## Referências

ANDERSON, Nash; BELAVY, Daniel L.; PERLE, Stephen M.; HENDRICKS, Sharief; HESPANHOL, Luiz; VERHAGEN, Evert; MEMON, Aamir R. AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport & Exercise Medicine*, v. 9, n. 1, e001568, fev. 2023. ISSN 2055-7647. DOI: 10.1136/bmjsem-2023-001568. Disponível em: https://bmjopensem.bmj.com/content/9/1/e001568. Acesso em: 21 nov. 2023.

CLAVEL-ARROITIA, Begônia; PENNOCK-SPECK, Barry. Analysing lexical density, diversity, and sophistication in written and spoken telecollaborative exchanges. *Computer Assisted Language Learning Electronic Journal (CALL-EJ)*, v. 22, n. 3, p. 230–250, 2021. Disponível em: http://callej.org/journal/22-3/Clavel-Speck2021.pdf. Acesso em: 23 jun. 2023.

COLOMBI, Maria Cecilia. Academic language development in Latino student's writing. *In:* SCHLEPPEGRELL, Mary J.; COLOMBI, Maria Cecilia (ed.). *Developing advanced literacy in first and second languages*. Mahwah: Lawrence Erlbaum Associates, 2000. p. 67–86.

DALE, Robert. GPT-3: What's it good for? *Natural Language Engineering*, v. 27, n. 1, p. 113–118, jan. 2021. ISSN 1351-3249, 1469-8110. DOI: 10.1017/S1351324920000601. Disponível em: https://www.cambridge.org/core/product/identifier/S1351324920000601/type/journal_article. Acesso em: 21 nov. 2023.

DEHOUCHE, N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, v. 21, p. 17–23, mar. 2021. ISSN 1863-5415, 1611-8014. DOI: 10.3354/esep00195. Disponível em: https://www.int-res.com/abstracts/esep/v21/p17-23/. Acesso em: 21 nov. 2023.

DÖRNYEI, Zoltán. *Research methods in Applied Linguistics*. New York: Oxford University Press, 2007.

FRÖHLING, Leon; ZUBIAGA, Arkaitz. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, v. 7, e443, abr. 2021. ISSN 2376-5992. DOI: 10.7717/peerj-cs.443. Disponível em: https://peerj.com/articles/cs-443. Acesso em: 21 nov. 2023.

GEHRMANN, Sebastian; STROBELT, Hendrik; RUSH, Alexander M. GLTR: Statistical Detection and Visualization of Generated Text, 2019. DOI: 10.48550/ARXIV.1906.04043. Disponível em: https://arxiv.org/abs/1906.04043. Acesso em: 21 nov. 2023.

GIL, Antônio Carlos. *Como elaborar projetos de pesquisa*. 4. ed. São Paulo: Atlas, 2002.

GONZÁLEZ FERNÁNDEZ, Adela. Big data y corpus lingüísticos para el estudio de la densidad léxica. *Skopos 9, 107-122 (2018)*, 2018. ISSN 2255-3703. Disponível em: http://helvia.uco.es/xmlui/handle/10396/19125. Acesso em: 21 nov. 2023.

GREGORI-SIGNES, Carmen; CLAVEL-ARROITIA, Begoña. Analysing Lexical Density and Lexical Diversity in University Students' Written Discourse. *Procedia - Social and Behavioral Sciences*, v. 198, p. 546–556, jul. 2015. ISSN 18770428. DOI: 10.1016/j.sbspro.2015.07.477. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S187704281504478X. Acesso em: 21 nov. 2023.

HALLIDAY, Michael Alexander Kirkwood. *Spoken and written language*. Geelong: Deakin University Press, 1985. (Language education).

HALLIDAY, Michael Alexander Kirkwood. Spoken and written modes of meaning. *In:* HOROWITZ, Rosalind; SAMUELS, S. Jay (ed.). *Comprehending oral and written language*. Orlando: Academic Press, 1987. p. 55–82.

HALLIDAY, Michael Alexander Kirkwood. Part A. *In:* HALLIDAY, Michael Alexander Kirkwood; HASAN, Ruqaiya (ed.). *Language, context and text*. 2. ed. Oxford: Oxford University Press, 1989. p. 3–49.

HALLIDAY, Michael Alexander Kirkwood. Some Grammatical Problems in Scientific English. *In:* HALLIDAY, Michael Alexander Kirkwood; MARTIN, Jim Robert (ed.). *Writing science: Literacy and discursive power*. London, New York: Routledge, 1993. p. 76–94.

HALLIDAY, Michael Alexander Kirkwood. The spoken language corpus: A foundation for grammatical theory. *In:* WEBSTER, Jonathan J. (ed.). *Computational and quantitative studies*. London; New York: Continuum, 2005. p. 157–190.

HALLIDAY, Michael Alexander Kirkwood; MATTHIESSEN, Christian Mathias Ingemar Martin. *An Introduction to Functional Grammar*. 4. ed. London: Edward Arnold, 2014.

JOHANSSON, Victoria. Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, v. 53, p. 61–79, 2008.

KASNECI, Enkelejda; SESSLER, Kathrin; KÜCHEMANN, Stefan; BANNERT, Maria; DEMENTIEVA, Daryna; FISCHER, Frank; GASSER, Urs; GROH, Georg; GÜNNEMANN, Stephan; HÜLLERMEIER, Eyke; KRUSCHE, Stephan; KUTYNIOK, Gitta; MICHAELI, Tilman; NERDEL, Claudia; PFEFFER, Jürgen; POQUET, Oleksandra; SAILER, Michael; SCHMIDT, Albrecht; SEIDEL, Tina; STADLER, Matthias; WELLER, Jochen; KUHN, Jochen; KASNECI, Gjergji. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, v. 103, p. 102274, abr. 2023. ISSN 10416080. DOI: 10.1016/j.lindif.2023.102274. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1041608023000195. Acesso em: 21 nov. 2023.

KEMBAREN, Farida Repelita; ASWANI, Ade Novira. Exploring Lexical Density in the New York Times. *ELLITE: Journal of English Language, Literature, and Teaching*, v. 7, n. 2, p. 109–119, nov. 2022. ISSN 25280066, 25274120. DOI: 10.32528/ellite.v7i2.8795. Disponível em: http://jurnal.unmuhjember.ac.id/index.php/ELLITE/article/view/8795. Acesso em: 21 nov. 2023.

KING, Michael R.; CHATGPT. A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. *Cellular and Molecular Bioengineering*, v. 16, n. 1, p. 1–2, fev. 2023. ISSN 1865-5025, 1865-5033. DOI: 10.1007/s12195-022-00754-8. Disponível em: https://link.springer.com/10.1007/s12195-022-00754-8. Acesso em: 21 nov. 2023.

KONDAL, Bonala. Effects of lexical density and lexical variety in language performance and proficiency. *International Journal of IT, Engineering and Applied Sciences Research (IJIEASR)*, v. 4, n. 10, p. 25–29, 2015.

KUMAR, Arun. Analysis of ChatGPT Tool to Assess the Potential of its Utility for Academic Writing in Biomedical Domain. *Biology, Engineering, Medicine and Science Reports*, v. 9, n. 1, p. 24–30, jan. 2023. ISSN 24546895. DOI: 10.5530/bems.9.1.5. Disponível em: https://www.bemsreports.org/index.php/bems/article/view/132. Acesso em: 21 nov. 2023.

LANCASTER, Thomas. Artificial intelligence, text generation tools and ChatGPT – does digital watermarking offer a solution? *International Journal for Educational Integrity*, v. 19, n. 1, p. 10, jun. 2023. ISSN 1833-2595. DOI: 10.1007/s40979-023-00131-6. Disponível em: https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00131-6. Acesso em: 21 nov. 2023.

MARTINS, Mário. Densidade lexical na escrita de textos escolares. *Signum: Estudos da Linguagem*, v. 20, n. 1, p. 218, maio 2017. ISSN 2237-4876. DOI: 10.5433/2237-4876.2017v20n1p218. Disponível em: http://www.uel.br/revistas/uel/index.php/signum/article/view/25225. Acesso em: 21 nov. 2023.

MITROVIĆ, Sandra; ANDREOLETTI, Davide; AYOUB, Omran. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. *arXiv:2301.13852v1 [cs.CL]*, 2023. DOI: 10.48550/ARXIV.2301.13852. Disponível em: https://arxiv.org/abs/2301.13852. Acesso em: 21 nov. 2023.

MOOHEBAT, Mohammadreza; RAJ, Ram Gopal; KAREEM, Sameem Binti Abdul; THORLEUCHTER, Dirk. Identifying ISI-Indexed articles by their lexical usage: A text analysis approach. *Journal of the Association for Information Science and Technology*, v. 66, n. 3, p. 501–511, mar. 2015. ISSN 2330-1635, 2330-1643. DOI: 10.1002/asi.23194. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23194. Acesso em: 21 nov. 2023.

NALESSO, Giulia. El desarrollo de la competencia léxica de estudiantes italianos universitarios de ELE. *Orillas: rivista d'ispanistica*, n. 7, p. 381–394, 2018. ISSN 2280-4390. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=7819128. Acesso em: 21 nov. 2023.

NASSERI, Maryam; THOMPSON, Paul. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, v. 47, p. 100511, jan. 2021. ISSN 10752935. DOI: 10.1016/j.asw.2020.100511. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1075293520300726. Acesso em: 21 nov. 2023.

NATION, I.S. Paul. *Learning vocabulary in another language*. 2. ed. [*S. l.*]: Cambridge University Press, 2013.

PERKINS, Mike. Academic integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, v. 20, n. 2, fev. 2023. ISSN 14499789, 14499789. DOI: 10.53761/1.20.02.07. Disponível em: https://ro.uow.edu.au/jutlp/vol20/iss2/07/. Acesso em: 21 nov. 2023.

RAMOS, Anatália Saraiva Martins. *Inteligência Artificial Generativa baseada em grandes modelos de linguagem - ferramentas de uso na pesquisa acadêmica*. [*S. l.*], maio 2023. DOI: 10.1590/SciELOPreprints.6105. Disponível em: https://preprints.scielo.org/index.php/scielo/preprint/view/6105/version/6463. Acesso em: 21 nov. 2023.

READ, John. *Assessing vocabulary*. Cambridge: Cambridge University Press, 2010.

RIFFO, Karina Fuentes; OSUNA, Sergio Hernández; LAGOS, Pedro Salcedo. Descripción de la diversidad y densidad léxicas en noticias escritas por estudiantes de periodismo. *Revista Brasileira de Linguística Aplicada*, v. 19, n. 3, p. 499–528, set. 2019. ISSN 1984-6398. DOI: 10.1590/1984-6398201914113. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1984-63982019000300499&tlng=es. Acesso em: 21 nov. 2023.

ROSPIGLIOSI, Pericles 'Asher'. Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT? *Interactive Learning Environments*, v. 31, n. 1, p. 1–3, jan. 2023. ISSN 1049-4820, 1744-5191. DOI: 10.1080/10494820.2023.2180191. Disponível em: https://www.tandfonline.com/doi/full/10.1080/10494820.2023.2180191. Acesso em: 21 nov. 2023.

SCHNUR, Erin; RUBIO, Fernando. Lexical complexity, writing proficiency and task effects in Spanish Dual Language Immersion. *Language Learning & Technology*, v. 25, n. 1, p. 53–72, fev. 2021. ISSN 1094-3501. Disponível em: http://hdl.handle.net/10125/73425. Acesso em: 21 nov. 2023.

URE, Jean. Lexical density and register differentiation. *In:* PERREN, George Ernest; TRIM, John Leslie Melville (ed.). *Applications of Linguistics: selected papers of the Second International Congress of Applied Linguistics*. Cambridge: Cambridge University Press, 1971. p. 443–452.

URE, Jean; ELLIS, Jeffrey. Register in descriptive linguistics and linguistic sociology. *In:* URIBE-VILLEGAS, Oscar (ed.). *Issues in Sociolinguistics*. The Hague: Mouton, 1977. p. 197–243.

### Contribuições dos autores

**Antonio Marcio da Silva**: Conceituação, Análise formal, Aquisição de financiamento, Metodologia, Recursos, Escrita – rascunho original, Escrita – revisão e edição; **Lucia Rottava**: Conceituação, Análise formal, Metodologia, Recursos, Escrita – rascunho original, Escrita – revisão e edição.

## A Anexo

Instruções das atividades realizadas pelos participantes

| **Write a text in Portuguese for the beginner level A1 following the instructions given below:** | **Write a text in Portuguese for the elementary level A2 following the instructions given below:** |
|---|---|
| Tema: **Amizade e fazer amigos**<br><br>Escreva um texto de 80 a 90 palavras sobre como fazer amigos. Fale sobre:<br>• Como conhecer pessoas novas.<br>• Como ser um bom amigo<br>• O que evitar em uma amizade<br>• Como fez um amigo e o que fez para mantê-lo próximo. | Tema: **Estresse**<br><br>Escreva um texto de 120 a 150 palavras sobre o estresse. Fale sobre os seguintes pontos:<br>• O que é estresse.<br>• As causas do estresse.<br>• Como lidar com o estresse.<br>• A importância de reduzir o estresse. |

| **Write a text in Portuguese for the intermediate level B1 following the instructions given below:** | **Write a text in Portuguese for the upper-intermediate level B2 following the instructions given below:** |
|---|---|
| Tema: **A educação no Brasil**<br><br>Escreva um texto de 180 a 200 palavras sobre a educação no Brasil. Inclua os seguintes pontos nas informações:<br>• Sistema educacional no Brasil.<br>• Desafios da educação no Brasil.<br>• Programas governamentais para melhorar a educação no país.<br>• O papel do professor. | Tema: **O impacto das novas tecnologias para a sociedade humana**<br><br>As novas tecnologias estão transformando rapidamente a nossa forma de viver, trabalhar e interagir com o mundo. Essas mudanças são tão profundas que podem afetar significativamente as relações sociais, a economia e a cultura. Nesse sentido, escreva um texto argumentativo que discuta o impacto das novas tecnologias para a sociedade humana, considerando os seguintes pontos:<br>• Oportunidades e desafios das novas tecnologias.<br>• A relação entre as pessoas e as máquinas.<br>• As mudanças na forma como produzimos e consumimos bens e serviços.<br>• Educação e cultura na preparação das pessoas para lidar com as novas tecnologias.<br>Ao escrever o seu texto, lembre-se de apresentar argumentos claros e coerentes, fundamentados em exemplos e evidências. O seu texto deve ter entre 200 e 250 palavras. |

| **Write a text in Portuguese for the advanced level C1 following the instructions given below:** | **Write a text in Portuguese for the proficiency level C2 following the instructions given below:** |
|---|---|
| Tema: **A arte de viver bem**<br><br>A busca pela felicidade e pelo bem-estar é uma constante na vida humana. Muitos procuram o segredo para viver bem, mas será que esse segredo existe? Nessa crônica, reflita sobre a arte de viver bem, considerando os seguintes pontos:<br>• O valor das pequenas coisas.<br>• A importância da gratidão.<br>• O equilíbrio entre trabalho e lazer.<br>• A importância das relações humanas.<br>Ao escrever a sua crônica, use sua experiência pessoal e observações do mundo ao seu redor para construir um texto pessoal e intimista. Lembre-se de usar recursos literários, metáforas e analogias, para tornar o seu texto mais expressivo e envolvente. O seu texto deve ter entre 250 e 300 palavras. | Tema: **Pontes culturais**<br><br>A interculturalidade é um tema cada vez mais presente no mundo globalizado em que vivemos. A diversidade cultural é uma riqueza que deve ser celebrada e valorizada, mas, muitas vezes, a falta de entendimento e de comunicação entre culturas diferentes pode gerar conflitos e incompreensões. Nessa atividade de escrita criativa, escreva um texto explorando o tema da interculturalidade, considerando os seguintes pontos:<br>• A beleza da diversidade.<br>• As barreiras da comunicação.<br>• A importância do intercâmbio cultural.<br>• A construção de pontes culturais.<br>Ao escrever o seu texto, use a sua criatividade e imaginação para explorar diferentes aspectos da interculturalidade. Utilize recursos literários, metáforas e analogias para tornar o seu texto mais expressivo e envolvente. O seu texto deve ter entre 300 e 350 palavras. |