UNIVERSIDADE FEDERAL DE ALAGOAS
CENTRO DE EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO - PPGE

ELIAN DA SILVA SANTOS LEITE

AVALIAÇÃO DA RIQUEZA LEXICAL EM MANUSCRITOS ESCOLARES DE ALUNOS RECÉM-ALFABETIZADOS: UMA ANÁLISE COMPARATIVA BRASIL-PORTUGAL COM O AUXÍLIO DO *LEXICANALYTICS WEB*

MACEIÓ-AL

2024

ELIAN DA SILVA SANTOS LEITE

AVALIAÇÃO DA RIQUEZA LEXICAL EM MANUSCRITOS ESCOLARES DE ALUNOS RECÉM-ALFABETIZADOS: UMA ANÁLISE COMPARATIVA BRASIL-PORTUGAL COM O AUXÍLIO DO *LEXICANALYTICS WEB*

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Educação Brasileira da Universidade Federal de Alagoas, como requisito para defesa de doutorado em Educação.

Orientador: Prof. Dr. Eduardo Calil.

Grupo de Pesquisa: Ensino, Texto, Criação (ETC)

MACEIÓ-AL

2024

Catalogação na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

L533a Leite, Elian da Silva Santos.

Avaliação da riqueza lexical em manuscritos escolares de alunos recém-alfabetizados : uma análise comparativa Brasil-Portugal com o auxílio do *Lexicanalytics Web* / Elian da Silva Santos Leite. – 2024.

111 f. : il.

Orientador: Eduardo Calil.

Tese (doutorado em Educação) – Universidade Federal de Alagoas. Centro de Educação. Programa de Pós-Graduação em Educação. Maceió, 2024.

Bibliografia: f. 67-72.

Anexos: f. 74-111.

1. Crianças - Escrita. 2. Conhecimento lexical. 3. Riqueza lexical. 4. Densidade lexical. 5. Diversidade Lexical. 6. *Lexicanalytics Web* (*Software*). I. Título.

CDU: 37.014.22:004.4’412

Universidade Federal de Alagoas
Centro de Educação
Programa de Pós-Graduação em Educação

AVALIAÇÃO DA RIQUEZA LEXICAL EM MANUSCRITOS ESCOLARES DE ALUNOS RECÉM-ALFABETIZADOS: UMA ANÁLISE COMPARATIVA BRASIL-PORTUGAL COM O AUXÍLIO DO LEXICANALYTICS WEB

ELIAN DA SILVA SANTOS LEITE

Tese de Doutorado submetida à banca examinadora, já referendada pelo Programa de Pós-Graduação em Educação da Universidade Federal de Alagoas e aprovada em 10 de abril de 2024.

Banca Examinadora:

gov.br Documento assinado digitalmente
EDUARDO CALIL DE OLIVEIRA
Data: 10/04/2024 13:55:29-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Eduardo Calil de Oliveira, Universidade Federal de Alagoas
Orientador

gov.br Documento assinado digitalmente
BRUNO ALMEIDA PIMENTEL
Data: 20/04/2024 16:14:28-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Bruno Almeida Pimentel, Universidade Federal de Alagoas
Avaliador Interno

ICP Brasil Documento assinado digitalmente
LEONARDO BRANDAO MARQUES
Data: 23/04/2024 09:54:42-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Leonardo Brandão Marques, Universidade Federal de Alagoas
Avaliador Interno

gov.br Documento assinado digitalmente
DEBORA AMORIM GOMES DA COSTA MACIEL
Data: 10/04/2024 14:20:06-0300
Verifique em https://validar.iti.gov.br

_____
Profa. Dra. Débora Amorim Gomes da Costa-Maciel, Universidade de Pernambuco
Avaliadora Externa À Instituição

gov.br Documento assinado digitalmente
PEDRO DE LEMOS MENEZES
Data: 24/04/2024 10:30:52-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Pedro De Lemos Menezes, Universidade Estadual de Ciências da Saúde de Alagoas
Avaliador Externo À Instituição

A Deus, aos meus pais, ao meu esposo Glauber, ao Qzyrk e Kiki, todo o meu amor e minha gratidão.

AGRADECIMENTOS

Agradeço a Deus por sua infinita bondade e por me fortalecer continuamente, permitindo-me seguir na realização dos meus sonhos. Sem a sua presença em minha vida, muitos dos desafios que enfrentei ao longo desta jornada teriam sido insuportáveis. A fé em seu amor me deu a coragem necessária para perseverar, mesmo nos momentos mais difíceis. Agradeço imensamente aos meus pais, Manoel e Maria, e ao meu irmão Eliel, por sonharem comigo através da educação. Lembro-me das longas horas de caminhada com minha mãe até a escola, das inúmeras estratégias para desviar dos obstáculos. Sempre do meu lado, meus pais me apoiaram em todas as fases. Que privilégio os ter comigo na concretização deste grande sonho.

Ao longo deste percurso acadêmico, encontrei pessoas que mudaram minha vida e contribuíram para a construção de novas perspectivas para enxergar o mundo e a vida. Uma dessas pessoas foi meu esposo Glauber. Compartilhamos as experiências e as angústias da graduação, mestrado e doutorado, sempre nos fortalecendo com a escuta, paciência, compreensão e amor. Ao meu esposo, meu amor e gratidão.

Tive a grande honra de ter como orientador o admirável pesquisador e professor Dr. Eduardo Calil, a quem eu agradeço imensamente por ter me conduzido nesse percurso e pela grande contribuição para a minha formação como pedagoga e pesquisadora. Suas aulas e ensinamentos ficarão guardados com muito afeto em minha memória. E por onde eu for, sempre terei o orgulho dizer: eu tive como orientador o professor Dr. Eduardo Calil.

Agradeço à professora Débora Amorim Gomes e aos professores Bruno Almeida Pimentel, Leonardo Brandão Marques e Pedro De Lemos Menezes, por aceitarem fazer parte da banca examinadora. O olhar atento de cada um e as ricas contribuições ajudaram significativamente no aprimoramento deste trabalho.

Aos meus amigos de longa data, Carlos e Lys, que me acompanharam desde o Ensino Médio, e aos amigos que ganhei na pós-graduação, Salezia, Vanessa, Petrus e Viviane, meus profundos agradecimentos. A amizade e o apoio de vocês fizeram a diferença em cada passo deste percurso.

Agradeço à Universidade Federal de Alagoas (UFAL) pelo acolhimento e assistência, e por se tornar minha segunda casa.

Agradeço aos coordenadores, professores e técnicos do Programa de Pós-Graduação em Educação (PPGE), especialmente ao Lucas Melo, pelo trabalho comprometido e apoio indispensável ao longo desses anos.

Agradeço ao grupo de pesquisa do Laboratório do Manuscrito Escolar por cada encontro e pelas discussões enriquecedoras que contribuíram significativamente para o avanço desta pesquisa.

RESUMO

A avaliação da riqueza lexical pode revelar informações valiosas sobre a proficiência dos estudantes em sua língua materna. No entanto, essa avaliação pode ser desafiadora, principalmente em um contexto de múltiplos textos, como em sala de aula, sendo necessário o auxílio de recursos computacionais para facilitar nesse processo. Diante disso, este trabalho tem como objetivo avaliar a riqueza lexical, a partir da densidade e diversidade lexical, em manuscritos escolares produzidos por alunos brasileiros e portugueses recém-alfabetizados, utilizando uma metodologia de escrita colaborativa. O *corpus* analisado é composto por um conjunto de 186 narrativas ficcionais produzidas por díades de alunos brasileiros e portugueses recém-alfabetizados, com idade entre seis e oito anos. Os alunos brasileiros cursavam o primeiro e o segundo ano do Ensino Fundamental em uma escola localizada na cidade de Maceió-Alagoas, enquanto os alunos portugueses cursavam o segundo ano do Ensino Básico em duas escolas públicas de Aveiro, sendo uma escola localizada na zona urbana e a outra na zona rural. Para a mensuração da riqueza lexical foram considerados como parâmetros o total de palavras escritas, a densidade e a diversidade lexical, os quais foram extraídos com o auxílio do software *Lexicanalytics Web*. Caracterizada como uma pesquisa exploratória, foi adotada uma abordagem qualiquantitativa para a análise dos dados. Na análise dos resultados, constatou-se que as díades do primeiro ano escreveram seus textos com uma média de 107 palavras, diversidade lexical de 72,06%, densidade lexical de 56,14% e uma riqueza lexical de 91,53. Nos textos das díades do segundo ano, por outro lado, a média de palavras escritas foi menor, aproximadamente 86 palavras, com uma diversidade lexical de 71,12% e uma densidade lexical de 56,97%, alcançando uma riqueza lexical de 91,23. Quanto aos resultados das díades portuguesas que cursavam o mesmo ano escolar, mas em escolas diferentes, os textos foram produzidos com uma média de 113 palavras, com densidade média de 57,82%, diversidade de 71,59% e riqueza de 93,15. Na comparação global entre as díades, observou-se que, embora os textos das díades portuguesas tenham apresentado uma riqueza lexical ligeiramente superior aos textos das díades brasileiras, o teste de hipótese não constatou diferença estatisticamente significativa entre os dois países. Logo, mesmo considerando que os textos foram produzidos por alunos de escolas e nacionalidades diferentes, a riqueza lexical obtida foi semelhante, indicando um mesmo domínio lexical em seus textos. Por fim, esperamos que os resultados obtidos neste trabalho contribuam para a produção de conhecimento acerca da riqueza lexical em textos de língua portuguesa, revelando aspectos importantes da escrita de alunos brasileiros e portugueses recentemente alfabetizados. Além disso, esperamos abrir novos caminhos para a avaliação e o acompanhamento do conhecimento lexical dos alunos, destacando a relevância de ferramentas como o *Lexicanalytics Web* nesses processos.

Palavras-chave: Escrita Infantil. Conhecimento Lexical. Riqueza Lexical. Densidade Lexical. Diversidade Lexical. *Lexicanalytics Web*.

ABSTRACT

The assessment of lexical richness can reveal valuable information about students' proficiency in their native language. However, this evaluation can be challenging, especially in a context of multiple texts, such as in a classroom, requiring the aid of computational resources to facilitate this process. Therefore, this work aims to assess lexical richness, based on lexical density and diversity, in school manuscripts produced by newly literate Brazilian and Portuguese students, using a collaborative writing methodology. The analyzed corpus consists of a set of 186 fictional narratives produced by dyads of newly literate Brazilian and Portuguese students, aged between six and eight years. The Brazilian students attended the first and second years of Elementary School in a school located in the city of Maceió-Alagoas, while the Portuguese students attended the second year of Basic Education in two public schools in Aveiro, one located in the urban area and the other in the rural area. For the measurement of lexical richness, the total number of words written, lexical density, and lexical diversity were considered as parameters, which were extracted with the aid of the Lexicanalytics Web software. Characterized as exploratory research, a qualitative and quantitative approach was adopted for data analysis. In the analysis of the results, it was found that the dyads of the first year wrote their texts with an average of 107 words, lexical diversity of 72.06%, lexical density of 56.14%, and a lexical richness of 91.53. In the texts of the dyads of the second year, on the other hand, the average number of words written was lower, approximately 86 words, with a lexical diversity of 71.12% and a lexical density of 56.97%, reaching a lexical richness of 91.23. As for the results of the Portuguese dyads who attended the same school year but in different schools, the texts were produced with an average of 113 words, with an average density of 57.82%, diversity of 71.59%, and richness of 93.15. In the overall comparison between the dyads, it was observed that, although the texts of the Portuguese dyads showed slightly higher lexical richness than the texts of the Brazilian dyads, the hypothesis test did not find a statistically significant difference between the two countries. Therefore, even considering that the texts were produced by students from different schools and nationalities, the obtained lexical richness was similar, indicating the same lexical domain in their texts. Finally, we hope that the results obtained in this work contribute to the production of knowledge about lexical richness in Portuguese language texts, revealing important aspects of the writing of newly literate Brazilian and Portuguese students. Furthermore, we hope to open new paths for the evaluation and monitoring of students' lexical knowledge, highlighting the relevance of tools such as Lexicanalytics Web in these processes.

Keywords: Children's Writing. Lexical Knowledge. Lexical Richness. Lexical Density. Lexical Diversity. Lexicanalytics Web.

SUMÁRIO

1 INTRODUÇÃO

A produção textual é uma atividade que demanda dos aprendizes um esforço contínuo para desenvolver e aprimorar os diferentes conhecimentos acerca da língua escrita, dentre eles, o conhecimento do léxico (Koch; Elias, 2010). No entanto, a avaliação e o acompanhamento do conhecimento lexical dos estudantes não é uma tarefa fácil para os professores (Isaacson, 1988), especialmente quando o objetivo é avaliar e comparar uma coleção de textos escritos por um único estudante ou por um grupo de estudantes, como em um contexto de sala de aula.

Compreendendo a importância e a complexidade da avaliação textual, o Laboratório do Manuscrito Escolar (LAME)[1], que é reconhecido por sua extensa trajetória dedicada à pesquisa sobre a escrita de alunos recém-alfabetizados de países como Brasil, França e Portugal, vem investigando temas como a rasura (Calil, 2012; Calil; Felipeto, 2014; Queiroz, 2019), o erro ortográfico (Lopes, 2005), a pontuação (Santos, 2023) e os comentários durante a produção textual (Silva, 2019). Recentemente, o LAME iniciou uma nova linha de pesquisa focada na avaliação do conhecimento lexical em textos escolares, com dois enfoques principais: (i) contribuir com o campo de estudo da escrita em língua materna, por meio da avaliação da Riqueza Lexical (RL) em textos de alunos recém-alfabetizados; (ii) desenvolver técnicas e recursos práticos para auxiliar os professores na avaliação da riqueza lexical da escrita dos alunos, utilizando dois indicadores principais: densidade e diversidade lexical.

Esse novo campo de pesquisa resultou na condução de novos estudos, incluindo os trabalhos de Santos *et al*. (2017; 2022), os quais se concentraram na criação de uma ferramenta de análise lexical denominada *Lexicanalytics Web*. Essa ferramenta foi desenvolvida para facilitar a avaliação da riqueza lexical e, posteriormente, foi empregada em um estudo para analisar a densidade e a diversidade lexical em 30 textos produzidos por cinco duplas de alunos brasileiros recém-alfabetizados, matriculados no 2º ano do Ensino Fundamental (Santos *et al*., 2018).

Visando ampliar nossas pesquisas acerca dessa temática e investigar a riqueza lexical em textos de alunos de diferentes nacionalidades, foi desenvolvida a presente pesquisa de doutorado, cujo objetivo geral é *avaliar a Riqueza Lexical em manuscritos escolares produzidos*

[1] O Laboratório do Manuscrito Escolar (LAME), fundado em 2010 e instalado no Centro de Pesquisa em Educação e Linguagem (CEPEAL), da universidade Federal de Alagoas (UFAL), reúne pesquisadores de diferentes partes do Brasil. Em seu escopo de pesquisa propõe-se: a) Investigar processos de escritura com foco sobre as relações entre sujeito, língua e sentido, delineando procedimentos metodológicos que permeiem a subjetividade e permitam a detecção da criação; b) Elaborar e analisar materiais didáticos com propostas de produção de textos que favoreçam a efetivação destes processos; c) desenvolvimento de técnicas e recursos para auxiliar os professores no processo de avaliação da riqueza lexical em produções textuais.

por díades de alunos brasileiros e portugueses recém-alfabetizados, utilizando a ferramenta Lexicanalytics Web.

Buscamos ao longo deste estudo responder à seguinte questão principal: Há diferença significativa da riqueza lexical nos manuscritos produzidos por díades de alunos brasileiros e portugueses recém-alfabetizados? Considerando que os textos foram elaborados por duplas de alunos com idades e níveis escolares próximos, utilizando a mesma metodologia de escrita e submetidos a mesma ferramenta para medição da riqueza lexical, nossa hipótese é que não haverá diferença significativa entre seus textos.

Além dessa questão principal, buscamos responder as seguintes perguntas adicionais: i) Há diferença significativa da riqueza lexical nas produções textuais de díades brasileiras matriculadas na mesma escola, mas cursando anos escolares diferentes (1º e 2º ano)? ii) Há diferença significativa da riqueza lexical nas produções textuais das díades portuguesas cursando o mesmo ano escolar (2º ano), porém matriculadas em escolas diferentes?

Com o propósito de alcançar o objetivo geral delineado e responder adequadamente às questões de pesquisa propostas, estabelecemos os seguintes objetivos específicos:

i) Estabelecer a base conceitual em torno da riqueza lexical, assim como os aspectos históricos e metodológicos que permeiam sua definição e mensuração;
ii) Apresentar a definição de densidade e diversidade lexical, bem como as técnicas utilizadas para sua mensuração;
iii) Discutir sobre as diferentes aplicações da densidade e diversidade lexical na avaliação de textos escolares;
iv) Apresentar as ferramentas disponíveis para extração da densidade, diversidade lexical e riqueza lexical;
v) Detalhar o processo de desenvolvimento da ferramenta *Lexicanalytics Web*;
vi) Descrever os resultados da medição da densidade lexical, diversidade lexical e riqueza lexical nos textos das díades brasileiras e portuguesas.

Com base nesses objetivos, a tese está estruturada em cinco seções, incluída esta introdução, que compreende a primeira seção. A segunda seção apresenta a base conceitual que fundamenta o conhecimento lexical, com ênfase na riqueza lexical que se caracteriza como o foco de análise deste estudo. Nela, são abordadas as bibliografias que exploram os aspectos históricos e metodológicos concernentes à definição e mensuração da riqueza lexical, além dos seus indicadores denominados densidade e diversidade lexical. A terceira seção aborda como esses indicadores vêm sendo aplicados no contexto educacional e traz uma apresentação das

ferramentas disponíveis para auxiliar os professores e pesquisadores no seu processo de extração.

A quarta seção detalha as duas etapas que compõem a metodologia adotada. Na primeira etapa, são enfatizadas as fases de desenvolvimento da ferramenta *Lexicanalytics Web*, que incluem definições de requisitos, arquitetura e validação. Na segunda etapa, o foco está no processo de composição do *corpus* usado para a demonstração da avaliação da riqueza lexical, onde são explicitados os sujeitos da pesquisa, os procedimentos adotados na coleta de dados, bem como as categorias de análise utilizadas.

Por fim, na quinta seção, são apresentados os resultados alcançados com base nos objetivos e hipóteses previamente estabelecidos, juntamente com a discussão dos principais achados e uma análise de como eles podem contribuir para o avanço do conhecimento no campo de estudo em questão.

2 FUNDAMENTACÃO TEÓRICA

Esta seção apresenta a base conceitual acerca da Riqueza Lexical, abordando os aspectos históricos e metodológicos que permeiam sua definição e mensuração em produções textuais escritas.

2.1 Conhecimento Lexical

O conhecimento lexical é amplamente reconhecido como um componente fundamental para a proficiência linguística dos estudantes (Nation, 1990) e, consequentemente, a base para suas produções escritas, uma vez que ele possui relação direta com o vocabulário (Koch; Elias, 2010). Dada sua relevância, diferentes estudos vêm sendo desenvolvidos para explorar sua natureza e formas avaliação.

Henriksen (1999) destaca que o conhecimento lexical está associado à habilidade de reconhecer uma palavra e associá-la ao seu significado em contextos receptivos, como durante uma leitura, e à capacidade de representar graficamente ou foneticamente uma palavra, vinculando-a ao seu significado apropriado em contextos de produção textual. Com base nessa abordagem, Henriksen sugere que compreender uma palavra vai além do seu reconhecimento, isto é, requer a habilidade de contextualizá-la adequadamente em situações reais de comunicação.

Corroborando com essa perspectiva, Nagy e Scott (2000) afirmam que a compreensão de uma palavra é intrinsecamente multidimensional, envolvendo diversas facetas de conhecimento qualitativamente distintas. Sob essa ótica, enfatizam que sua avaliação deve transcender a mera contagem de palavras, exigindo uma abordagem mais holística que considere as diferentes dimensões do conhecimento lexical dos aprendizes.

O desenvolvimento do conhecimento lexical transita por diferentes etapas (Anderson, 2009), delineadas da seguinte maneira: i) etapa cognitiva, na qual os aprendizes conscientemente integram novas palavras ao seu vocabulário. Essa fase é marcada pelo esforço para entender e memorizar novas palavras; ii) etapa associativa, na qual os aprendizes começam a assimilar as regras de uso das palavras em diversos contextos. Nessa fase, eles estabelecem conexões entre as palavras e suas aplicações, o que é essencial para uma compreensão mais profunda e um uso proficiente do vocabulário; iii) etapa de autonomia, que é caracterizada pela automatização dessas operações. Ou seja, os aprendizes são capazes de utilizar seu vocabulário de forma proficiente de acordo com suas necessidades comunicativas.

Logo, a progressão de etapas no aprendizado lexical não apenas ressalta a complexidade inerente ao processo de aquisição do vocabulário, mas também sublinha a importância de práticas de ensino que sejam adaptadas para apoiar os alunos em cada uma delas.

Considerando essa perspectiva, podemos compreender que diferentes aspectos cognitivos e linguísticos são acionados durante a recepção e produção de uma palavra. Isso nos leva a considerar a importância, no contexto de ensino-aprendizagem, de uma avaliação detalhada acerca das escolhas lexicais dos alunos em suas produções textuais, pois como salienta Leffa (2000):

> Conhecer uma palavra não é apenas estabelecer a conexão rígida entre forma e conteúdo, como se fosse dois monólitos que se encaixam um ao outro, impossíveis de serem analisados. Conhecer uma palavra é despi-la de sua embalagem, descobrir as partes que a compõem e ver como cada uma dessas partes tem repercussões lá fora, com elementos internos de outras palavras (Leffa, 2000, p. 40).

Frente à complexidade da aquisição do conhecimento lexical, o diagnóstico e o acompanhamento pedagógico se tornam fundamentais não apenas para avaliar a progressão dos estudantes, mas também para diagnosticar possíveis dificuldades em relação à aprendizagem de sua língua materna. Adicionalmente, esse acompanhamento contribuirá para o direcionamento do trabalho docente, especialmente na tomada de decisão das metodologias de ensino a serem adotadas em sala de aula.

No entanto, esse acompanhamento e diagnóstico traz consigo um desafio: como avaliar o conhecimento lexical? Na busca de uma resolução para essa problemática, estudiosos das diferentes áreas do conhecimento (Linguística Aplicada, Psicolinguística e Computação) vêm propondo alternativas metodológicas para auxiliar na extração e avaliação do conhecimento lexical dos estudantes a partir de suas produções textuais. Uma dessas alternativas é o índice denominado de Riqueza Lexical (RL), que mede o conhecimento lexical através de indicadores linguísticos como densidade e diversidade lexical (Read, 2000).

2.2 Riqueza Lexical (RL): conceitos e métricas

A riqueza lexical[2] é um índice que contribui para uma maior compreensão do conhecimento lexical dos aprendizes (Skehan, 2009). Existem várias técnicas para sua mensuração, apoiadas em diferentes fundamentos teóricos. Uma dessas abordagens sugere que a RL está diretamente relacionada à extensão do texto, indicando que textos mais longos tendem

[2] Este conceito é também conhecido como Complexidade Lexical. No entanto, neste trabalho, em consonância com Read (2000), adotamos o termo Riqueza Lexical.

a demonstrar maior riqueza ou complexidade lexical (Menard, 1983). Há abordagens que consideram a RL como sinônimo da diversidade lexical, podendo ser mensurada pela quantidade de palavras ortograficamente diferentes do texto (Arnaud, 1984).

Nesse panorama, a contribuição de Read (2000) se sobressai com uma proposta ampla para a definição da RL, proporcionando uma base mais sólida para sua compreensão e avaliação. De acordo com sua definição, a RL é considerada como uma característica multidimensional da escrita, composta por quatro indicadores linguísticos já estabelecidos na literatura. São eles: densidade lexical, diversidade lexical, número de erros e sofisticação lexical.

A Densidade Lexical (DeL) se refere à proporção de palavras com valores lexicais do texto, os chamados itens lexicais (Ure, 1971; Laufer; Nation, 1994). A Diversidade Lexical (DiL) corresponde à variedade do vocabulário de um texto (Mccarthy; Jarvis, 2007). Enquanto a Sofisticação Lexical (SL), por sua vez, corresponde à proporção de palavras avançadas (Laufer; Nation, 1994). E, por fim, o Número de Erros (NE), que abrange os problemas ortográficos e de morfologia derivacional (Read, 2000).

Embora esses quatro indicadores sejam importantes para avaliar a riqueza lexical, é possível observar que a densidade e, principalmente, a diversidade lexical tem sido frequentemente utilizada em estudos sobre o desenvolvimento lexical. Essa prevalência pode ser atribuída à sua "[...] estabilidade enquanto indicador confiável de verificação do desenvolvimento linguístico do repertório lexical" (Martins, 2016, p. 1069) e a sua natureza quantificável, tornando esses indicadores mais apropriados para a aplicação de ferramentas computacionais em avaliação de larga escala (Johansson, 2008).

Dessa maneira, o uso da DeL e DiL pode ser observado em estudos que avaliam os estágios do desenvolvimento linguístico de crianças e adolescentes (Rodrigues, 2008), em investigações sobre a progressão escolar (Johansson, 2009; Martins, 2016), em estudos comparativos entre textos orais e escritos (Ure, 1971; Stromqvist *et al*., 2002), em trabalhos que avaliam a proficiência em segunda língua (L2) (Raimes, 1985; Uzawa; Kondal, 2015; Elgobshawi; Aldawsari, 2022). Além disso, esses indicadores podem ser usados para diferenciar os elementos da qualidade da escrita e o conhecimento do vocabulário dos estudantes (Laufer; Nation, 1994).

Portanto, reconhecendo as diversas aplicabilidades da densidade e da diversidade lexical, assim como sua consolidação na literatura, o presente estudo também adota esses indicadores para a avaliação da riqueza lexical em um *corpus* composto por produções textuais de estudantes recém-alfabetizados.

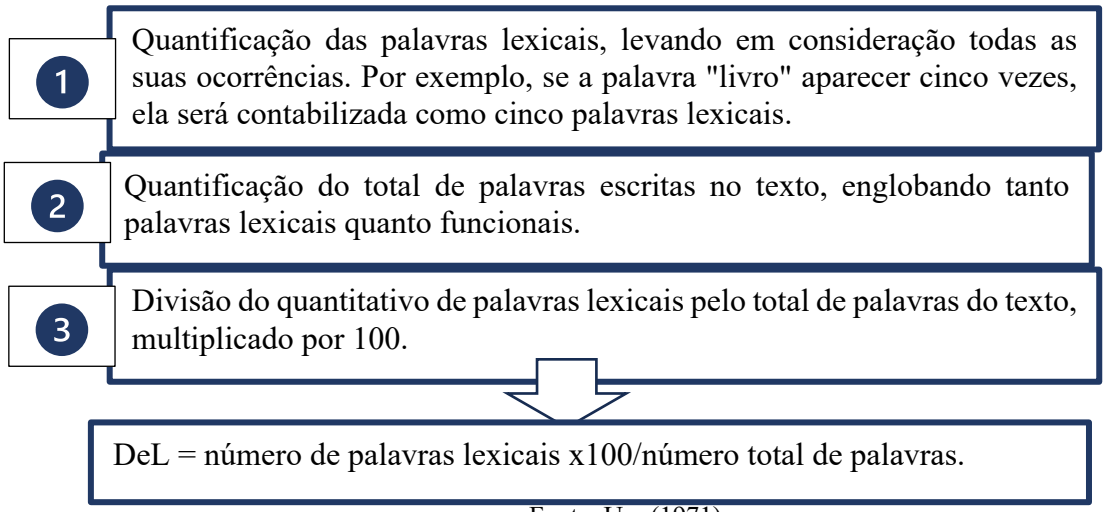
2.2.1 Densidade Lexical (DeL): natureza e mensuração

O conceito de Densidade Lexical foi proposto por Ure (1971) para descrever a proporção de palavras lexicais em relação ao total de palavras do texto. A DeL também pode ser considerada como um índice da quantidade de informações (Ishikawa, 2015). A definição da densidade lexical baseia-se na diferenciação das funções que as palavras exercem no texto, podendo assumir funções lexicais ou apenas funcionais, como destacado por Laufer e Nation (1994).

As palavras lexicais, também denominadas de palavras de conteúdo ou itens lexicais, são responsáveis por transmitir as informações do texto, elas englobam os substantivos, adjetivos, verbos e os advérbios de modo terminados em "-mente" (Ure, 1971; Halliday, 1985; 1993). Enquanto as palavras funcionais (pronomes, artigos, numerais etc.) exercem o papel de acompanhar as palavras lexicais ou conectá-las uma as outras. Sobre essa diferenciação, Johansson (2008, p. 64) ressalta que um texto com uma alta proporção de palavras lexicais contém mais informações do que um texto com uma alta proporção de palavras funcionais.

Segundo o método proposto por Ure (1971), a densidade pode ser calculada por meio de uma operação matemática, seguindo os passos abaixo:

Figura 1 – Método de Mensuração da Densidade Lexical



Fonte: Ure (1971).

O resultado dessa aplicação pode ser expresso em porcentagem, no intervalo de 0 a 100%. Quanto maior o valor resultante dessa operação, mais denso o texto será em termos de proporção de palavras lexicais e, consequentemente, de concentração de informações.

2.2.2 Diversidade Lexical (DiL): técnicas de mensuração

A diversidade lexical é definida como a variedade de palavras presente no texto (Mccarthy; Jarvis, 2007), sendo diretamente associada ao vocabulário do aprendiz. Esse indicador é frequentemente adotado como um reflexo da habilidade do indivíduo em usar uma variedade de palavras de forma proficiente e adequada, bem como um importante indicador de uma comunicação criativa (Yu, 2009) e do desenvolvimento linguístico (Rodrigues, 2009).

Para compreender com maior profundidade sobre os aspectos históricos e metodológicos da DiL, é essencial consultarmos os autores pioneiros, responsáveis por estabelecer os primeiros conceitos e técnicas para sua mensuração, sendo eles: Zipf (1935), Carroll (1938) e Johnson (1944).

Em sua obra intitulada "A Psicobiologia da Linguagem"[3], o linguista Zipf (1935) propôs a Lei de Zipf para explicar, através de um modelo estatístico, a frequência das palavras em uma amostra textual. Segundo essa lei, a frequência de qualquer palavra em uma amostra de dados textuais é inversamente proporcional à sua posição em uma tabela de frequência.[4] Ou seja, a palavra mais frequente ocorrerá duas vezes mais que a segunda palavra mais frequente, três vezes mais que a terceira palavra mais frequente e assim sucessivamente.

Corroborando com a lei de Zipf, o psicólogo e linguista Carroll (1938) também analisou a frequência das palavras em textos de língua inglesa e observou a predominância de um pequeno número de palavras altamente frequentes no texto, a exemplo de "e" e "de*"*. Nesse mesmo trabalho, Carroll (1938) introduz o conceito de diversidade lexical, definindo-a como a quantidade relativa do vocabulário encontrada no uso da linguagem produtiva de um falante ou escritor, medida pela razão entre *types* e *tokens.*

De acordo com Johnson (1944), os *tokens* se referem às palavras individuais em uma sequência de texto. Por exemplo, a frase "Palavras que transportam ideias, palavras que alimentam a imaginação" possui nove *tokens:* "Palavras", "que", "transportam", "ideias", "palavras", "que", "alimentam", "a", "imaginação". Por outro lado, os *types* representam as formas únicas e distintas de uma palavra, independentemente de quantas vezes ela apareça no texto. Usando a mesma frase como exemplo, a classificação de *types* seria: "Palavras", "que", "transportam", "ideias", "alimentam", "a", "imaginação". Logo, a frase possui sete t*ypes* e nove *tokens.*

[3] Versão original: *The Psychobiology of Language.*
[4] É uma forma de representação da frequência de cada valor distinto de uma determinada variável.

Os conhecimentos produzidos pelos estudos Zipf (1935), Carroll (1938) e Johnson (1944) desempenharam um papel fundamental no estabelecimento de conceitos que orientaram o desenvolvimento de investigações subsequentes na área, especialmente no que diz respeito às métricas de mensuração da diversidade lexical.

Quanto à mensuração da DiL, a literatura apresenta uma gama de técnicas que abrange desde métodos matemáticos mais simples, como a aplicação direta de equações baseadas em operações elementares, até as abordagens mais avançadas que empregam algoritmos e conceitos de estatística inferencial. Um aspecto importante a ser ressaltado sobre a evolução desses métodos é a integração das tecnologias computacionais para promover a avaliação da DiL em contextos de larga escala.

Traçando uma linha cronológica do desenvolvimento dessas métricas, um marco importante é a introdução da *Type-Token Ratio* (TTR) desenvolvida por Johnson (1944), considerada uma das medidas mais influentes e recorrentes nos trabalhos que avaliam a diversidade lexical. A TTR é teoricamente definida como:

> Uma medida da "flexibilidade" ou variabilidade do vocabulário projetada para indicar certos aspectos da adequação da linguagem, expressa pela proporção de diferentes palavras *(types)* para o total de palavras (*tokens*) em uma determinada amostra de língua (Johnson, 1944, p.1, tradução nossa)[5].

Em síntese, a TTR é representada matematicamente pela equação expressa a seguir, em que t equivale a *types* e n equivale a *tokens,* tendo seus resultados demonstrados no intervalo de 0 a 100%.

$$TTR = \frac{t}{n}\ x100 \qquad (1)$$

Embora a TTR seja uma métrica amplamente utilizada, sua confiabilidade tornou-se objeto de análise e questionamento em diversos estudos. Um exemplo notável é o estudo realizado por Jarvis e McCarthy (2010), que examinou o desempenho da TTR na avaliação da diversidade lexical em diferentes amostras textuais. Como resultado, eles observaram uma limitação no desempenho da TTR, principalmente quando aplicada em textos de diferentes comprimentos, revelando que essa métrica pode não ser consistentemente eficaz em todas as

[5] Versão original: *Type-Token Ratio (TTR) is a measure of vocabulary "flexibility" or variability, designed to indicate certain aspects of language adequacy. It expresses the ratio of different words (types) to total words (tokens) in a given language sample. (JOHNSON, 1944, p.1)*

situações de avaliação. De acordo com esses autores, essa limitação se deve ao seguinte fenômeno:

> À medida que um texto se torna mais longo (novas palavras são adicionadas), o número de palavras gerais no texto (ou *tokens*) aumenta. Mas, embora o aumento de *token* seja linear (uma nova palavra, um novo *token*), a taxa de aumento do número de palavras diferentes no texto (ou *types*) diminui constantemente. A desaceleração do aumento de *type* ocorre porque com cada nova instância de um token há uma diminuição correspondente na probabilidade de um novo *type*. (Jarvis; Mccarthy. 2010, p.10, tradução nossa).[6]

O distanciamento entre o número de types e tokens promove uma relação inversa entre DiL e o tamanho do texto, levando à falsa convicção de que textos longos possuem baixa DiL. Apesar dessa limitação da TTR, ainda é possível observar que alguns trabalhos a utilizam para mensurar a DiL, a exemplo Scherer e Souza (2011) e Riffo (2019).

Ao longo do tempo, foram propostas diversas alternativas para corrigir essa fragilidade. No Quadro 1, é possível conferir duas delas:

Quadro 1 - Propostas para correção da TTR

Root Type-Token Ratio (RTTR) [7]	$RTTR = \frac{t}{\sqrt{n}}$ *(2)*
Corrected Type-Token Ratio (CTTR)[8]	$CTTR = \frac{t}{\sqrt{2\,n}}$ *(3)*

No estudo de Malvern (2004), essas propostas para ajuste da TTR foram exploradas, e observou-se que os valores de *tokens* continuavam a crescer nas primeiras centenas de palavras dos textos analisados. Além disso, Malvern notou uma tendência de diminuição nos valores da diversidade lexical em textos extensos. Essa queda, embora lenta e muitas vezes quase imperceptível, provocava uma aparente consistência nos valores obtidos, levando a uma falsa percepção de precisão nessas medidas corretivas da TTR.

Novas alternativas continuaram sendo desenvolvidas, desta vez, apoiadas em modelagens matemáticas e computacionais mais avançadas. Essa evolução resultou no surgimento de técnicas cada vez mais inovadoras para superar o problema de dependência do

---

[6] Versão original: *Text becomes longer (that is, adding new words to it), the number of general words in the text (or tokens) increases. But while the token increase is linear (a new word, a new token), the rate of increase in the number of different words in the text (or types) decreases steadily. The increase in type deceleration occurs because each new instance of a token has a corresponding decrease in the likelihood of a new type.*

[7] Raiz da TTR

[8] Correção da TTR

comprimento do texto na medição da DiL. Uma delas é o coeficiente D, baseado em um modelo matemático mais sofisticado.

O coeficiente D usa uma técnica de ajuste de curva (*curve-fitting*) para o comportamento decrescente da TTR, ou seja, a tendência de diminuição de *types* à medida que o tamanho do texto aumenta (Mccarthy; Jarvis, 2007 *apud* Malvern; Richard, 1997). Esse coeficiente é determinado pela seguinte equação:

$$TTR = \frac{D}{N}\left[\left(1 + 2\frac{N}{D}\right)^{\frac{1}{2}} - 1\right] \quad (4)$$

Sendo:

TTR representa a relação entre *types* e *tokens*.

N é o número total de *tokens*.

D é o parâmetro de ajuste da TTR.

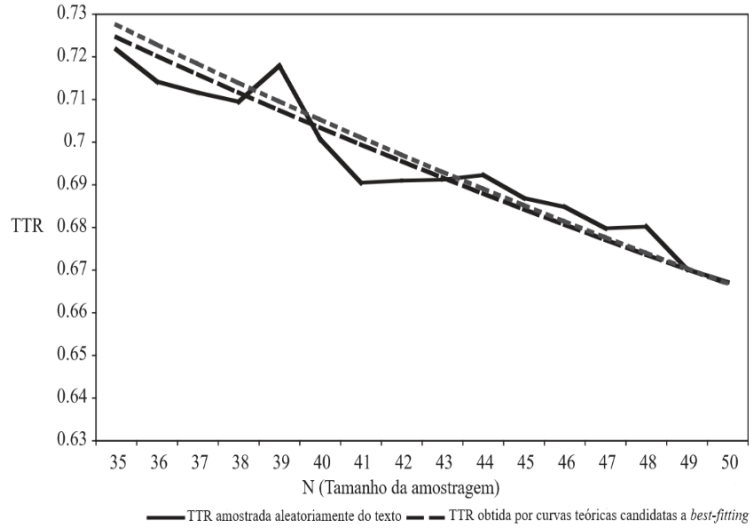
Como resultado dessa equação, é obtido o valor de D, que representa o índice de diversidade lexical. Valores altos de D refletem um alto nível de diversidade lexical de um texto, enquanto valores mais baixos indicam uma diversidade menor (Malvern *et al*., 2004). No entanto, por ser baseada em um modelo matemático mais avançado, essa técnica passou a contar com o auxílio de um algoritmo, denominado vocd, para computar o parâmetro D que aproxima a curva teórica à curva empírica. O algoritmo funciona da seguinte maneira:

> Primeiro, o vocd estima o nível de Diversidade Lexical (LD) de um texto pegando 100 amostras aleatórias de 35 tokens retirados do texto e calculando uma média TTR para essas amostras. Esse procedimento é repetido para amostras de 36 tokens, 37 tokens, e assim por diante, até amostras de 50 tokens. O programa então plota os valores médios do TTR para cada tamanho de amostra a fim de criar uma curva TTR de amostragem aleatória para o texto. (Mccarthy; Jarvis, 2007, p. 462. Tradução nossa).[9]

Após esse primeiro procedimento, com os valores médios da TTR computados para as amostras de 35 a 50 *tokens*, o algoritmo vocd aplica a equação do coeficiente D para produzir uma curva teórica que mais se aproxime da curva TTR de amostragem aleatória, isto é, para encontrar a diversidade lexical. O Gráfico 1 apresenta um exemplo visual que demonstra a representação das curvas geradas pelo processo descrito

[9] Versão original: *First, vocd estimates a text's level of LD by taking 100 random samples2 of 35 tokens drawn from the text and calculating a mean TTR for these samples. This procedure is repeated for samples of 36 tokens, 37 tokens, and so on, all the way to samples of 50 tokens. The program then plots the mean TTR values for each sample size in order to create a random-sampling TTR curve for the text.*

Gráfico 1 - Gráfico referente ao processo de ajuste de curva da vocd.



Fonte: Adaptação de Mcchathy e Jarvis (2007).

Quanto à confiabilidade dos resultados dessa técnica, os estudos de Marvern *et al*. (2004) e Jarvis (2002) confirmaram que o índice D se mostrou mais resistente ao efeito do tamanho do texto, sendo indicado como uma alternativa confiável para a mensuração da diversidade lexical em textos de diferentes comprimentos.

Em um estudo posterior, focado na análise da complexidade do algoritmo vocd(D), McCarthy e Jarvis (2007) observaram que essa técnica de mensuração da DiL se aproximava de uma distribuição hipergeométrica. Diante disso, eles descreveram e propuseram uma forma mais direta de calcular a diversidade lexical, utilizando um algoritmo menos complexo do ponto de vista computacional, denominado de índice HD-D[10].

Em síntese, é possível compreender que a mensuração da diversidade lexical vai muito além de uma simples contagem das palavras diferentes no texto. Como foi evidenciado ao longo desta seção, ela é uma tarefa complexa que demanda uma abordagem matemática analítica e refinada. Partindo da medida pioneira TTR até o desenvolvimento de métodos mais avançados como o coeficiente D e o índice HD-D, é possível observar um grande progresso, resultado dos esforços contínuos da comunidade acadêmica para aperfeiçoar as técnicas e ferramentas para uma análise precisa da diversidade lexical.

Com os conceitos de riqueza, densidade e diversidade lexical apresentados, bem como seus aspectos históricos e metodológicos, avançaremos para a próxima seção, que abordará as

[10] Hypergeometric distribution D.

aplicações desses indicadores linguísticos no contexto educacional, além das ferramentas disponíveis para sua mensuração.

3 APLICAÇÃO DOS INDICADORES DA RIQUEZA LEXICAL NO CONTEXTO EDUCACIONAL

Nesta seção, conheceremos como os indicadores linguísticos, densidade e diversidade lexical vêm sendo aplicados no contexto educacional e as ferramentas disponíveis para auxiliar os professores e pesquisadores no seu processo de extração.

3.1 Tipos de aplicação

A densidade e diversidade lexical vêm sendo utilizadas no contexto educacional com diferentes propósitos. Entre eles, destacam-se a avaliação da qualidade da escrita dos alunos, a análise do domínio dos diferentes gêneros textuais, a avaliação da progressão escolar e da aprendizagem de língua estrangeira (L2). Vale destacar que esses propósitos convergem para um entendimento mais amplo e profundo do conhecimento lexical dos estudantes, proporcionando uma base sólida para o aprimoramento de práticas pedagógicas voltadas para o ensino em sala de aula.

3.1.1 Avaliação da escrita

A diversidade lexical é um forte preditor da qualidade da escrita (Mellor, 2010). Ela está relacionada com as escolhas lexicais dos alunos, ou seja, com o uso do seu vocabulário. De acordo com Isaacson (1988), a seleção do vocabulário representa um fator essencial no processo de escrita, no qual o aluno precisa selecionar as palavras apropriadas para transmitir a informação que deseja comunicar.

Nesse contexto, a avaliação do domínio do vocabulário é um passo fundamental para a compressão dos conhecimentos linguísticos e habilidades de escrita dos alunos, porém, algumas dificuldades ainda precisam ser superadas, dentre elas o suporte empírico:

> O fraco suporte empírico apresenta um problema diferente para os professores. Não saber como avaliar com precisão as habilidades de escrita de um aluno torna difícil derivar objetivos instrucionais correspondentes. Embora a pontuação holística comumente tenha sido usada para avaliar amostras de escrita, ela não é tão útil do ponto de vista diagnóstico para o professor em sala de aula quanto às

> escalas analíticas ou medidas simples e contáveis do produto de escrita (Isaacson, 1988. p. 533, tradução nossa).[11]

Nesse sentido, podemos compreender que a avaliação tradicionalmente adotada em sala de aula pode falhar em detectar ou enfatizar aspectos específicos da escrita dos alunos, especialmente no contexto em que múltiplos textos são considerados. Diante disso, a aplicação de métricas, como a análise da diversidade e densidade lexical, emerge como uma estratégia valiosa para os professores explorarem mais profundamente essas produções. Além disso, essas métricas podem relevar informações valiosas que poderão ser usadas como suporte para o planejamento de estratégias pedagógicas em sala de aula.

3.1.2 Avaliação do domínio dos diferentes gêneros textuais

Na pesquisa desenvolvida por Olinghouse e Wilson (2012), a diversidade lexical foi usada para avaliar o domínio dos alunos em determinados gêneros textuais escritos. Para isso, foi analisado um conjunto de composições textuais narrativas, persuasivas e informativas de 105 estudantes americanos. Dentre os gêneros produzidos, o narrativo obteve a maior diversidade lexical. Como conclusão do estudo, os autores destacaram que quanto maior o domínio em um determinado gênero textual, maior tende a ser sua diversidade lexical.

Sadeghi e Dilmaghani (2013) realizaram um estudo semelhante sobre a diversidade lexical nos gêneros textuais argumentativos, narrativos e comparativos. Os resultados obtidos também sugeriram uma conexão entre os gêneros textuais e a diversidade lexical, indicando que a variedade de palavras em um texto pode ser afetada pelo tipo de gênero produzido.

Em síntese, esses estudos reforçam a importância de analisar a escrita dos alunos na produção dos diferentes gêneros textuais para diagnosticar quais deles os alunos estão enfrentando mais dificuldades. Dentre as dificuldades que podem ser encontradas durante a produção, destaca-se a repetição de palavras, o que pode influenciar na baixa diversidade lexical do texto. Portanto, a diversidade lexical pode servir como um indicador para os professores investigarem os aspectos que precisam ser potencializados no ensino de um determinado gênero textual.

---

[11] Versão original: *Weak empirical support presents a different problem for teachers. Not knowing how to accurately assess a student's writing skills makes it difficult to derive corresponding instructional objectives. Although holistic scoring commonly has been used to evaluate writing samples, it is not as useful diagnostically to the classroom teacher as analytic scales or simple countable measures of the writing product.*

3.1.3 Avaliação da progressão escolar

Em um trabalho de larga escala, Berman e Verhoeven (2002) avaliaram a diversidade lexical em textos de estudantes de diversos idiomas, como inglês, espanhol, hebraico, francês, islandês, sueco e holandês. Participaram da pesquisa estudantes de quatro níveis de escolaridade: quarto ano (9 e 10 anos de idade); sétimo ano (12 e 13 anos); ensino médio (16 a 17 anos) e estudantes universitários (20 a 30 anos). Os resultados relativos à diversidade lexical, obtidos por meio da medida D, revelaram diferença significativa da diversidade lexical apenas entre os textos dos alunos do sétimo ano e do ensino médio. Também foi observado que o gênero e o nível escolar foram fatores que influenciaram na diversidade lexical. Em suas conclusões, os atores reforçam que a diversidade é uma medida que pode servir como ponto de partida para compreensão dos valores socioculturais, cognitivos e linguísticos envolvidos na construção de um texto (Berman; Verhoeven, 2022).

Seguindo essa mesma linha de investigação, Johansson (2009) conduziu um estudo cujo objetivo era medir a densidade e a diversidade lexical em textos de quatro grupos etários de estudantes suecos. Esses grupos incluíam crianças de 10 anos, adolescentes de 13 e 17 anos e adultos (estudantes universitários). O estudo avaliou um conjunto de composições em dois gêneros textuais diferentes, narrativos e expositivos, e seus resultados apontaram que não houve diferença significativa entre os grupos de estudantes em relação à densidade lexical. Entretanto, para a diversidade lexical foi observada diferença apenas para o grupo estudantes de 13 e 17 anos. Outro achado importante foi a constatação de uma tendência de desenvolvimento mais expressiva para a diversidade lexical do que para a densidade, sugerindo que a diversidade pode ser uma medida mais eficaz para identificar diferenças na habilidade de escrita entre grupos de alunos com faixas etárias distintas.

Partindo para o contexto da investigação em textos de alunos monolíngues do português europeu, temos o estudo de Rodrigues (2008), que adotou a densidade e diversidade lexical para investigar a capacidade de composição da escrita de alunos portugueses do primeiro ao quarto ano escolar, a fim de examinar como suas habilidades de escrita evoluíam ao decorrer desses períodos. Rodrigues observou nos resultados um aumento progressivo na média de palavras escritas pelos alunos: 101 palavras para o primeiro ano, 108 palavras para o segundo ano, 214 palavras para o terceiro e 201 palavras para os alunos do quarto ano. Quanto à diversidade e densidade lexical, as médias para todos os anos de escolaridade superaram 50%, alinhando-se com os achados de Ure (1971), que apontou uma densidade lexical inferior a 40% para textos orais e superior a 40% para os textos escritos. Além disso, Rodrigues constatou que

não foram encontradas diferenças significativas na densidade lexical entre os alunos do segundo e quarto ano, e entre o terceiro e quarto ano.

Martins (2016) também examinou a diversidade lexical na escrita de alunos monolíngues em português europeu. Seu objetivo era avaliar a relação entre a diversidade lexical e a progressão escolar nos textos dos alunos do quinto, sétimo e décimo ano. Utilizando o índice D para analisar a diversidade lexical, foi avaliado um conjunto de textos narrativos e argumentativos. Os resultados revelaram uma correlação positiva entre a diversidade lexical e a progressão escolar em ambos os gêneros textuais, porém, apenas para os alunos do quinto e do décimo ano. Quanto ao valor médio de diversidade lexical para todos os anos escolares, foi observado um valor superior a 70%. No entanto, na comparação entre os grupos, não foi detectada diferença significativa da diversidade entre o quinto e o sétimo ano, como explica Martins (2016):

> Em termos estritamente matemáticos, a variação no uso do vocabulário identificada no quinto ano se assemelha à variação identificada no sétimo ano, havendo mudanças apenas na comparação com a variação no décimo ano da escolaridade, pelo que é possível afirmar que, em algum momento entre o sétimo e o décimo ano dos grupos estudados, se inicia, de fato, um processo de maior diversificação do uso das palavras (Martins, 2016, p. 1079).

O supracitado autor ressalta que embora o vocabulário dos alunos tenda a se tornar mais diversificado à medida que avançam nos anos escolares, esse desenvolvimento não segue um padrão linear. Ao invés disso, ocorrem saltos significativos em fases específicas do percurso escolar.

Em resumo, os estudos sobre progressão escolar aqui reportados sublinham que a diversidade lexical é uma característica da escrita que pode ser influenciada por fatores como idade, nível de escolaridade e gênero textual. Por outro lado, a densidade lexical é mais constante entre as diferentes faixas etárias. Isso implica que a diversidade lexical pode ser um indicador mais eficaz de desenvolvimento da habilidade de escrita, conforme destaca Johansson (2009).

Esses estudos também nos revelam que o desenvolvimento e ampliação do vocabulário não ocorre de maneira linear, mas em momentos específicos ao longo dos anos escolares (Martins, 2016). Esses achados têm implicações significativas para o ensino de língua materna, pois destacam a importância de se dedicar atenção ao desenvolvimento e à evolução do vocabulário dos alunos nos diferentes estágios do seu percurso educacional.

3.1.4 Avaliação da aprendizagem de língua estrangeira

Embora esta pesquisa seja voltada para a análise da língua materna, é necessário reconhecer a relevância desses indicadores para o campo de avaliação da aprendizagem da segunda língua (L2), refletida pela considerável quantidade de pesquisas dedicadas à diversidade e densidade lexical em contextos de L2, seja para a avaliação da proficiência dos estudantes (Wang*, 2014;* Ishikawa; Kondal; 2015; Elgobshawi; Aldawsari, 2022) ou para examinar a qualidade da escrita (Engber, 1995).

O estudo de Ishikawa (2015) propôs-se a examinar o desenvolvimento lexical em textos escritos por estudantes asiáticos que estavam cursando inglês como segunda língua, com o objetivo de avaliar como a densidade e a diversidade lexical se modificavam à medida que a proficiência dos alunos aumentava. Os resultados revelaram que, à medida que o nível de proficiência dos aprendizes avançava, a diversidade lexical diminuía e, posteriormente, aumentava, enquanto a densidade lexical permanecia constante.

Na pesquisa de Elgobshawi e Aldawsari (2022), por sua vez, foi utilizado o método de Ure (1971) para analisar a densidade lexical em redações de dois grupos de estudantes universitários sauditas que estavam cursando inglês como segunda língua. Os resultados apontaram uma maior densidade entre os alunos de níveis mais avançados do curso, sugerindo que há um aumento na densidade lexical à medida que os alunos progridem para estágios mais elevados de estudo.

Em uma perspectiva semelhante, Kondal (2015) também aplicou o método de Ure (1971) para calcular a densidade em textos escritos em inglês por estudantes indianos com idades entre 14 e 16 anos. O objetivo era analisar o impacto da densidade lexical na proficiência em L2 desses aprendizes. Os resultados também indicaram que os textos dos alunos mais avançados apresentavam maior densidade lexical, assemelhando-se aos achados de Elgobshawi e Aldawsari (2022).

Ao analisar de forma concisa esses três estudos, é possível identificar alguns pontos comuns na avaliação em L2, especialmente relacionados à densidade e à diversidade lexical. Em primeiro lugar, destaca-se a valorização desses indicadores para o diagnóstico e acompanhamento da proficiência dos alunos. Em segundo lugar, é notável a adoção da métrica de Ure (1971) no cálculo da densidade lexical, reforçando sua relevância e confiabilidade.

Seguindo os objetivos traçados no presente trabalho, apresenta-se algumas aplicações dos indicadores da riqueza lexical (DeL e DiL), e discute-se sua relevância no contexto escolar. É importante destacar que, de modo geral, essas aplicações apresentadas estão inseridas em

duas principais linhas de estudo: i) avaliação da produção textual em língua materna; ii) avaliação da produção textual em língua estrangeira.

No entanto, apesar do reconhecido potencial desses indicadores para avaliar a riqueza lexical na língua escrita, no âmbito da pesquisa relacionada à língua materna, ainda existem lacunas que precisam ser exploradas. Uma delas é a necessidade de ampliar o conhecimento sobre a riqueza lexical na escrita de alunos brasileiros, assim como a aplicação desses indicadores em diferentes contextos de produção.

Visando contribuir com a produção de conhecimento acerca dessa temática, Santos *et al*. (2018) se propuseram a analisar o comportamento desses dois indicadores a partir de 30 manuscritos escolares produzidos por cinco duplas de alunos brasileiros recém-alfabetizados, matriculados no segundo ano do Ensino Fundamental. Como principais resultados desse estudo inicial, os testes estatísticos aplicados indicaram que os textos produzidos pelas díades não diferiram significativamente em termos de densidade e diversidade lexical, sugerindo que os textos produzidos por alunos de uma mesma sala de aula, submetidos a uma mesma metodologia de escrita, tendem a apresentar características lexicais similares.

Em um estudo posterior, voltado para as ferramentas de extração da riqueza lexical, Santos et al. (2022) ressaltou que uma das razões que pode contribuir para a escassez de pesquisas acerca da riqueza lexical/complexidade lexical na escrita de alunos brasileiros é a falta de ferramentas específicas para extração direta da densidade e diversidade para a língua portuguesa:

> Contudo, apesar do potencial desses indicadores para a avaliação da riqueza lexical dos textos, verifica-se a existência de um número limitado de programas voltados para sua mensuração, especialmente aqueles que adotam métricas atualizadas para o cálculo da diversidade e densidade lexical (Santos *et al*., 2022, p. 1078).

A escassez desses programas adaptados às propriedades especificadas da língua portuguesa pode representar um desafio para os professores e pesquisadores que se dedicam à análise da riqueza lexical da escrita dos alunos. Diante desse cenário, o presente trabalho também busca contribuir com soluções para essa problemática apontada por Santos *et al.* (2022), assim como somar esforços com pesquisadores como e Rodrigues (2008) e Martins (2016) na produção de conhecimento sobre a densidade e diversidade lexical na escrita em língua portuguesa.

Para um maior aprofundamento e compreensão do cenário delineado por Santos *et al*. (2022), é apresentado a seguir um levantamento das ferramentas disponíveis para auxiliar na avaliação da riqueza lexical, por meio da extração de seus indicadores (DeL e DiL).

3.2 Ferramentas disponíveis para extração da DeL e DiL

O uso dos recursos computacionais é um valioso aliado no processo de extração, análise e visualização dos aspectos lexicais presentes nos textos dos alunos. Seja através de técnicas básicas, como a quantificação do número de palavras do texto, ou de técnicas mais complexas, como a classificação morfológica e sintática, usando técnicas de Inteligência Artificial (IA), a tecnologia vem oferecendo ao campo educacional novas possibilidades para o ensino e avaliação da produção textual dos alunos

As métricas para calcular a densidade e a diversidade lexical aqui estudadas têm sido incorporadas em diversas ferramentas computacionais, provenientes de vários domínios de aplicação, principalmente para análise linguística em produções textuais. A seguir, serão apresentadas algumas ferramentas existentes.

3.2.1 *CLAN (Computerized Language Analysis)*

O CLAN é uma ferramenta desenvolvida pelo projeto CHILDES (Whinney, 2000), que oferece diversas funcionalidades para análises linguísticas em língua inglesa, incluindo cálculos de frequência de palavras e comandos para a medição da diversidade lexical. O CLAN é um *software* gratuito e disponível para *download*, mas ao contrário das ferramentas e aplicativos *web* acessíveis via navegador, ele requer a instalação no computador do usuário.

Como *software* especializado em análise conversacional, o CLAN[12] exige que os usuários aprendam a utilizar seus diversos recursos e comandos para efetuar as análises. O programa não oferece funcionalidades para calcular a densidade lexical, apenas a diversidade lexical. Apesar de não suportar a métrica HD-D (*Hypergeometric Distribution-based Diversity*), o CLAN é capaz de realizar a extração da diversidade utilizando a metodologia VOCD por meio de um comando específico. Por fim, devido ao seu foco principal na análise conversacional, que frequentemente envolve diferentes formatos, como áudio, texto e vídeo, sua interface é projetada para trabalhar com textos provenientes de anotações (chat) de uma única ocasião. Ou seja, ele não foi desenvolvido para lidar com múltiplos registros.

[12] Disponível para download em: https://dali.talkbank.org/clan/

3.2.2 TAALED (Tool for the Automatic Analysis of Lexical Diversity)

TAALED[13] é uma ferramenta especializada para a computação da diversidade lexical, usando medidas tradicionais, como a TTR e a RTTR, e abordagens mais atuais, como a HD-D (KYLE et al. 2020). O TAALED é um *software* gratuito e de código aberto que oferece uma interface gráfica de usuário (GUI) intuitiva, compatível com sistemas operacionais Mac e Windows. Além disso, seu código de análise também está acessível como um pacote em *Python*. A versão GUI do TAALED é em inglês e otimizada para trabalhar com pesquisadores proficientes nessa língua, mas a versão em *Python*, como não se trata de um programa, mas de uma biblioteca de funções para desenvolver programas próprios, permite calcular índices de diversidade lexical em programas customizados de diferentes idiomas. Contudo, ele não possui em suas funcionalidades a extração da densidade lexical dos textos.

3.2.3 *Coh-Metrix*

O Coh-Metrix[14] fornece uma análise multidimensional de textos, incluindo a capacidade de medir a diversidade lexical com a implementação de técnicas como a TTR (Mcnamara, 2014). Essa ferramenta também se concentra em outras características textuais, como coesão e coerência, utilizando técnicas de inteligência artificial para análise semântica. Um aspecto relevante do Coh-Metrix é sua adaptação para a língua portuguesa. A versão adaptada, chamada de Coh-Metrix-PTBR, foi desenvolvida com o objetivo de extrair dados linguísticos de textos em língua português, visando aplicações educacionais (Camelo *et al*., 2020).

3.2.4 *Textalyzer*

O *Textalyzer*[15] é uma ferramenta online dedicada à análise de textos para identificar padrões do uso de palavras e analisar a complexidade do vocabulário. O *Textalyzer* examina apenas um texto por vez, não sendo possível comparar vários textos simultaneamente. Dentre suas funcionalidades está a contagem total de palavras, a análise de sentimentos expressos no texto, a diversidade lexical e a densidade lexical, sendo está última apenas para textos em língua inglesa.

---

[13] Disponível em: https://www.linguisticanalysistools.org/taaled.html
[14] Disponível em: http://fw.nilc.icmc.usp.br:23380/cohmetrixport
[15] Disponível em: https://seoscout.com

3.2.5 *VocabProfile*[16]

Outro programa online gratuito para análise linguística é o *VocabProfile*. Ele segmenta o texto em listas de palavras com base em sua frequência e nível de dificuldade. Além disso, realiza a medição da densidade e diversidade lexical, utilizando a métrica TTR. Embora este programa forneça uma abordagem simplificada da medição da diversidade lexical, seu uso é específico para textos em língua inglesa.

3.2.6 *Text Inspector*

O *Text Inspector*[17] é um analisador de texto online, que fornece um leque de recursos analíticos avançados (Bax, 2012). Dentre as funcionalidades oferecidas, destaca-se o cálculo da diversidade lexical, empregando a metodologia VOCD. Sua versão gratuita limita a análise para textos de até 250 palavras (usuário sem cadastro) e 400 palavras (usuários com cadastro). A versão gratuita com cadastro também permite a análise de múltiplos textos, porém, sem oferecer uma análise comparativa entre eles. Caso o usuário necessite realizar a análise de textos extensos, ultrapassando o limite de 250 ou 400 palavras, bem como a exportação dos resultados, é necessário realizar a assinatura do serviço.

3.2.7 *Lexical Complexity Analyzer*

Lexical Complexity Analyzer[18] é um analisador da complexidade lexical para textos em língua inglesa, capaz de computar a diversidade lexical, usando VOCD, e a densidade por meio da razão entre palavras lexicais e gramaticais (Lu, 2012). Operando em Python 3.0, o analisador funciona via interface de linha de comando, requerendo o *download* e a instalação no computador do usuário. No entanto, essa característica pode representar uma barreira, pois exige uma familiaridade com comandos de programação e procedimentos de instalação de *software* em ambientes de desenvolvimento.

[16] Disponível em: https://www.lextutor.ca/vp/eng/
[17] Disponível em: https://textinspector.com
[18] Disponível para download em: https://sites.psu.edu/xxl13/lca/

3.3 Análise comparativa das ferramentas para extração da DeL e DiL

Com a apresentação dessas sete ferramentas, procede-se, agora, com uma análise comparativa entre elas, focando em suas funcionalidades, aplicações e limitações. O Quadro 2 traz uma visão detalhada de cinco fatores observados em cada ferramenta.

Quadro 2 - Comparação das ferramentas.

Ferramentas	Extração da DiL	Extração da DeL	Suporta múltiplos textos	Versão em português	Disponibilidade
CLAN	✓	-	-	-	*Download*
TAALED	✓	-	✓	-	*Download*
Coh-Metrix	✓	-	-	✓	*Download*
Textalyzer	✓	✓	-	-	Online
VocabProfile	✓	✓	-	-	Online
Text Inspector	✓	-	✓	-	Online
Lexical Complexity Analyzer	✓	✓	-	-	*Download*

Fonte: A autora (2024).

Como mostra o quadro comparativo, das setes ferramentas apresentadas, apenas a *Textalyzer, VocabProfile* e *Lexical Complexity Analyzer* oferecem a extração simultânea da densidade e diversidade lexical. Contudo, por elas serem originalmente projetadas para a análise de textos em língua inglesa, isso acaba impondo uma restrição para o seu uso em textos de língua portuguesa. Esse cenário é particularmente evidente no que se refere à extração da densidade lexical, uma vez que os algoritmos envolvidos nessas ferramentas levantadas foram desenvolvidos para realizar a classificação morfológica na língua inglesa.

Dentre todas as ferramentas apresentadas, apenas a ferramenta *Coh-Metrix* possui uma versão adaptada para textos em língua portuguesa. Porém, ela utiliza a medida TTR para calcular a diversidade, que, conforme discutido na seção anterior, é uma medida sensível ao tamanho do texto.

Quanto ao suporte para a avaliação de múltiplos textos, a TAALED e o *Text Inspector* permitem a análise simultânea, embora essa funcionalidade esteja disponível apenas para a análise da diversidade lexical. Por fim, em termos de disponibilidade, há uma divisão entre as ferramentas para *download*/instalação, como o CLAN, TAALED e *Lexical Complexity Analyzer*, e aquelas acessíveis online, como a *Textalyzer, VocabProfile* e *Text Inspector*. Vale

destacar que essa diferença pode impactar na escolha das ferramentas, dependendo do domínio do usuário e da infraestrutura disponível para o seu uso.

O panorama apresentado corrobora com o apontamento de Santos *et al*. (2022), ou seja, as limitações dos *softwares* disponíveis podem ser um grande obstáculo para os professores e pesquisadores que têm interesse na avaliação da riqueza lexical em textos produzidos em língua portuguesa, especialmente em cenários que demandem a análise de múltiplos textos, como é o caso deste estudo.

Diante disso, como alternativa para ajudar nessa problemática, desenvolveu-se o *Lexicanalytics web,* um *software* voltado para a medição da riqueza lexical em textos de língua portuguesa. Na próxima seção, destinada à metodologia de trabalho, são detalhadas suas fases de desenvolvimento e como esse programa pode ser usado em um *corpus* com múltiplos textos.

4 METODOLOGIA

O percurso metodológico da presente tese compreende duas etapas distintas. A primeira etapa apresenta as fases de desenvolvimento da ferramenta *Lexicanalytics web* utilizada para a extração da riqueza lexical dos textos que compõem nosso *corpus*, incluindo as definições de requisitos, arquitetura e o passo a passo de sua utilização. A segunda etapa, por sua vez, aborda o detalhamento do *corpus* analisado, os sujeitos da pesquisa e os procedimentos adotados na coleta de dados, bem com as técnicas para aferição e interpretação dos resultados.

4.1 *Lexicanalytics Web*: etapas de desenvolvimento

Ao considerar o desenvolvimento de um *software*, inicialmente, imagina-se uma tela repleta de códigos. Porém, é fundamental destacar que existem etapas que precedem esse desenvolvimento. Dentre essas etapas, estão a identificação de técnicas e programas existentes, a definição dos requisitos e a criação da arquitetura, todas essenciais para a consistência e execução do programa. A seguir, serão descritas cada uma das etapas que fizeram parte do desenvolvimento do *lexicanalytics web.*

Etapa 1: Mapeamento de técnicas e programas

O mapeamento para identificar os programas existentes foi fundamental para compreender as tecnologias disponíveis e identificar as limitações que poderiam ser aprimoradas nessa nova ferramenta, conforme demonstrado na seção anterior.

Vale destacar que a primeira versão do *Lexicanalytics*[19] começou a ser desenvolvida por Santos *et al*. (2017), inicialmente contando com recursos básicos como a contagem do número de linhas do texto, contagem de palavras e a mensuração da DiL por meio da TTR. Com o aprofundamento na bibliografia especializada e com o mapeamento das ferramentas existentes, foi possível aprimorar o programa e adicionar funcionalidades mais complexas para a extração da riqueza lexical.

No entanto, essas funcionalidades demandaram a adoção de técnicas mais avançadas, como o Processamento de Linguagem Natural (PLN), um ramo da Inteligência Artificial (IA)

[19] As imagens da primeira versão estão disponíveis no ANEXO 3 deste trabalho.

que busca aprender, entender e produzir conteúdo na linguagem humana, a fim de interpretá-la e automatizá-la (Filatro, 2021).

Conforme mencionado anteriormente, a extração da densidade lexical tem como ponto de partida a classificação morfológica das palavras para diferenciá-las de acordo com suas propriedades lexicais (substantivos, adjetivos, verbos e advérbios) ou funcionais (artigos, pronomes, numerais etc.). Para automatização desse processo de identificação e classificação morfológica no l*exicanalytics,* usou-se algumas técnicas de PNL, como a *Part of Speech* (*POS-Tagging),* que tem como tarefa associar cada palavra a sua classe gramatica*l.*

Etapa 2: Análise e definição dos requisitos

Os requisitos são descrições dos recursos específicos que serão fornecidos pelo programa (Sommerville, 2015). Essa etapa de análise e definição é fundamental, não apenas para estabelecer esses recursos, mas também para estabelecer as restrições quanto à operação do sistema e quanto aos processos de desenvolvimento. Diante disso, foram estabelecidos os seguintes requisitos para o *lexicanalytics web:*

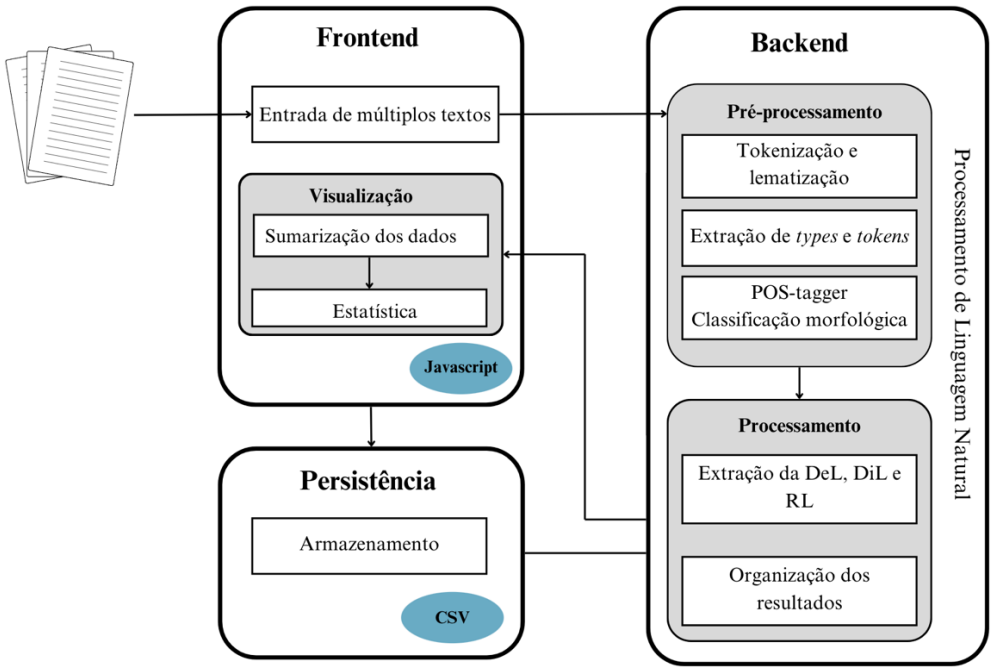
1. Suporte para textos em língua portuguesa;
2. Importação de múltiplos textos;
3. Contagem de palavras do texto (*tokens)*;
4. Computação do vocabulário (*types)*;
5. Mensuração da Densidade Lexical;
6. Mensuração da Diversidade Lexical;
7. Mensuração da Riqueza Lexical;
8. Comparação dos resultados de múltiplos textos;
9. Exportação dos dados para análises aprofundadas (estatística inferencial);
10. Funcionamento em interface *web* (para não necessitar de instalação local);
11. Armazenamento persistente das produções, garantindo assim manter análises feitas pelo usuário.

Etapa 3: Definição da arquitetura

Após a definição dos requisitos, foi iniciado o projeto de arquitetura de *software*, sendo esta definida como a estrutura composta pelos diferentes componentes de um programa, suas

inter-relações, bem como os princípios e diretrizes que orientam o design e a evolução do *software* (Sommerville, 2015). Cabe frisar que essa arquitetura desempenha um papel essencial na concepção de como o programa atenderá os requisitos funcionais previamente estabelecidos. A figura 2 apresenta a arquitetura projetada para o *lexicanalytics web*.

Figura 2 - Arquitetura do *Lexicanalytics Web.*



Fonte: A autora (2024).

Entre os elementos dessa estrutura, destaca-se o *backend*, um componente fundamental do sistema. Sua função é gerenciar o pré-processamento do texto (antes da mensuração da DeL e DiL) e o processamento propriamente dito (mensuração da DeL e DiL), além de gerenciar o armazenamento das informações extraídas por meio de um serviço de persistência.

No pré-processamento, são realizadas a tokenização e a lematização; a extração de *types* e *tokens*; e a classificação morfológica das palavras. A tokenização é o processo no qual todas as palavras do texto são separadas, ocorrendo a remoção dos caracteres como pontos, vírgulas, ponto de interrogação, ponto de exclamação, aspas, reticências, parênteses, chaves, colchetes, dentre outros. Já na lematização, as palavras são analisadas e agrupadas de acordo com suas formas flexionadas. Essa fase é fundamental para o sistema diferenciar os *types,* que representam a primeira ocorrência de cada palavra no texto, e os *tokens,* que correspondem ao total de palavras do texto. Em seguida, o programa faz a contagem do número de *types* e *tokens* de cada texto.

Para concluir, o sistema realiza a classificação morfológica das palavras, usando como aporte o *POS-tagger*, um classificador especializado para a língua portuguesa, treinado com base em técnicas de Processamento de Linguagem Natural. Esse classificador é fundamental para a realização da extração da densidade lexical dos textos.

Com a finalização do pré-processamento, os textos avançam para a etapa de processamento. Nesta fase, são extraídas: a densidade lexical, calculada pela razão entre os itens lexicais (substantivos, verbos, adjetivos e advérbios de modo terminados em -mente) e o total de palavras do texto; a diversidade lexical, utilizando o algoritmo HD-D, e, por fim, a riqueza lexical. A partir disso, o *backend* organiza os dados e os encaminham para o *frontend*, que é um componente do sistema responsável pela exibição dos resultados processados para o usuário por meio de tabelas e gráficos.

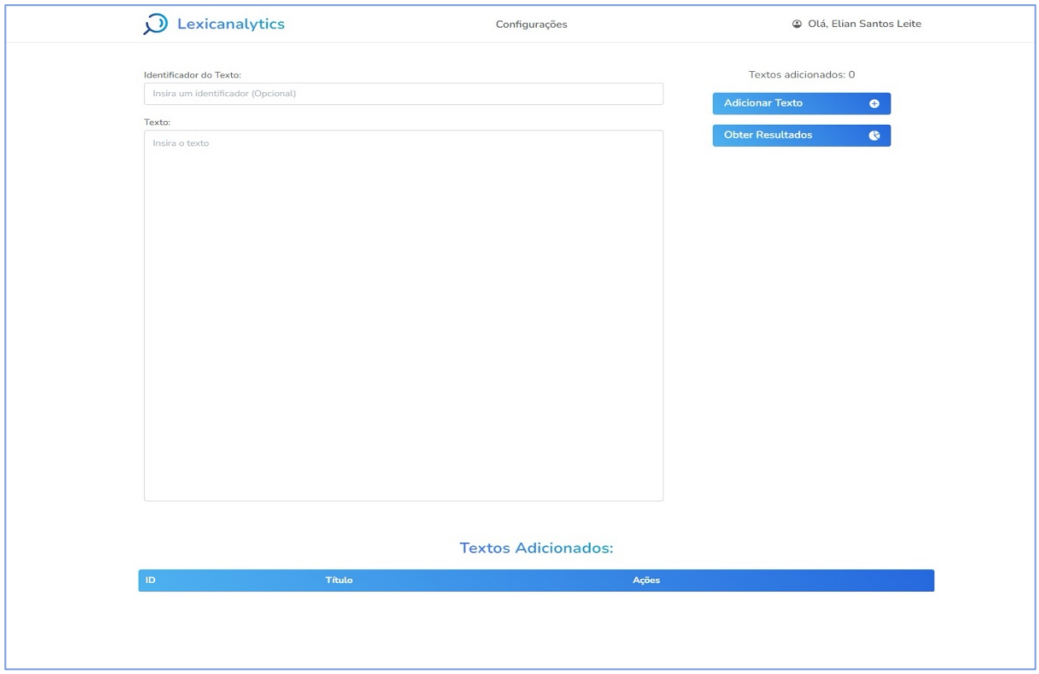
Na última etapa do processo, todas as informações geradas no *Lexicanalytics web* são encaminhadas e armazenadas na persistência, para estarem disponíveis para acessos futuros, sem a necessidade de reprocessamento.

4.1.1 *Lexicanalytics web*: passo a passo da avaliação textual

Para demonstrar a execução de uma avaliação textual usando o *Lexicanalytics web*, selecionou-se um conjunto de manuscritos escolares pertencentes ao banco de dados do LAME. É importante salientar que os resultados aqui reportados são provenientes da versão funcional do programa, executada em um navegador de acesso restrito para os seus desenvolvedores. Dito isto, para iniciar o processo de avaliação é necessário, inicialmente, inserir a versão do texto transcrito (Figura 3). Essa ação pode ser realizada apenas copiando o texto e colando-o na aba inicial do programa. Em seguida, é necessário selecionar a opção “Adicionar texto”.

O programa também conta com o campo denominado “Identificador do texto”, que serve como uma legenda para nomear os textos adicionados, facilitando o controle da análise e a intepretação dos resultados obtidos. Para uma análise de múltiplos textos, cada texto deve ser adicionado por vez. À medida que essa ação é executada, eles vão sendo exibidos de forma ordenada na parte inferior da tela. Caso algum texto seja adicionado erroneamente, há duas ações disponíveis: editar legenda ou remover o texto.

Figura 3 - Tela inicial do *Lexicanalytics Web.*

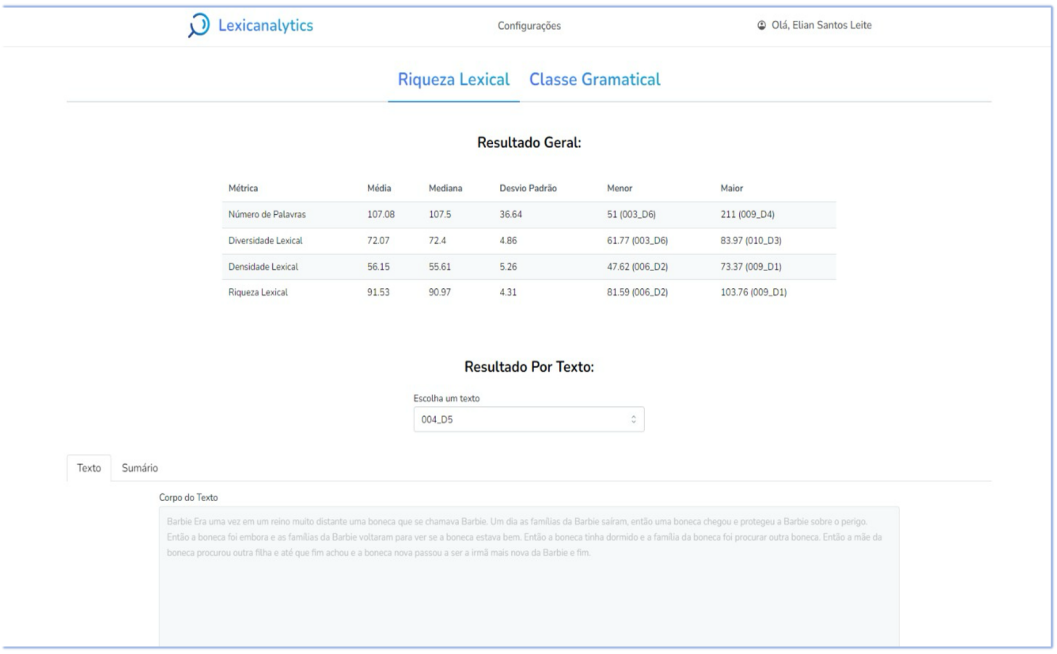


Fonte: A autora (2024).

Com os textos adicionados, ao clicar no campo “Obter resultados”, é exibida uma nova tela com os resultados gerais da riqueza lexical, da densidade lexical, da diversidade lexical e do número de palavras. Esses resultados são apresentados de forma estatística por meio de medidas de posição e variabilidade (Figura 4).

Cabe pontuar que as medidas de posição, também conhecidas como medidas de primeira ordem, selecionam um valor representativo do conjunto de dados analisados. Entre elas estão a média aritmética e a mediana, conforme explicam Bussad e Morettin (2017). No entanto, utilizar apenas uma dessas medidas para representar todos os dados pode não revelar os aspectos importantes da distribuição ou da dispersão do conjunto de dados analisados. Por isso, recorre-se também às medidas de variabilidade ou medidas de segunda ordem, como o desvio padrão. Essas medidas ajudam a compreender melhor a variação e a dispersão dos dados, fornecendo uma visão mais detalhada das características dos textos analisados.

Figura 4 - Tela de Resultado Geral do *Lexicanalytics Web.*

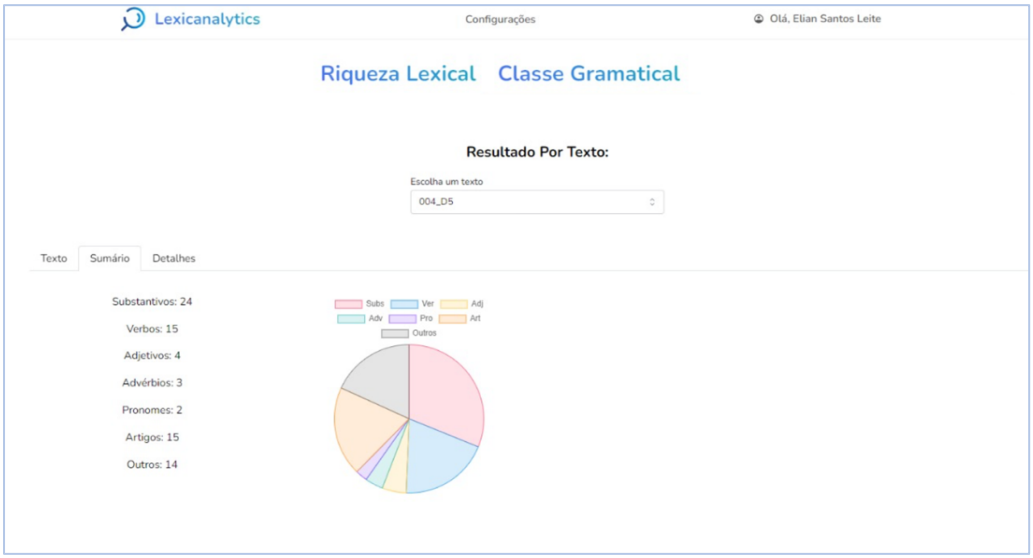


Fonte: A autora (2024).

Outra informação importante gerada na tela de resultados é a classificação dos textos com base nos maiores e menores valores obtidos. Essa funcionalidade permite identificar facilmente, por exemplo, qual texto apresentou a maior densidade lexical e qual registrou a menor densidade. Além de proporciona uma visão panorâmica dos aspectos lexicais de um conjunto de textos, o programa gera os resultados individuais para cada produção, contribuindo para o acompanhamento da progressão da riqueza lexical dos alunos.

O programa possui um sumário que reúne informações referente às classes gramaticais utilizadas em cada texto. Essas informações são apresentadas de forma gráfica e por meio de um resumo quantitativo (Figura 5). As classes gramaticais como preposição, numeral, conjunção e interjeição são agrupadas na categoria “Outros” dentro do resumo quantitativo, conforme representado a seguir.

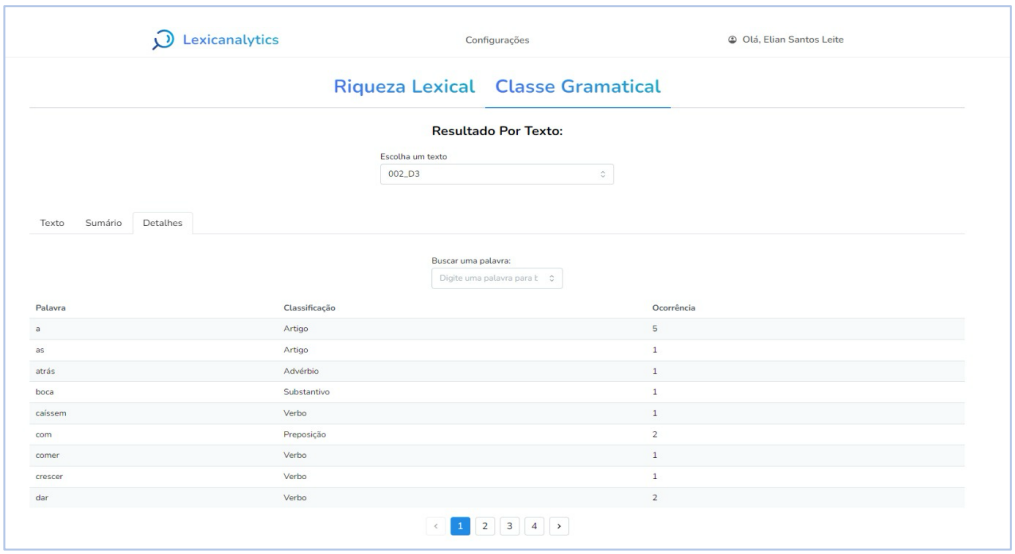
Figura 5 - Tela de Resultado por classe gramatical.



Fonte: A autora (2024).

Na aba “Detalhes”, os usuários têm acesso à classificação morfológica e à ocorrência de cada palavra no texto, o que permite uma verificação precisa das palavras mais utilizadas em cada produção textual ou por classe gramatical (Figura 6). Além disso, nessa mesma aba, está disponível a opção de “Buscar”, que viabiliza o filtro para analisar uma palavra específica. Esse recurso não apenas facilita o mapeamento das ocorrências de cada palavra, mas também permite verificar a precisão da classificação realizada pelo sistema para cada palavra, garantindo uma análise mais confiável.

Figura 6 - Tela de ocorrência de palavras.



Fonte: A autora (2024).

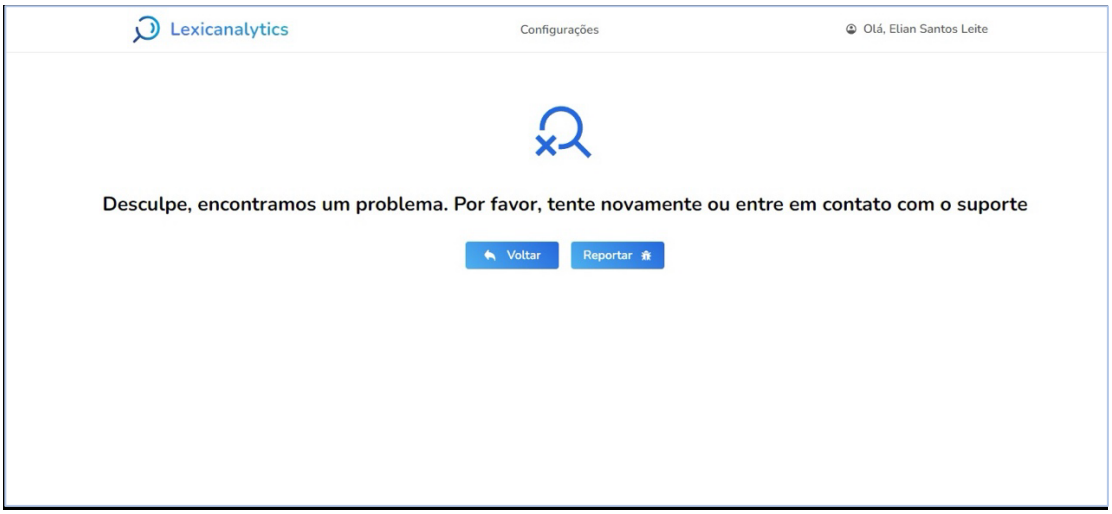
Em síntese, as informações extraídas pelo *lexicanalytics web* podem auxiliar no diagnóstico dos aspectos da escrita dos alunos que necessitam de uma maior atenção. A partir disso, os professores podem desenvolver atividades e estratégias de ensino voltadas para a melhoria das habilidades de escritas e para a ampliação do repertório lexical dos alunos.

4.1.2 *Lexicanalytics Web*: etapa de validação

Para validação dessa ferramenta foi realizada uma etapa de testes utilizando textos escolares pertencentes ao banco de dados do LAME. Essa etapa foi essencial para a identificação e correção de eventuais problemas que os usuários poderiam encontrar durante o processo de extração das informações lexicais dos textos.

Para promover a identificação de eventuais erros do programa, foi implementada uma configuração para o reconhecimento das palavras. Com isso, caso alguma palavra do texto inserido para analise não fosse identificada pelo programa, ele não seguiria para a etapa de resultados e, posteriormente, seria exibida uma mensagem sinalizando o erro (Figura 7). Essa funcionalidade se mostrou fundamental durante a fase de testes, pois permitiu uma rápida detecção do problema. Além disso, foi implementada na configuração do programa a opção "Reportar problemas", que encaminha o erro identificado diretamente para a equipe de desenvolvedores.

Figura 7 - Tela com mensagem de erro.



Fonte: A autora (2024).

A mensagem "Volte à página e tente novamente", exibida nessa tela de erro (Figura 7), atua como um alerta para a necessidade de revisar o texto inserido. A maioria dos erros

detectados durante essa etapa de testes eram relacionados ao reconhecimento de alguns caracteres. O quadro 3 detalha esses erros:

Quadro 3 -Tipos de erros.

Tipos de erro	Detalhamento
Palavras com travessão junto	-Eu sou Branca de Neve
Palavras entre aspas	“Os dinossauros"
Número decimal 0,5	Ele perdeu 0,5 pontos
Texto com pontuação repetida	E falou… Fim!!!
Palavras entre colchetes	[Falou]
Palavras entre chaves	{Carro}

Fonte: A autora (2024).

É importante ressaltar que, sendo um sistema baseado em Inteligência Artificial (IA), o programa é treinado para classificar as palavras por meio de técnicas de aprendizagem de máquina. Nesse sentido, há ocasiões em que certas palavras ou elementos textuais, como sinais de pontuação e caracteres especiais, podem não ser imediatamente reconhecidos.

Para lidar com essa questão, uma de nossas prioridades foi ensinar continuamente o programa a identificar e processar adequadamente esses elementos para que a presença deles não interrompesse ou prejudicasse o processo de análise. Após essa etapa de testes, com os problemas identificados, realizamos uma série de ajustes no filtro do programa para corrigi-lo. Esses ajustes não apenas aumentaram a funcionalidade do programa, mas também melhoraram sua precisão e confiabilidade.

Em suma, o *Lexicanalytics web* se mostrou uma ferramenta de grande potencial para auxiliar os professores e pesquisadores que desejam avaliar a riqueza lexical dos textos dos alunos produzidos em língua portuguesa, pois, diferente de outros programas já apresentados, ele se destaca pela capacidade de extrair tanto a densidade e diversidade lexical, quanto a riqueza lexical, além de suportar a avaliação de múltiplos textos de forma simultânea, oferecendo resultados comparativos entre eles.

Outro fator que merece destaque é que sua projeção para estar disponível online elimina a necessidade de assinaturas de serviços ou de instalação complexa, facilitando seu acesso e uso. Além disso, sua interface intuitiva busca simplificar o processo de análise, tornando-o mais simples e direto.

4.2 Procedimentos metodológicos para análise dos dados

Nesta segunda etapa da metodologia, focaremos na apresentação do *corpus* do presente trabalho e no detalhamento dos passos seguidos para a análise dos textos. Como ponto de partida, é fundamental destacar que este trabalho possui natureza exploratória (Gil, 2022). Desse modo, para a análise dos dados, utilizou-se as seguintes abordagens:

1. Quantitativa: para extração e mensuração das categorias de análise, assim como o uso de instrumentos estatísticos para auxiliar na organização e interpretação dos dados.
2. Qualitativa: para descrever, explicar e refletir sobre os resultados obtidos e suas implicações para o ensino da produção textual em sala de aula.

4.2.1 Natureza do *corpus*

O *corpus* desta pesquisa é fruto de uma colaboração internacional, envolvendo pesquisadores do Brasil, representados pelo Laboratório do Manuscrito Escolar (LAME) e pesquisadores da Universidade de Aveiro, em Portugal. Esse conjunto de textos foi coletado no âmbito do projeto *InterWriting* entre os anos de 2012 e 2015. O projeto teve como foco principal o intercâmbio educacional entre Brasil e Portugal, visando a criação de um banco de dados comum, centrado no processo de escrita infantil.

4.2.2 Escolas participantes

A seleção das escolas foi conduzida pelos coordenadores do projeto, tendo como critério de seleção: escolas que atendessem alunos recém-alfabetizados, com faixa etária de seis a oito anos; e escolas que tivessem interesse e disposição para participar do estudo.

No Brasil, a coleta aconteceu no ano de 2012, em uma escola da rede privada, que atendia alunos do maternal ao 5º ano do Ensino Fundamental, localizada na cidade de Maceió/Alagoas. Em Portugal, a coleta ocorreu no ano de 2015 em duas escolas públicas de Aveiro, sendo uma escola localizada na zona urbana e a outra na zona rural. Ambas atendiam alunos desde o Jardim de Infância até o 1º Ciclo do Ensino Básico.[20]

4.2.3 Sujeitos da Pesquisa

[20] Compreende o primeiro dos três ciclos que constituem o Ensino básico de Portugal. Esse ciclo tem duração de quatro anos e atende crianças do 1º ao 4º ano de escolaridade.

Os sujeitos da pesquisa são crianças brasileiras e portuguesas recém-alfabetizadas, com idade entre seis e oito anos. Em relação ao nível escolar, no Brasil, as crianças estavam matriculadas no primeiro e segundo ano do Ensino Fundamental, enquanto as crianças portuguesas cursavam o segundo ano do Ciclo do Ensino Básico.

4.2.4 Processo de coleta de dados

Em ambos os países, a coleta foi conduzida por integrantes do LAME, cada uma com duração de aproximadamente três meses. A metodologia de escrita adotada para produção dos textos foi baseada na escrita colaborativa em pares[21], uma metodologia que vem sendo aplicada em diferentes estudos desenvolvidos pelos LAME (Calil, 2012; Braga; Lira, 2017; Queiroz; Silva, 2019; Santos, 2023). Uma das motivações para o uso dessa metodologia é a descoberta de que os estudantes tendem a escrever melhor quando trabalham em duplas, pois a colaboração entre eles facilita a mediação e a articulação de ideias durante o processo de escrita (Daiute; Dalton, 1992; Vass, 2002).

Outro aspecto relevante da coleta foi a preservação das condições ecológicas de sala de aula, assegurando que os aspectos ambientais, curriculares e interacionais fossem integralmente respeitados, assim como a prática didática da professora, no que se refere às atividades de produção textual.

Seguindo esse protocolo pedagógico, os alunos foram agrupados em díades para escreverem juntos uma história inventada (narrativa ficcional), assumindo um deles a função de "escrevente" (aquele responsável pelo registro na folha de papel) e o outro a função de "ditante" (responsável por ditar o texto a ser escrito). Desse modo, uma vez por semana, cada díade produzia um texto a partir de temas sugeridos pela professora ou de temas de livre escolha. Salienta-se que para a formação das díades, os alunos tiveram liberdade para escolher seu/sua parceiro(a).

Para garantir que cada componente das díades tivesse as mesmas oportunidades de escrita, ao decorrer de cada produção textual havia uma alternância da função de escrevente e ditante. No entanto, a relação entre essas variáveis ("escrevente e ditante", "temas livre ou sugeridos") não será discutida neste estudo, pois sua ênfase se concentra apenas na avaliação da riqueza lexical dos manuscritos produzidos nos dois países.

[21] Esta técnica de coleta de dados, a partir da escrita em díades, vem sendo desenvolvida por Calil desde 1989 e aproxima-se metodologicamente dos trabalhos desenvolvidos por um grupo interdisciplinar de pesquisadores dedicados a analisar o processo de escritura a dois, nomeados como redação conversacional (Bouchard; Mondada, 2010).

Após as coletas, os textos passaram por uma etapa de catalogação e, atualmente, encontram-se disponíveis no acervo do LAME para os seus membros, servindo de matéria-prima para o desenvolvimento de diferentes estudos sobre a escrita de alunos recém-alfabetizados.

4.2.5 Especificação do *corpus*

O *corpus* analisado no presente trabalho de doutoramento é composto por 186 manuscritos, sendo 86 produzidos por díades de alunos brasileiros e 100 produzidos por díades de alunos portugueses, conforme especifica a tabela a seguir:

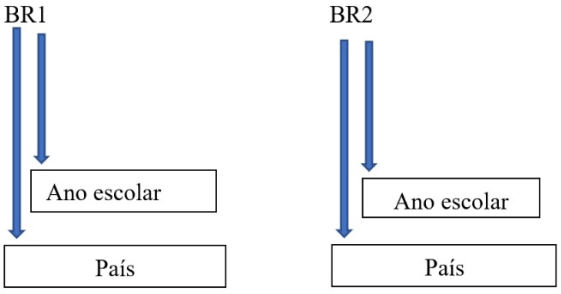
Tabela 1 - Especificação do *corpus* de acordo com o país de origem.

País de origem	**Ano escolar**	**Nº de díades**	**Nº de textos**
Brasil	1º ano	12	36
Brasil	2º ano	10	50
Portugal	2° ano (Escola zona urbana)	20	50
Portugal	2° ano (Escola zona rural)	20	50

Fonte: Acervo do LAME (2024).

Para garantir um maior controle e organização dos dados ao longo da análise, realizamos a catalogação de todos os textos. Essa catalogação foi fundamental para facilitar o acesso e a recuperação das informações. Para o *corpus* brasileiro, utilizamos a seguinte legenda para diferenciar as díades do primeiro e do segundo ano do Ensino Fundamental:

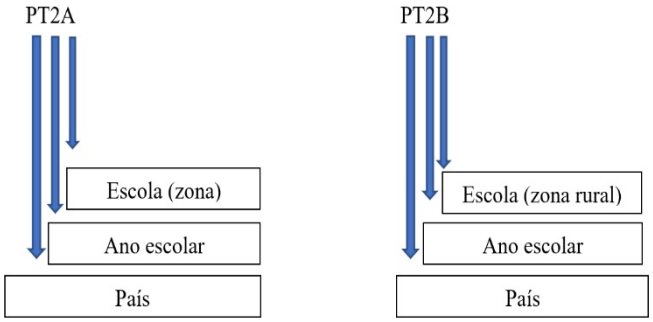
Figura 8 – Legenda para os textos das díades brasileiras.



Fonte: A autora (2024).

Ao contrário do Brasil, os textos das díades portuguesas foram produzidos por estudantes que cursavam o mesmo ano escolar, mas pertenciam a duas escolas diferentes. Para organizar esses textos, foi adotado o seguinte sistema de legenda visando refletir esta diferenciação entre as instituições.

Figura 9 – Legenda para os textos das díades portuguesas



Fonte: A autora (2024).

4.2.6 Categorias de análise

Serão verificadas em cada texto as seguintes categorias/variáveis de análise:

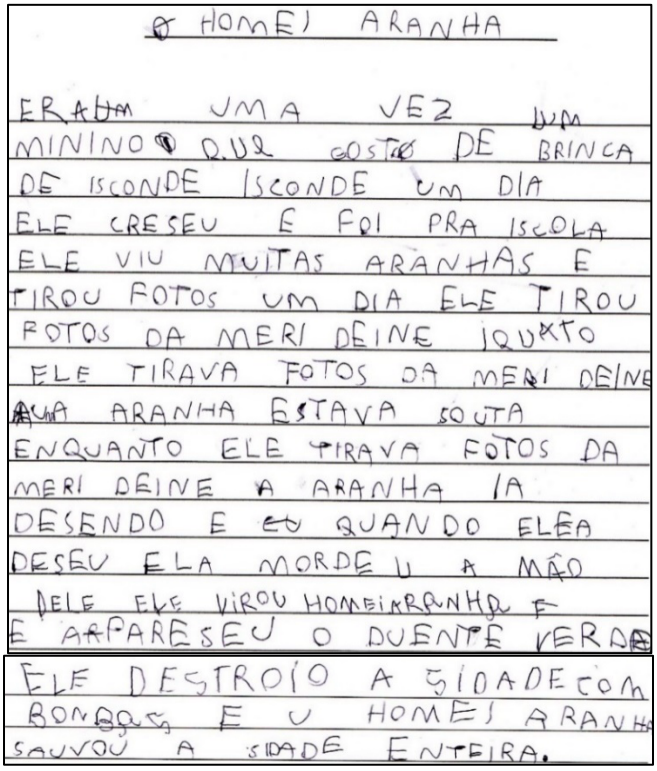
- Total de Palavras Escritas (PE): representa o total de palavras escritas no texto, considerando todas as ocorrências;
- Diversidade Lexical (DiL): corresponde à variedade de palavras presente no texto (Mccarthy; Jarvis, 2007);
- Densidade Lexical (DeL): refere-se à razão de palavras lexicais pelo total de palavras do texto (Ure, 1971), considerando como palavras lexicais os substantivos, verbos, adjetivos e advérbios de modo que terminam em "-mente".
- Riqueza Lexical (RL): composta por indicadores linguísticos como a densidade e diversidade lexical (Read, 2000).

4.2.7 Extração e análise dos dados

Para extração das categorias de análise por meio do lexicanalytics, os 186 textos passaram por uma transcrição normativa, utilizou-se um *software* de edição (*word*). Essa transcrição procedeu da seguinte maneira: i) leitura do texto; ii) mapeamento das rasuras, das palavras separadas e erros ortográficos; iii) transcrição com os ajustes. Para demonstrar o funcionamento desse processo, apresenta-se um texto produzido por uma díade brasileira do 1º

ano (Figura 10). Nele, é possível acompanhar todas as marcas gráficas deixadas no texto original.

Figura 10 - Manuscrito de uma díade brasileira do 1º ano do Ensino Fundamental.



Fonte: Acervo do LAME (2012).

Por se tratar de manuscritos de alunos recém-alfabetizados, nota-se a presença recorrente de marcas de oralidade na grafia de algumas palavras, como “iscola”, “isconde”, “homei”. Ademais, também foram observadas rasuras e a ausência de letras em algumas palavras, como “deseu” e “desendo”. Na etapa de leitura, todas essas ocorrências foram identificadas.

Todas as transcrições foram cuidadosamente iniciadas a partir do título de cada texto, assegurando que o número de palavras escritas permanece inalterado, sem qualquer exclusão ou inclusão de palavras. É importante mencionar que, devido à ausência de tecnologias de transcrição automática no momento da transcrição desses manuscritos, o processo acabou sendo demorado, especialmente devido ao volume do *corpus*. Na Figura 11, é possível conferir o mesmo texto após a sua transcrição.

Figura 11 - Transcrição do texto produzido por uma díade brasileira do 1º ano do Ensino Fundamental.

O homem-aranha

Era uma vez um menino
que gostava de brincar
de esconde-esconde. Um dia
ele cresceu e foi para escola
ele viu muitas aranhas e
tirou fotos. Um dia ele tirou
fotos da Mary Jane enquanto
ele tirava fotos da Mary Jane,
uma aranha estava solta
enquanto ele tirava fotos da
Mary Jane a aranha ia
descendo e quando ela
desceu ela mordeu a mão
dele. Ele virou homem-aranha
e apareceu o duende verde.
Ele destruiu a cidade com
bombas e o homem-aranha
salvou a cidade inteira.

Fonte: A autora (2024).

Após a fase de transcrição, utilizou-se o *Lexicanalytics Web* para extrair a densidade, diversidade e riqueza lexical dos textos. A densidade foi medida pelo método de Ure (1971), enquanto a diversidade foi mensurada utilizando o índice HD-D (*Hypergeometric Distribution-based Diversity).* Optou-se por utilizar a medida HD-D, devido à sua maior precisão e menor complexidade computacional em comparação com outras abordagens existentes (Mccarthy; Jarvis, 2007).

Para calcular a riqueza lexical dos textos e alcançar uma comparação mais segura e equitativa entre eles, adotou-se uma técnica de normalização para a DeL e DiL, representada pela fórmula matemática:

$$RL = \sqrt{DiL^2 + DeL^2} \qquad (5)$$

Essa fórmula é conhecida como norma euclidiana. Ela combina os dois indicadores (DeL e DiL) para calcular uma medida unificada de riqueza lexical. Com isso, foi possível medir e avaliar a riqueza das produções de forma mais segura e direta, possibilitando também

a comparação entre os textos de diferentes comprimentos e valores de DeL e DiL, e ordenando-os com base em sua riqueza lexical.

Com a finalização do processo de mensuração de PE, DiL, DeL e RL, os valores foram organizados em tabelas para a análise de estatística inferencial, adotando como aporte teórico Magalhães e Lima (2010), Morettin e Bussab (2010). Esse procedimento foi necessário para o contexto desta investigação, tendo em vista auxiliar a responder a questão de pesquisa, especialmente considerando que o *lexicanalytics web* só conta com a análise de estatística descritiva.

De acordo com Magalhães e Lima (2010), a estatística inferencial é uma técnica que possibilita a exploração de um grande conjunto de dados, a partir das informações e conclusões obtidas de um subgrupo de valores, usualmente de dimensão muito menor. Neste trabalho, foram utilizadas as técnicas de estatística inferencial para estimar a existência de um comportamento padrão das nossas categorias de análise, além de verificar se há uma diferença significativa entre os textos analisados.

Desse modo, foi aplicado o coeficiente de correlação linear (r), também conhecido como correlação de Pearson, para quantificar a força da relação linear entre as variáveis estudadas, o teste de normalidade *Shapiro-Wilk*, para identificar se os dados seguiam ou não uma distribuição de probabilidade normal (ou gaussiana), e o teste paramétrico ANOVA, para avaliar a diferença da riqueza lexical nos textos das díades portuguesa e brasileiras. Para tanto, adotou-se um nível de significância $\alpha = 0{,}05$, se $p > 0{,}05$, não rejeitando-se a hipótese nula, já se $p \leq 0{,}05$, a hipótese nula é rejeitada.

Para responder a questão de pesquisa, foram estruturados 3 cenários paras as hipóteses:

1. Hipóteses para o conjunto de textos das díades brasileiras.

- **Hipótese nula (H0BR):** Não há diferença significativa da Riqueza Lexical entre os textos das díades brasileiras (BR1 e BR2).
- **Hipótese alternativa (HaBR):** Há diferença significativa da Riqueza Lexical entre os textos das díades brasileiras (BR1 e BR2).

2. Hipóteses para o conjunto de textos das díades portuguesas.

- **Hipótese nula (H0PT):** Não há diferença significativa da Riqueza Lexical entre os textos das díades portuguesas (PT2A e PT2B).
- **Hipótese alternativa (HaPT):** Há diferença significativa da Riqueza Lexical entre os textos das díades portuguesas. (PT2A e PT2B).

3. Hipóteses para comparação da Riqueza Lexical entre textos de díades brasileiras e portuguesas.

- **Hipótese nula ($H0$):** Não há diferença significativa da Riqueza Lexical entre os textos das díades brasileiras e portuguesas.
- **Hipótese alternativa (Ha):** Há diferença significativa da Riqueza Lexical entre os textos das díades brasileiras e portuguesas.

4.2.8 Visualização dos resultados da análise de estatística inferencial

O tratamento, a computação e a visualização dos dados foram feitas por meio da linguagem *Python*[22], através do *software Google Colab*[23]. Dentre as bibliotecas que funcionam como extensões do *Python*, foram usados *NumPy, SciPy* e *Pandas*[24] para manipular os dados brutos e extrair medidas estatísticas, desde as mais básicas como a média e mediana até as mais complexas, como a correlação e testes de hipótese para inferência. Além disso, foram usados o *Matplotlib* e o *Seaborn*[25] para visualização dos gráficos.

[22] Linguagem de programação.
[23] É um ambiente de desenvolvimento integrado baseado em nuvem para a linguagem de programação *Python.*
[24] são bibliotecas de software de código aberto para a linguagem de programação *Python.*
[25] *Matplotlib* e *Seaborn* são bibliotecas de visualização de dados para a linguagem de programação *Python.*

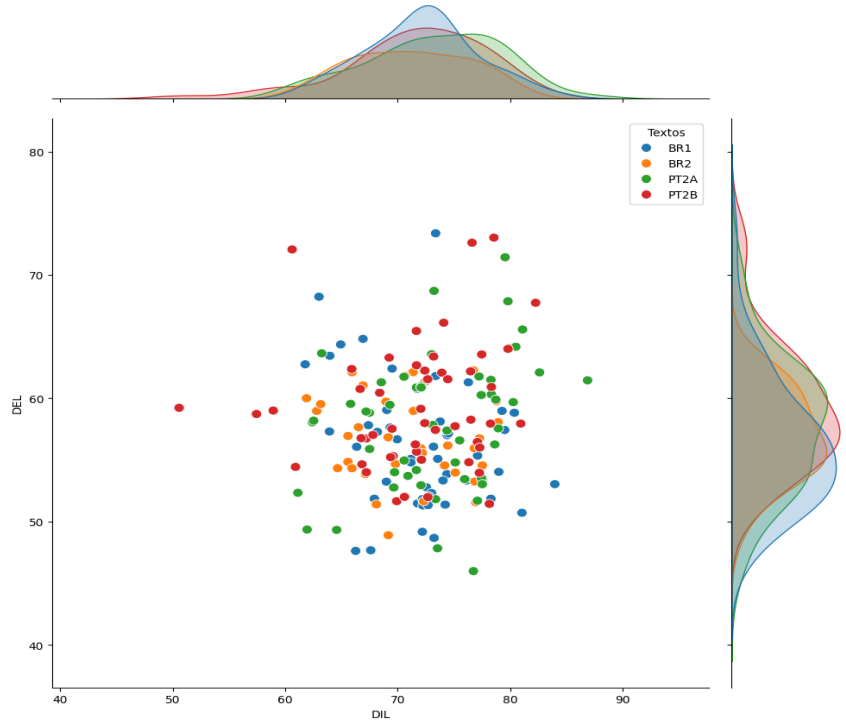
5 RESULTADOS E DISCUSSÕES

Nesta seção, apresenta-se os resultados alcançados com base nos objetivos e hipóteses previamente estabelecidos. Também se discute os principais achados e analisa-se como eles contribuem para o avanço do conhecimento no campo de estudo em questão. Ressalta-se que para tornar a análise computacional mais clara e organizada, adotou-se nomenclaturas específicas para os conjuntos de dados das diferentes turmas, escolas e países. Desse modo, os dados das produções dos alunos brasileiros do 1º ano foram nomeados como BR1. Já as produções dos alunos do 2º ano da mesma escola brasileira, como BR2. De maneira similar, os dados referentes às duas turmas portuguesas de alunos do 2º ano foram identificados pelas respectivas siglas PT2A e PT2B.

5.1 Panorama geral dos textos produzidos pelas díades brasileiras e portuguesas

Como introdução aos resultados, oferece-se uma visão ampla dos textos produzidos por todas as díades. No Gráfico 2, pode ser observada a distribuição geral das produções, em que cada ponto no gráfico representa um texto e seus respectivos valores de DiL e DeL.

Gráfico 2 - Gráfico de dispersão de DeL e DiL.



Fonte: A autora (2024).

Nos eixos da DeL e DiL, presente no gráfico, pode-se observar que há uma maior concentração de textos com densidade lexical com valores de 50% e 60%. Do mesmo modo, há uma maior concentração de textos com diversidade lexical com valores entre 60% e 80%.

Usando esses conceitos representados no gráfico, é possível identificar que a produção com a maior diversidade lexical foi produzida por uma díade portuguesa pertencente ao grupo PT2A, indicada pela cor verde, e a produção com a menor DiL foi escrita por uma díade de PT2B, representada na cor vermelha. Quanto à densidade lexical, a maior DeL foi produzida por uma díade brasileira (BR1), em azul, e a menor DeL foi identificada na produção de uma díade portuguesa (PT2A).

5.2 Medidas descritivas para os conjuntos de textos

As medidas-resumo ferentes aos 186 textos analisados estão dispostas na Tabela 2. Vale destacar que essas medidas são fundamentais para a compressão das características gerais da escrita das díades.

Tabela 2 - Resultados gerais das produções de todas as díades.

	PE	**DIL (%)**	**DEL (%)**
Média	106,46	72,08	57,62
DP	35,27	5,62	5,08
Mín.	51,00	50,56	45,98
25%	79,25	68,43	54,00
50%	102,0	72,28	57,34
75%	126,75	76,70	60,96
Máx.	252,0	86,88	73,37

Fonte: A autora (2024).

A média geral de palavras escritas por produção foi de 106,46 PE, com um desvio padrão de 35,27, sugerindo uma variação considerável no comprimento dos textos. O menor texto tinha 51 palavras, enquanto o maior atingiu 252 palavras.

Quanto à DiL, a média computada foi de 72,08%, com um desvio padrão de 5,62, o que aponta para uma consistência na diversidade lexical entre os textos. O valor mínimo para DiL foi de 50,56% e o máximo 86,88%. Os percentis (25%, 50%, 75%) também mostram uma distribuição equilibrada da DiL entre os textos analisados.

Vale destacar que esses resultados de PE e DiL evidenciam a importância do uso de métricas resistentes ao comprimento do texto, como a HD-D (Mccarthy; Jarvis, 2007). Como é possível constatar nos resultados ilustrados na Tabela 2, apesar da acentuada diferença entre o valor máximo e mínimo de PE, o desvio padrão da DiL foi baixo (5,62), indicando, de maneira geral, apenas uma leve variação nos textos.

Quanto à DeL, os textos apresentaram uma média de 57,62% e um desvio padrão de 5,08. O valor mínimo registrado para DeL foi de 45,98%, enquanto o máximo atingiu 73,37%. Em resumo, esses resultados indicam que as díades demonstraram um desempenho superior na diversidade lexical (72,09%) em comparação com a densidade lexical (57,62%).

Para finalizar a análise dos resultados da Tabela 2, nota-se que o valor da mediana, indicada pelo quartil de 50%, está próximo da média nas variáveis PE, DiL e DeL. Isso corrobora com a análise de que a média, nesses casos, é uma métrica com boa representatividade para o conjunto de textos analisados.

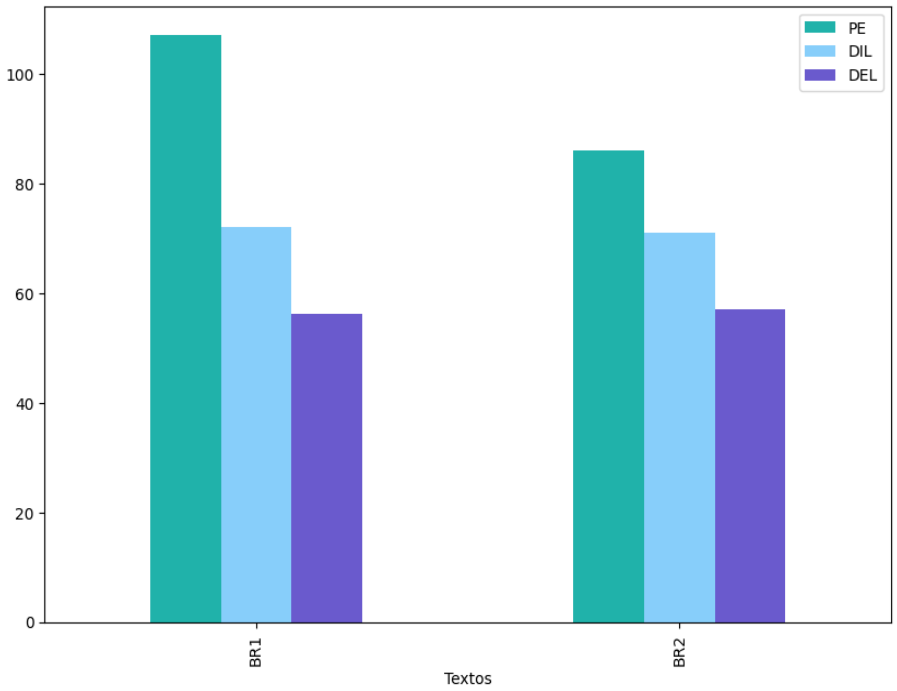
5.3 Resultados para cada grupo de díades

Para uma visualização dos resultados individuais das díades de ambos os países, fora elaborado um gráfico de setores para cada grupo. Isso permitiu uma melhor compreensão das médias de PE, DeL e DiL.

5.3.1 Resultados dos textos das díades brasileiras

Conforme demonstrado no Gráfico 3, temos a comparação dos resultados obtidos pelos dois grupos de díades brasileiras. Os textos das díades do primeiro ano (BR1) apresentaram uma média de aproximadamente 107 palavras, com uma diversidade lexical de 72,06% e uma densidade de 56,14%. Por outro lado, os textos das díades do segundo ano (BR2) foram escritos com uma média de aproximadamente 86,08 PE, com DiL de 71,12% e DeL de 56,97%.

Gráfico 3 - Resultados dos textos das díades brasileiras.



Fonte: A autora (2024).

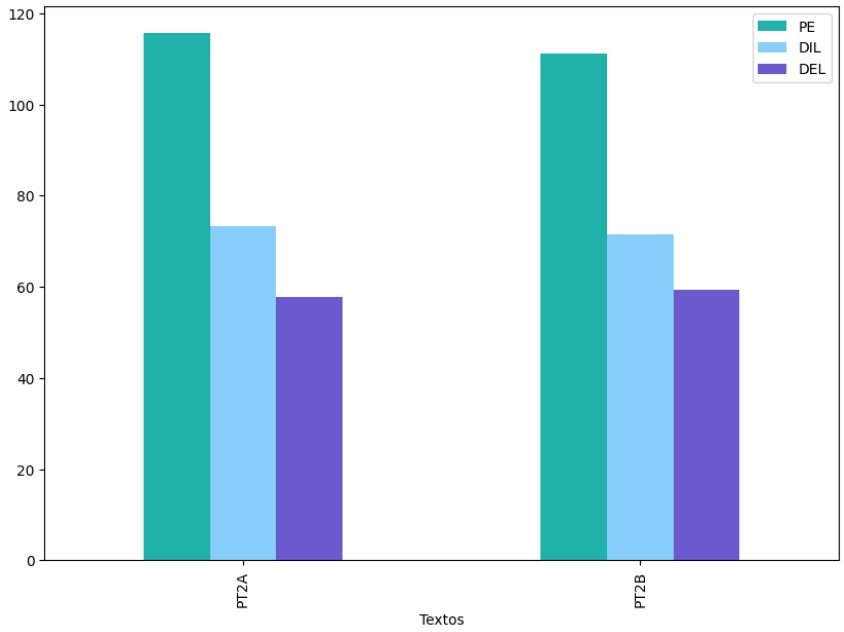
Nota-se, através do gráfico, que os textos produzidos pelas díades do primeiro ano, em média, foram mais extensos. No entanto, quando se trata da densidade e diversidade lexical, ambos os grupos apresentaram médias aproximadas. Essa proximidade sugere que, apesar da diferença no comprimento dos textos, a riqueza lexical se mostrou similar.

5.3.2 Resultados dos textos das díades portuguesas

As díades do grupo PT2A produziram seus textos com uma média de aproximadamente 115 palavras, tendo uma diversidade de 73,28% e densidade de 57,82%. Já os textos das díades do grupo PT2B foram escritos com uma média de aproximadamente 111 palavras, com diversidade lexical de 71,59% e densidade de 59,36%, conforme ilustrado o gráfico a seguir.

Os dois grupos apresentaram médias bastante próximas, tanto no número de palavras escritas, quanto na densidade e diversidade lexical. Esses resultados também sugerem uma similaridade na riqueza lexical entre os textos. É importante destacar que, embora os dois grupos de díades portuguesas sejam de duas escolas diferentes, ambos cursavam o mesmo ano escolar (segundo ano) e possuíam a mesma faixa etária.

Gráfico 4 - Resultados dos textos das díades portuguesas.



Fonte: A autora (2024).

5.3.3 Comparação dos textos das díades brasileiras e portuguesas

Fazendo uma análise comparativa inicial dos textos produzidos pelas díades de ambos os países, é possível observar que todos os grupos (BR1, BR2, PT2A, PT2B) obtiveram uma média de densidade acima de 50%. Ou seja, 50% das palavras dos textos, eram substantivos, adjetivos, verbos e advérbios. Esses resultados estão alinhados aos achados do estudo de Rodrigues (2008), que investigou a escrita de alunos portugueses do primeiro ao quarto ano e observou que a média de densidade dos seus textos esteve acima de 50% em todos os grupos escolares. Além disso, nossos resultados estão em consonância com a pesquisa de Ure (1971), que constatou que a densidade em textos escritos tende a ser superior a 40%.

Na análise da diversidade lexical dos textos dos quatro grupos de díades, foi identificada uma notável consistência das médias em torno de 71,12% a 73,28%. Isso demonstra que o vocabulário usado pelas díades é diversificado, pois, em média, apenas 30% das palavras escritas em seus textos foram repetidas. Esses resultados se aproximam dos achados de Martins (2016), que também observou médias de diversidade lexical acima de 70% nos textos de alunos portugueses do quinto, sétimo e décimo ano.

Em relação ao total de palavras escritas, dos quatro grupos de díades analisados, três apresentaram média superior a 100 palavras por texto: BR1 com 107 palavras, PT2B com 111 palavras e PT2A com 115 palavras. Esses resultados também demonstram uma similaridade

com a média de palavras por texto identificada no estudo de Rodrigues (2008) com alunos portugueses do segundo ano.

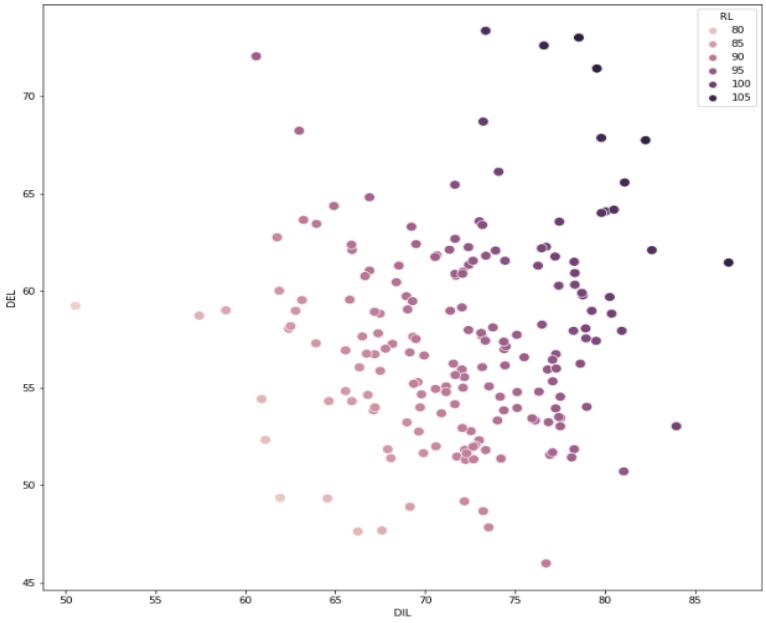
Apenas com os resultados obtidos desta análise comparativa entre as díades analisadas, não é possível responder a questão de pesquisa. Para determinar se existe diferença significativa ou não entre os textos das díades, foi necessário realizar uma análise mais detalhada, a qual será apresentada nos tópicos seguintes.

5.4 Resultado geral da medição da Riqueza Lexical (RL)

Os resultados referentes à extração da riqueza lexical dos textos podem ser visualizados no Gráfico de dispersão (Gráfico 5). Esse gráfico possui dois eixos principais: DeL e DiL. Cada ponto no gráfico, representado por diferentes cores, corresponde a um texto analisado no estudo, e a posição de cada ponto na intersecção dos eixos DeL e DiL indica a riqueza lexical desse texto.

Como pode ser observado na legenda do gráfico, a variação das cores nos pontos permite a identificação do valor da riqueza lexical dos textos, que pode variar de uma taxa de 80 para textos com menor riqueza lexical e uma taxa de 105 para textos com maior riqueza lexical.

Gráfico 5 - Resultado geral da riqueza lexical das díades.

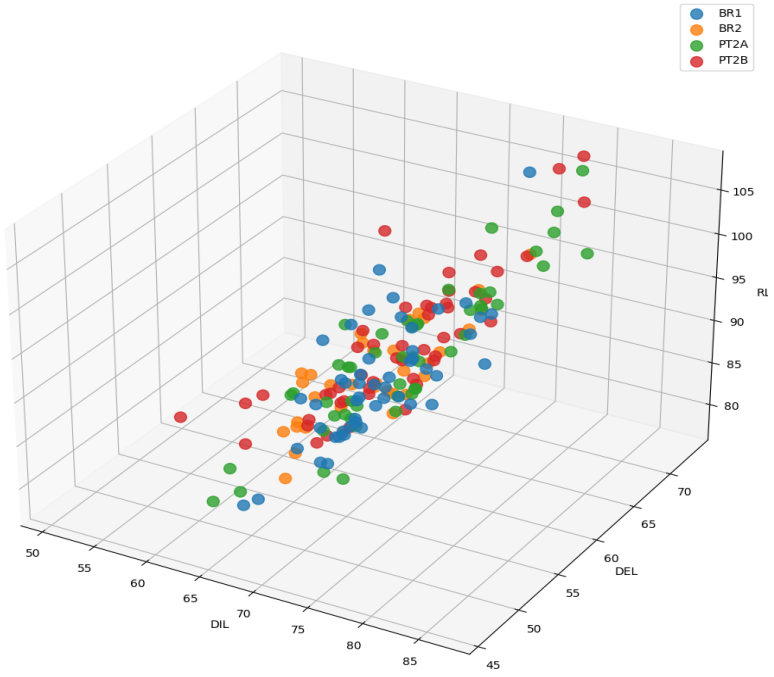


Fonte: A autora (2024).

Para ilustrar com mais clareza a riqueza lexical por grupo de díades, agora, visualiza-se no Gráfico 6, a seguir, o conjunto de textos representados em três eixos: DiL, DeL e riqueza

lexical. Tem-se, então, a posição de cada texto, diferenciando-os de acordo com seus respectivos grupos. Com isso, é possível observar que os textos com maior riqueza lexical foram produzidos pelas díades portuguesas (PT2B), bem como o texto com menor riqueza lexical (PT2A).

Gráfico 6 – Riqueza lexical por grupo de díades.



Fonte: A autora (2024).

Tanto no Gráfico 5 quanto no Gráfico 6, é evidente que os textos com maior diversidade e densidade lexical são os mais ricos. Porém, para entender de forma quantitativa a relação entre a diversidade, a densidade e riqueza lexical, aplicou-se o coeficiente de correlação de Pearson.

Nessa análise, encontrou-se que a correlação entre DiL e RL é de aproximadamente r=0,81, indicando uma forte correlação positiva. Já entre DeL e RL, a correlação foi de r=0,62, sugerindo uma correlação positiva moderada. Estes resultados confirmam que tanto a DiL quanto a DeL têm uma relação diretamente proporcional com a RL: quanto maior a DiL e a DeL de um texto, maior tende a ser sua riqueza lexical. Embora esse resultado seja esperado, já que a RL é calculada com base nesses dois indicadores, ele enfatiza a forte ligação entre eles.

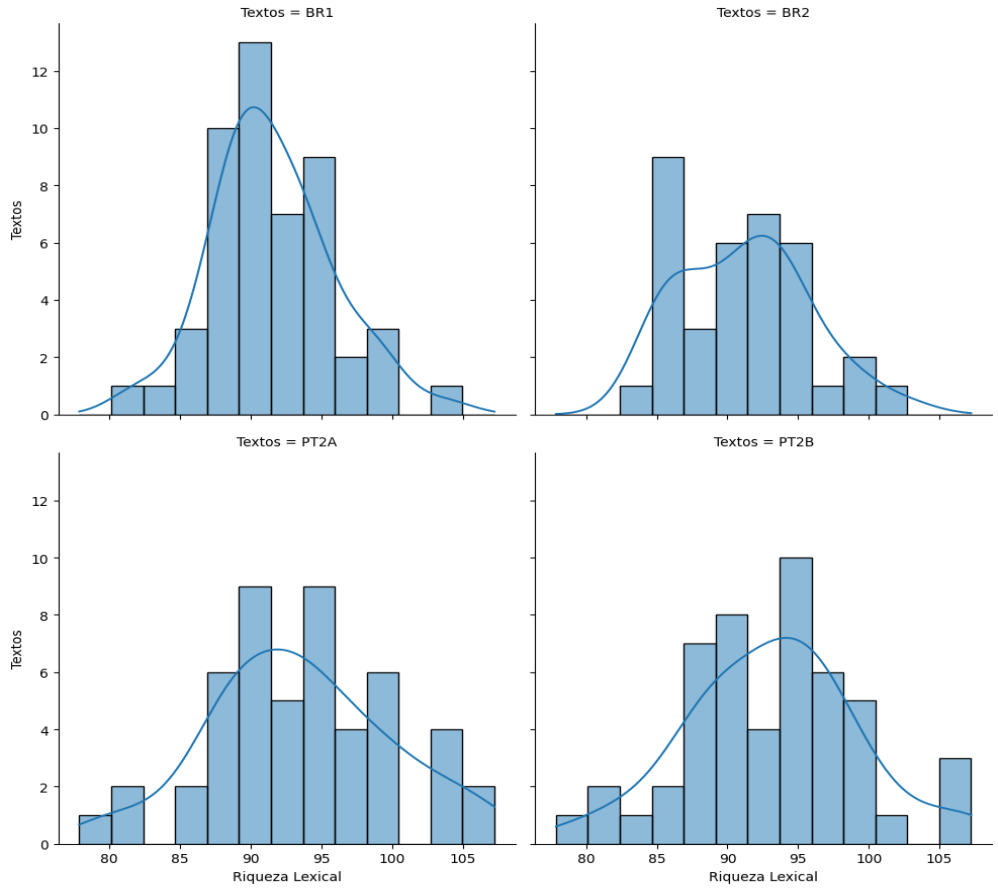
Com base nesses dados, foi possível estabelecer uma ordem para os textos utilizando a RL como critério unificado, o que permitiu compará-los de forma consistente e quantitativa.

5.5 Resultados da Riqueza lexical por grupo de díades

O Gráfico 7 traz o comportamento da RL nos textos de cada grupo de díades, por meio do histograma das frequências absolutas, dispostas nos gráficos de barras. Para cada grupo, computamos o estimador para a respectiva densidade de probabilidade, evidenciada pela linha suave que acompanha as barras. Essa linha permite uma visão mais contínua de como se comportariam as frequências em um número maior de díades, caso elas seguissem o mesmo padrão.

Em relação às médias de RL, o grupo BR1 apresentou uma média de 91,53, enquanto o grupo BR2 teve uma média de 91,23. Já para os grupos portugueses, PT2A e PT2B, as médias foram, respectivamente, 93,37 e 93,15.

Gráfico 7 - Histogramas de distribuição da Riqueza Lexical.



Fonte: A autora (2024).

A distribuição dos quatro grupos, pelo formato de sino, remete à distribuição gaussiana, sendo a distribuição de BR2 a mais duvidosa dessas. É interessante observar que todos os grupos apresentaram médias de RL próximas. Ao comparar essas médias, constata-se que os

textos produzidos pelas díades portuguesas apresentaram uma média geral (93,60) de RL maior que os textos produzidos pelas díades brasileiras (91,50). Entretanto, para assegurar quantitativamente essa característica, é necessário recorrer a uma das técnicas de estatística inferencial, que será discutida nos tópicos a seguir

5.6 Avaliação da normalidade das distribuições

Para identificar o tipo de distribuição que cada grupo de díades seguia, aplicou-se o teste de *Shapiro-Wilk*, e obtivendo os seguintes resultados:

Tabela 3 - Resultados do teste de normalidade das distribuições usando *Shapiro-Wilk.*

Turma	BR1	BR2	PT2A	PT2B
p-valor	0,4926	0,2101	0,4470	0,8138

Fonte: A autora (2024).

É possível verificar que todas as amostras apresentaram *p-valor* maior que 0,05. Isso significa que, para todos os testes, não foi rejeitada a hipótese nula de *Shapiro-Wilk*, ou seja, indicando que elas podem ser consideradas distribuições normais. Admitindo os valores computados de média e variância de cada um dos grupos, tem-se as seguintes distribuições:

- RL de BR1 é uma distribuição normal com média 91,53 e variância de 18,64;
- RL de BR2 é uma distribuição normal com média 91,23 e variância de 20,65;
- RL de PT2A é uma distribuição normal com média 93,47 e variância de 41,25;
- RL de PT2B é uma distribuição normal com média 93,15 e variância de 37,55.

5.7 Avaliação da similaridade das distribuições

Uma vez que as amostras seguiam uma distribuição normal, aplicou-se o teste ANOVA para avaliar se a mediana dos grupos relacionados apresentava similaridade estatística. Para isso, considerou-se três cenários de testes: i) comparação entre os grupos de díades de alunos brasileiros (BR1 e BR2); ii) comparação entre díades de alunos portugueses (PT2A e PT2B); iii) comparação entre as díades de alunos brasileiros e portugueses (BR1, BR2 e PT2A, PT2B). A Tabela 4 apresenta os p-valores resultantes dos respectivos testes.

Tabela 4 – Resultados do teste ANOVA[26].

Teste	BR1 e BR2	PT2A e PT2B	BR1, BR2 e PT2A, PT2B
p-valor	0,748	0,800	0,130

Fonte: A autora (2024).

Conforme expresso na tabela acima, os p-valores resultantes foram maiores que 0,05. Portanto, estes resultados confirmam a hipótese de que não há diferença estatisticamente significativa nos três cenários examinados, respondendo assim a questão de pesquisa.

Outro aspecto que merece destaque é que, apesar dos textos analisados neste estudo pertencerem a díades do primeiro e segundo ano do Ensino Fundamental, a proximidade nos resultados sugere que a diversidade lexical, a densidade lexical e a riqueza lexical dos textos não apresentaram diferenças em relação ao nível escolar. Esse mesmo padrão foi identificado por Johansson (2008) quando avaliou textos de alunos de diferentes idades e níveis escolares e observou que não houve diferença da densidade em suas produções.

Além disso, mesmo considerando que os textos foram produzidos a partir de uma proposta de escrita colaborativa, os resultados se aproximam das descobertas de Rodrigues (2008) e Martins (2016) em seus estudos sobre a diversidade lexical em produções individuais de alunos portugueses. Essa congruência sugere que, independentemente do formato da atividade de escrita - seja em díade ou individual -, a diversidade lexical em textos de alunos do mesmo ano escolar ou de anos escolares próximos tende a não apresentar diferenças significativas em relação à riqueza lexical

[26] Resultados com arredondamento de 4 casas decimais.

CONSIDERAÇÕES FINAIS

Este trabalho teve como propósito avaliar a riqueza lexical em produções textuais escritas por díades de alunos brasileiros e portugueses recém-alfabetizados, utilizando a ferramenta *Lexicanalytics web.* O *corpus* analisado, proveniente do projeto de cooperação internacional *InterWriting*, reuniu textos de alunos do 1º e 2º ano do Ensino Fundamental no Brasil e de alunos do 2º segundo ano do Ciclo do Ensino Básico de duas escolas de Portugal.

Ao longo do trabalho, apresentou-se a base teórica que fundamenta a riqueza lexical, bem como seus indicadores. Explorando as diferentes métricas voltadas para a mensuração da densidade e diversidade lexical, foi possível perceber os esforços contínuos da comunidade acadêmica para aperfeiçoá-las e garantir uma maior confiabilidade em seus resultados, principalmente na comparação entre textos com diferentes comprimentos.

Também foi possível compreender as diferentes aplicações da densidade e diversidade lexical na avaliação da produção textual de alunos de diferentes níveis de escolaridade, faixas etárias e nacionalidades. Contudo, ficou sublinhada a carência de ampliar os conhecimentos acerca da riqueza lexical no contexto da língua portuguesa, principalmente na escrita de alunos brasileiros, bem como a aplicação dos seus indicadores em diferentes contextos de produção.

Salienta-se que a escassez de ferramentas que realizam a extração da densidade, diversidade e riqueza lexical em textos em língua portuguesa é um dos fatores que podem contribuir para essa lacuna. Das sete ferramentas mapeadas e analisadas, apenas uma possui versão adaptada para a língua portuguesa, porém, não possui em suas funcionalidades a extração conjunta da densidade, diversidade e riqueza lexical.

Esse cenário motivou o desenvolvimento do *Lexicanalytics web,* um programa voltado exclusivamente para a medição da densidade, diversidade e riqueza lexical em textos de língua portuguesa. Com a capacidade de processar múltiplos textos simultaneamente e apresentar resultados comparativos entre eles, essa ferramenta destacou-se como um importante aliado para exiliar no processo de extração das informações lexicais dos textos.

Na avaliação dos resultados do *corpu*s da pesquisa, utilizando o *Lexicanalytics web,* fpo possível constatar que os textos produzidos pelas díades de alunos brasileiros do primeiro ano apresentavam uma média de cerca de 107 palavras, caracterizando-se por uma diversidade lexical de 72,06%, uma densidade lexical de 56,14% e riqueza lexical de 91,53. Já nos textos das díades do segundo ano, a média de palavras escritas foi ligeiramente menor,

aproximadamente de 86,08 palavras, com uma diversidade lexical de 71,12% e uma densidade lexical de 56,97%, alcançando uma riqueza lexical de 91,23.

Mesmo sendo esperado que essas medidas aumentassem com a progressão escolar, na comparação entre os dois grupos, apenas a diversidade lexical dos textos das díades do segundo ano demonstrou uma pequena elevação em sua média, enquanto as demais variáveis apresentaram uma pequena queda, sendo a de PE visivelmente maior. Apesar desse comportamento, a análise estatística por meio de teste de hipótese indicou que essas diferenças entre as produções das díades brasileiras do primeiro e segundo ano não são estatisticamente significativas, sugerindo um nível similar do conhecimento lexical em seus textos.

Na análise dos resultados das produções das díades portuguesas, revelou-se que seus textos foram escritos com uma média de 111 a 115 palavras, com densidade média de 57,82% a 59,36%, diversidade de 71,59% a 73,28%, alcançando uma riqueza de 93,15 a 93,47. A análise comparativa entre esses dois grupos de díades portuguesas mostrou algumas implicações importantes. Primeiramente, a proximidade entre as quatro métricas sugere que, apesar de estudarem em escolas diferentes, seus textos apresentaram domínio lexical similar. Logo, essas características lexicais na escrita dos alunos podem ser um reflexo da proposta pedagógica de ensino de língua materna do país.

Um aspecto de grande relevância a ser destacado é que mesmo os textos sendo produzidos em díade, conforme a metodologia adotada neste trabalho, os resultados se assemelharam aos achados de Rodrigues (2008) e Martins (2016) quando analisaram a diversidade lexical em produções individuais de alunos portugueses. Essa similaridade sugere que, independentemente do formato da atividade de escrita (seja em díade ou individual), a diversidade lexical em textos de alunos de mesmo ano escolar ou de anos escolares próximos tende a não apresentar diferenças significativas.

Por fim, na comparação global entre as produções das díades brasileiras e portuguesas, foi observado que, embora os textos das díades portuguesas tenham apresentado uma riqueza lexical ligeiramente superior, variando de 93,15 a 93,47, em comparação com a das díades brasileiras, que ficou entre 91,23 e 91,53, essa diferença não se mostrou estatisticamente significativa, como confirmou o teste de hipótese.

Assim, estes resultados respondem à questão de pesquisa, confirmando a hipótese de que não existe diferença significativa da riqueza lexical em textos produzidos pelas díades de alunos portugueses e brasileiros recém-alfabetizados. Apesar das pequenas variações observadas, os alunos de ambos os países demonstram um nível similar de domínio lexical em suas composições.

Contudo, é importante salientar que fatores como a proposta de ensino de língua materna dos países envolvidos e a metodologia de escrita aplicada neste trabalho podem ter tido um papel relevante nos resultados alcançados. Portanto, se reconhece que esses fatores merecem uma análise mais aprofundada em estudos futuros, o que permitirá compreender suas implicações pedagógicas no processo de ensino-aprendizagem da produção textual escrita.

Por fim, espera-se que este trabalho de doutoramento contribua para a expansão do conhecimento acerca da riqueza lexical em textos de língua portuguesa, revelando os aspectos lexicais da escrita de alunos brasileiros e portugueses recém-alfabetizados, bem como contribuir com novas perspectivas para a avaliação e o acompanhamento do conhecimento lexical dos alunos durante seu percurso escolar.

REFERÊNCIAS

ANDERSON, M. **Memory**. New York: Psychology Press, 2009.

ARNAUD, P. J. The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests. **Practice and Problems in Language Testing**, v. 7, p. 14-28, 1984.

BAX, S. **Text Inspector**. Online text analysis tool. 2012. Disponível em: https://textinspector.com/. Acesso em: 20 jul. 2024.

BRAGA, K. **Discurso reportado e a escritura colaborativa de histórias inventadas em sala de aula**. 2017. 230 f. Tese (Doutorado em Educação) – Centro de Educação, Programa de Pós-Graduação em Educação, Universidade Federal de Alagoas, Maceió, 2017.

BERMAN, R.; VERHOEVEN, L. Cross-Linguistic Perspectives on the Development of Text-Production Abilities in Speech and Writing. **Written Languages**, v. 5, p. 1-43, 2002.

CALIL, E.; DEL RÉ, A. Análise multimodal de uma história inventada: o caso da onomatopeia visual. **Revista da Anpoll**, Belo Horizonte, v. 2, n. 27, p. 13 -41, 2009.

CALIL, E.; FELIPETO, C. Rasuras orais semânticas na escritura a dois: a metaenunciação em histórias inventadas. **Intersecções**, Jundiaí, v. 2, p. 188-202, 2014.

CAMELO, R.; JUSTINO, S.; AND MELLO, R. Coh-metrix pt-br: Uma api web de análise textual para a educação. In: **Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação**, pages 179–186, Porto Alegre, RS, Brasil. SBC. XI Congresso Brasileiro de Informática na Educação (CBIE 2022) Anais do XXXIII Simpósio Brasileiro de Informática na Educação (SBIE 2022) 941, 2020.

CARROLL, J. B. Diversity of vocabulary and the harmonic series law of word frequency distribution. **The Psychological Record**, v. 2, p. 379-386, 1938.

COVINGTON, M.; MCFALL, J. Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR). **Journal of Quantitative Linguistics**, v. 17, n. 2, p. 94-100, 2010.

CRUZ, W. R. Linguística Computacional e suas subáreas. **Revista DisSoL - Discurso, Sociedade E Linguagem**, n. 2, 2015.

DAIUTE, C.; DALTON, B. Collaboration between children learning to write: con novices be masters. **Technical Report**, n. 60, 1992.

DALLER, H.; VAN HOUT, R.; TREFFERS-DALLER, J. Lexical Richness in the Spontaneous Speech of Bilinguals. **Applied Linguistics**, v. 24, n. 2, p. 197-222, junho de 2003. DOI: 10.1093/applin/24.2.197.

ELGOBSHAW, A; ALDAWSARI, M. Lexical Density as Improvement Indicator in the Written Performance of EFL Majors. **International Journal of English Language and Literature Studies**, v. 11, n. 4, p. 181-190, 2022.

ENGBER, C. The relationship of lexical proficiency to the quality of ESL compositions. **Journal of Second Language Writing**, v. 4, n. 2, p. 139-155, 1995.

FAYYAD, U. *et al.* **Advances in Knowledge Discovery and Data Mining**. American Association for Artificial Intelligence, 1996.

FILATRO, A. **Data Science na Educação:** presencial, a distância e corporativa. 1. ed. São Paulo: Saraiva Educação, 2021.

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

HALLIDAY, M. A. K. Spoken and written modes of meaning. *In:* HOROWITZ, R.; SAMUELS, S. J. (Org.). **Comprehending oral and written language**. Academic Press. Orlando, 1985.

HARMON, J., WOOD, K. D., MEDINA, A. Vocabulary learning in the content areas: Research-based practices for middle and secondary school classrooms. *In:* K. D. Wood.; W. E. Blanton (Eds.), **Literacy instruction for adolescents:** Research based practice. New York, p. 344–367, 2009.

HENRIKSEN, B. Three Dimensions of Vocabulary Development. **Studies in Second Language Acquisition (SSLA)**, v. 21, p. 303-317, 1999.

ISHIKAWA, S. Lexical Development in L2 English Learners' Speeches and Writings. **Procedia - Social and Behavioral Sciences**, 202–210, 2015. doi:10.1016/j.sbspro.2015.07.041

ISAACSON, S. Assessing the Writing Product: Qualitative and Quantitative Measures. **Exceptional Children**, v. 54, n. 6, pp. 528-534, 1988.

JARVIS, S. **Short texts, best fitting curves, and new measures of lexical diversity**. Language Testing, 2002.

JOHANSSON, V. **Lexical density and lexical variation in speech and writing:** a developmental perspective. Working Papers, Lund University, p. 61-79, 2009.

JOHNSON, Wendell. Studies in language behavior: A program of research. **Psychological Monographs**, v. 56, n. 2, p. 1-15, 1944.

LOPES, A. **A singularidade do erro ortográfico e os efeitos do funcionamento da língua**. Tese de doutoramento. Programa de Pós-graduação de Letras e Linguística. Universidade Federal de Alagoas, 2005.

KIM, J. Predicting L2 Writing Proficiency Using Linguistic Complexity Measures: A Corpus-Based Study. **English Teaching**, v. 69, n. 4, p. 27-51, 2014.

KOCH, I. G.; ELIAS, V. M. **Ler e escrever:** estratégias de produção textual. 2. Ed. São Paulo: Contexto, 2010.

KONDAL, B. Effects of Lexical Density and Lexical Variety in Language Performance and Proficiency. International Journal of IT, **Engineering and Applied Sciences Research**, v. 4, n.10, p. 2319-441, 2015.

KYLE, K.; CROSSLEY, S. A.; JARVIS, S. (2021). Assessing the validity of lexical diversity using direct judgements. **Language Assessment Quarterly**, v. 18, n. 2, p. 154-170, 2021. Disponível em: https://doi.org/10.1080/15434303.2020.1844205. Acesso em: 20 jul. 2024.

LEFFA, V. Aspectos externos e internos da aquisição lexical. *In:* LEFFA, Vilson (Org). **As palavras e sua companhia:** o léxico na aprendizagem das línguas. Pelotas: EDUCAT, p. 15-44, 2000.

LINNARUD, M. **Lexical density and lexical variation** – An analysis of the lexical texture of Swedish students' written work. University of Lund, 1973.

LIRA, L. **A construção do discurso reportado em processos de escritura de manuscritos escolares por duas díades de alunos do 2º ano do Ensino Fundamental – fronteiras entre o oral e o escrito**. 2017. 177 f. Tese (Doutorado em Educação) – Centro de Educação, Programa de Pós-Graduação em Educação, Universidade Federal de Alagoas, Maceió, 2017.

LU, X. The relationship of lexical richness to the quality of ESL learners' oral narratives. **The Modern Language Journal**, v. 96, n. 2, p. 190-208, 2012.

MACWHINNEY, B.; SNOW, C. The Child Language Data Exchange System: An update. **Journal of Child Language**, v. 17, n. 2, p. 457– 472, 1990.

MACWHINNEY, B. **The Childes Project:** Tools for Analyzing Talk. 3ª Edition. Lawrence Erlbaum Associates: Mahwah, New Jersey, 2000.

MAGALHÃES, M. **Noções de Probabilidade e Estatística.** 7 ed. São Paulo: Editora da Universidade de São Paulo, 2010.

MARTINS, M. A Diversidade Lexical na escrita de textos escolares. **Fórum Linguístico**, Florianópolis, v.13, n. 1, p. 1068-1082, 2016. DOI:10.5007/1984-8412.2016v13n1p1068. Acesso em: 20 jul. 2024.

MARZANO, R.; PICKERING, D. **Building academic vocabulary:** Teacher's manual. Alexandria, VA: Association for Supervision and Curriculum Development, 2005.

MCCARTHY, M., JARVIS, S. VOCD: A theoretical and empirical evaluation. **Language Testing**. University of Memphis, USA, and Ohio University, USA. v. 24, p. 459-488, 2007.

MCCARTHY, M., JARVIS, S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. **Behavior Research Method,** v. 42 (2), p. 381-392, 2010.

MCNAMARA, D. S. *et al*. **Automated evaluation of text and discourse with Coh-Metrix**. Cambridge University Press, 2014.

MALVERN, D. *et al.* **Lexical Diversity and Language Development:** Quantification and Assessment. Basingstoke, Hampshire: Palgrave Macmillan, 2004.

MENARD, N. **Mesure de la richesse lexicale**. Geneva: Slatkine, 1983.

MORETTIN, A., BUSSAB, W. **Estatística Básica.** Ed. 9. São Paulo: Saraiva, 2017.

NAGY, W. E.; SCOTT, J. A. Vocabulary processes. *In:* M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.). **Handbook of reading research**, vol. 3, p. 269–284, 2000.

NATION, I. S. P. **Teaching and Learning Vocabulary**. Newbury House, 1999.

OLINGHOUSE, N. G.; LEAIRD, J. T. The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. **Reading and Writing**, v. 22, n. 5, p. 545–565, 2008. doi:10.1007/s11145-008-9124-z. Acesso em: 20 jul. 2024.

PALMER, D. Text Preprocessing. **Handbook of natural language processing**, v.2, p. 9-30, 2010.

QUEIROZ, J. **Rasuras escritas comentadas e sua importância para a compreensão da revisão em ato:** um estudo de caso de uma díade recém-alfabetizada. 2019. 202 f. Tese (Doutorado em Linguística) – Faculdade de Letras, Programa de Pós-graduação em Linguística e Literatura, Universidade Federal de Alagoas, Maceió, 2019.

OLINGHOUSE, N. G.; WILSON, J. The relationship between vocabulary and writing quality in three genres. **Reading and Writing**, v. 26, n. 1, p.45 - 65, 2012.

RAIMES, A. What unskilled ESL students do as they write: A classroom study of composing. **TESOL Quarterly**, v. 19, n. 2, 229-258, 1985.

REHDER, B. *et al*. **Using latent semantic analysis to assess knowledge:** Some technical considerations. Discourse Processes, 1998.

READ, J. **Assessing vocabulary.** Cambridge University Press, p.188-210, 2000.

RIFFO, K.; OSUNA, S.; LAGOS, P. S. Lexical Diversity and Lexical Density Description of News Written by Journalism Student. **Rev. Bras. Linguíst. Apl**., v. 19, n. 3, p. 499- 528, 2019.

RODRIGUES, S. **Escrita espontânea:** Desenvolvimento das capacidades de composição escrita de em crianças do 1º ao 4º ano de escolaridade. (Dissertação de Mestrado). Universidade Fernando Pessoa-Porto, p.107, 2008.

SADEGHI, K.; DILMAGHANI, S. The Relationship between Lexical Diversity and Genre in Iranian EFL Learners Writings. **Journal of Language Teaching and Research**, v. 4, n. 2, p. 328-334, 2013. doi:10.4304/jltr.4.2.328-334, 2013. Acesso em: 20 jul. 2024.

SANTOS, E. S.; LEITE, G. R.; CALIL, E. Lexicanalytics: uma ferramenta para auxiliar na análise lexical em produções textuais. *In:* **Anais do Congresso Acadêmico Integrado de Inovação e Tecnologia (CAIITE)**, v1, n1, p. 1468-1469, 2017.

SANTOS, E. **Manuscritos escolares de alunos recém-alfabetizados:** um estudo sobre a densidade e diversidade lexical. 2020. 116 f. Dissertação (Mestrado em Educação) – Centro de Educação, Programa de Pós-graduação em Educação, Universidade Federal de Alagoas, Maceió, 2020.

SANTOS, E.; CARVALHO FILHO, F.; LEITE, G. R.; CALIL, E., Lexicanalytics web: uma ferramenta de análise lexical para auxiliar no processo de avaliação da escrita escolar. *In:* **Cultura Escolar em Tempos de Pandemia**. Campina Grande: Realize Editora, 2022. Disponível em: https://www.editorarealize.com.br/artigo/visualizar/92065. Acesso em: 11 jan. 2024.

SANTOS, J. **Marcas de pontuação em manuscritos escolares: estudo sobre os usos de vírgulas por alunos de uma turma do 9º ano do ensino fundamental**. 2023. 111 f. Dissertação (Mestrado em Letras e Linguística: Linguística) - Faculdade de Letras, Programa de Pós-Graduação em Letras e Linguística, Universidade Federal de Alagoas, Maceió, 2023.

SCHERER, S.; SOUZA, A. S. Types e tokens na aquisição típica de linguagem por sujeitos de 18 a 32 meses falantes do português brasileiro. **Revista CEFAC**, v. 12, n. 5, p. 838- 845, 2011.

SILVA, M. **Rasuras orais:** estudo comparativo dos processos de escritura de duas díades de alunas recém-alfabetizadas. 2014.108 f. Dissertação (Mestrado em Educação Brasileira) - Centro de Educação, Programa de Pós-graduação em Educação, Universidade Federal de Alagoas, Maceió, 2019.

SOMMERVILLE, I. **Engenharia de Software**. 9. ed. Pearson Ed Superior, 2011.

SKEHAN, P. **Lexical Performance by Native and Non-Native Speakers on Language-Learning Tasks**. Vocabulary Studies in First and Second Language Acquisition, 2009, p. 107-124.

STROMQVIST, V. *et al*. Toward a cross-linguistic comparison of lexical quanta in speech and writing. **Written Language and Literacy**, v. 5, p. 45-67, 2002.

URE, J. lexical density and register differentiation. *In:* **Applications of linguistics**. selected papers of the second international congress of applied linguistics. Cambridge, p, 443-452, 1971.

UZAWA, K.; CUMMINGS, A. Writing strategies in Japanese as a foreign language: Lowering or keeping up the standards. **The Canadian Modern Language Review**, v. 46, p.179–191, 1989.

YU, G. Lexical Diversity in Writing and Speaking Task Performances. **Applied Linguistics**, v. 31, n. 2, p. 236–259, 2009.

ZIPF, G. K. **The Psychobiology of Language**. Houghton Mifflin, 1935.

WANG, X. The relationship between lexical diversity and EFL writing proficiency. **University of Sydney Papers in TESOL**, v. 9, p. 65-88, 2014.

ANEXOS

ANEXO 1 – Dados brutos extraídos dos textos analisados

Textos BR1

Textos	PE	DiL	DeL
1	108	78,3	51,85
2	151	64,9	64,36
3	211	79	54,03
4	118	73,6	55,08
5	120	74	53,33
6	59	69,3	57,63
7	113	72,7	51,33
8	86	67,6	47,67
9	62	69	53,23
10	83	69	59,04
11	135	76,1	53,33
12	68	80,4	58,82
13	51	61,8	62,75
14	103	77,1	55,34
15	81	67,9	51,85
16	125	69,5	62,4
17	89	73,4	61,8
18	130	71,7	60,77
19	93	64	63,44
20	78	79,3	58,97
21	62	76,3	61,29
22	72	72,6	52,78
23	89	63,9	57,3
24	102	72	55,88
25	54	66,9	64,81
26	78	74,4	53,85
27	130	73	52,31
28	84	66,3	47,62
29	120	70	56,67
30	113	73,2	48,67
31	118	71,2	55,08

32	148	73,8	58,11
33	64	67,4	57,81
34	110	68,2	57,27
35	110	72,2	51,82
36	120	72,2	49,17
37	73	71,2	54,79
38	146	74,2	51,37
39	107	73,2	56,07
40	100	74,4	57
41	115	72,3	51,3
42	175	73,4	73,37
43	73	72,8	52,05
44	211	81,1	50,71
45	84	74,5	57,14
46	155	79,5	57,42
47	136	71,8	51,47
48	132	84	53,03
49	66	66,4	56,06
50	143	63	68,23

Textos BR2

Texto	PE	DIL	DEL
1	128	77	51,56
2	93	78,9	58,06
3	78	71,4	58,97
4	124	65,9	62,1
5	77	66,9	61,04
6	72	65,6	56,94
7	78	62,8	58,97
8	80	61,9	60
9	84	63,2	59,52
10	72	68,1	51,39
11	81	65,9	54,32
12	82	78,8	59,76
13	55	70,7	61,82
14	53	76,7	62,26
15	84	76,8	55,95
16	81	72,2	55,56
17	77	74,2	54,55
18	63	75,1	53,97
19	54	71,4	62,11
20	52	73,2	57,69
21	122	72,3	51,64
22	75	72,4	61,33
23	84	72,1	55,95

24	62	65,6	54,84
25	141	77,3	56,74
26	73	74,5	56,16
27	103	80	64,08
28	75	69,8	54,67
29	72	69	59,72
30	139	76,9	53,24
31	85	66,5	57,65
32	90	69,2	48,89
33	91	67,1	53,85
34	81	64,7	54,32
35	183	69,2	56,83
36	55	77,5	54,55

Textos PT2A

Texto	PE	DiL	DeL
1	153	82,6	62,09
2	131	76	53,44
3	129	73	63,57
4	127	69,7	52,76
5	184	72,1	60,87
6	115	70,6	61,74
7	119	72,1	52,94
8	73	64,6	49,32
9	119	79,6	71,43
10	168	79,8	67,86
11	136	67,5	55,88
12	111	70,6	54,95
13	96	78,6	56,25
14	110	63,2	63,64
15	99	81,1	65,57
16	69	73,5	47,83
17	143	62,4	58,04
18	166	86,9	61,45
19	108	70,9	53,7
20	106	79	57,55
21	73	75,1	54,79
22	131	65,8	59,54
23	61	74,4	57,38
24	114	77,4	53,51
25	178	77,1	51,69
26	78	77,4	60,26
27	131	73,2	68,7
28	100	69,7	54
29	101	77,5	53,47
30	85	67,5	58,82

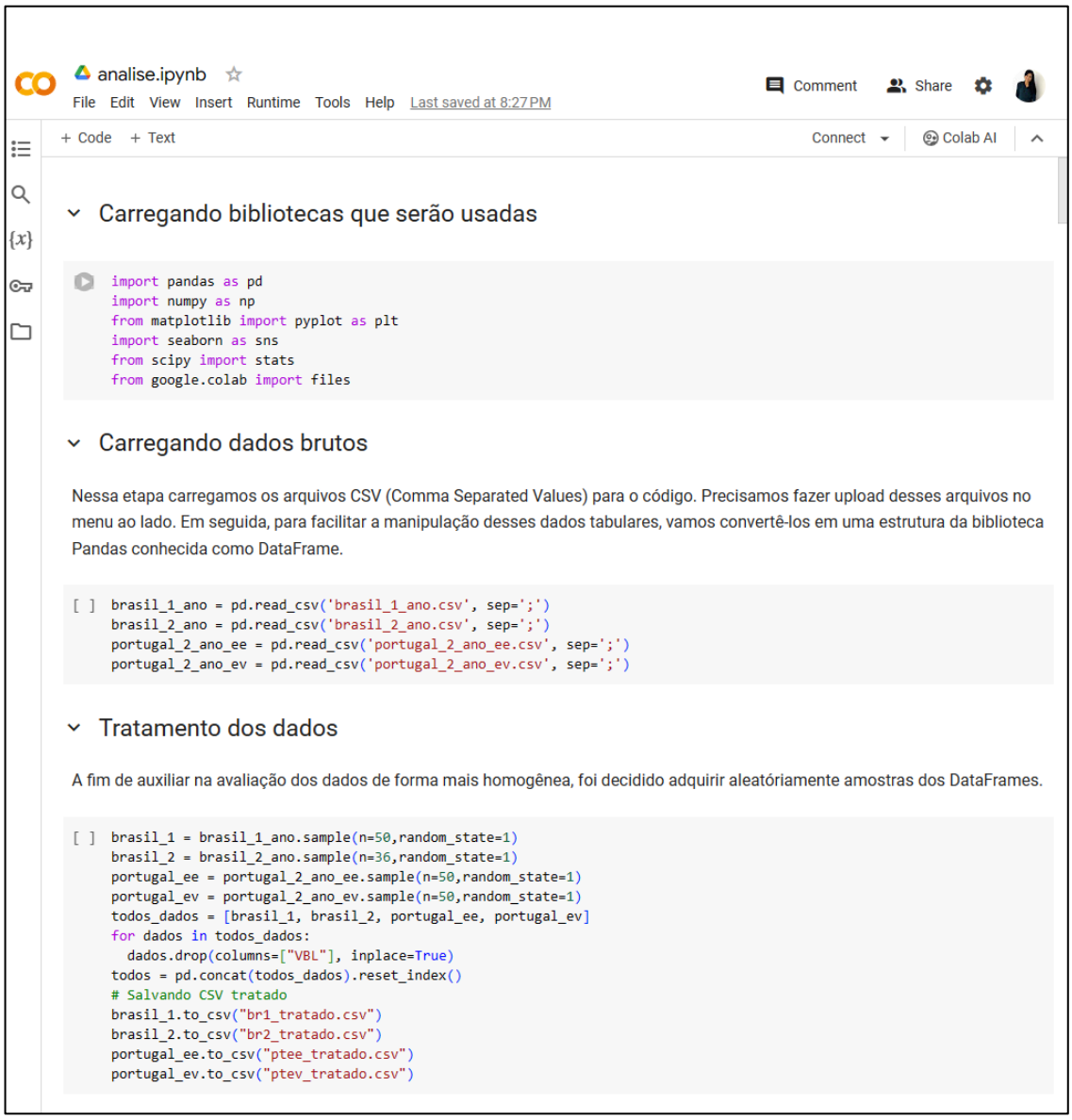
31	131	78,3	60,31
32	148	78,3	61,49
33	72	71,7	54,17
34	182	78,7	59,89
35	68	80,5	64,17
36	154	61,9	49,35
37	110	62,5	58,18
38	83	73,4	51,81
39	87	76,7	45,98
40	123	72,1	60,98
41	93	68,5	61,29
42	186	80,3	59,68
43	185	67,2	58,92
44	102	77,2	61,76
45	83	73,1	57,83
46	74	69,3	59,46
47	161	71,7	60,87
48	66	77,5	53,03
49	76	75,5	56,58
50	86	61,1	52,33

Textos PT2B

Texto	PE	DiL	DeL
1	121	74,1	66,12
2	127	72,1	55,01
3	74	66,8	56,76
4	79	60,9	54,43
5	93	72,1	59,14
6	121	67,8	57,02
7	91	69,9	51,65
8	87	73,9	62,07
9	103	50,6	59,22
10	75	70,6	52
11	252	72,7	51,98
12	126	78,6	73,02
13	101	73,4	57,43
14	135	76,3	54,81
15	107	78,3	57,94
16	96	71,6	56,25
17	100	77,3	56
18	113	69,5	57,52
19	126	80,9	57,94
20	132	69,6	55,3
21	134	69,4	55,22
22	80	76,6	72,61
23	97	66,8	54,64
24	93	82,3	67,74
25	142	73,2	63,38
26	55	71,7	65,45
27	100	58,9	59
28	104	67,2	56,73
29	75	71,7	62,67
30	106	71,7	55,66
31	62	77,1	56,45

32	119	76,5	62,18
33	152	77,3	53,95
34	119	72,4	57,98
35	107	66,7	60,75
36	117	72,7	61,54
37	156	74,5	61,54
38	105	78,2	51,43
39	97	75,1	57,73
40	110	78,3	60,91
41	91	68,4	60,44
42	79	69,2	63,29
43	150	79,8	64
44	136	60,6	72,06
45	109	57,4	58,72
46	196	72,4	62,24
47	93	65,9	62,37
48	100	67,2	54
49	115	76,5	58,26
50	107	77,5	63,55

ANEXO 2 – Ambiente de tratamento e inferência dos dados (Colab)


analise.ipynb
File Edit View Insert Runtime Tools Help Last saved at 8:27 PM
Comment
Share
+ Code + Text
Connect
Colab AI
Carregando bibliotecas que serão usadas
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from scipy import stats
from google.colab import files
Carregando dados brutos
Nessa etapa carregamos os arquivos CSV (Comma Separated Values) para o código. Precisamos fazer upload desses arquivos no menu ao lado. Em seguida, para facilitar a manipulação desses dados tabulares, vamos convertê-los em uma estrutura da biblioteca Pandas conhecida como DataFrame.
[ ] brasil_1_ano = pd.read_csv('brasil_1_ano.csv', sep=';')
brasil_2_ano = pd.read_csv('brasil_2_ano.csv', sep=';')
portugal_2_ano_ee = pd.read_csv('portugal_2_ano_ee.csv', sep=';')
portugal_2_ano_ev = pd.read_csv('portugal_2_ano_ev.csv', sep=';')
Tratamento dos dados
A fim de auxiliar na avaliação dos dados de forma mais homogênea, foi decidido adquirir aleatóriamente amostras dos DataFrames.
[ ] brasil_1 = brasil_1_ano.sample(n=50,random_state=1)
brasil_2 = brasil_2_ano.sample(n=36,random_state=1)
portugal_ee = portugal_2_ano_ee.sample(n=50,random_state=1)
portugal_ev = portugal_2_ano_ev.sample(n=50,random_state=1)
todos_dados = [brasil_1, brasil_2, portugal_ee, portugal_ev]
for dados in todos_dados:
dados.drop(columns=["VBL"], inplace=True)
todos = pd.concat(todos_dados).reset_index()
# Salvando CSV tratado
brasil_1.to_csv("br1_tratado.csv")
brasil_2.to_csv("br2_tratado.csv")
portugal_ee.to_csv("ptee_tratado.csv")
portugal_ev.to_csv("ptev_tratado.csv")

analise.ipynb

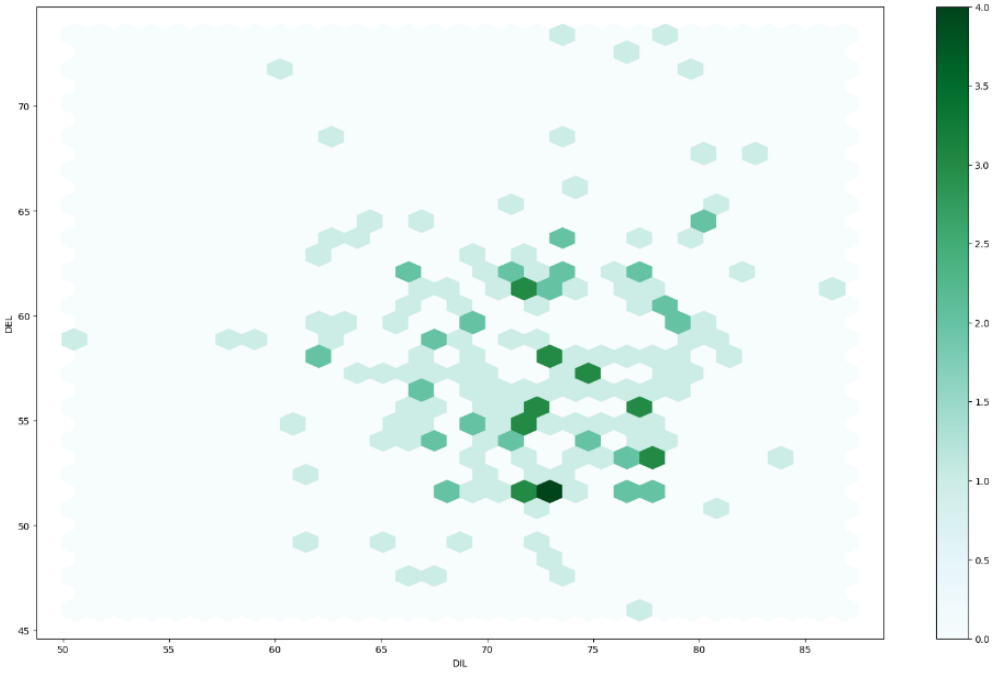
File Edit View Insert Runtime Tools Help Last saved at 8:27 PM

Comment Share

+ Code + Text Connect Colab AI

Descrição gráfica dos dados

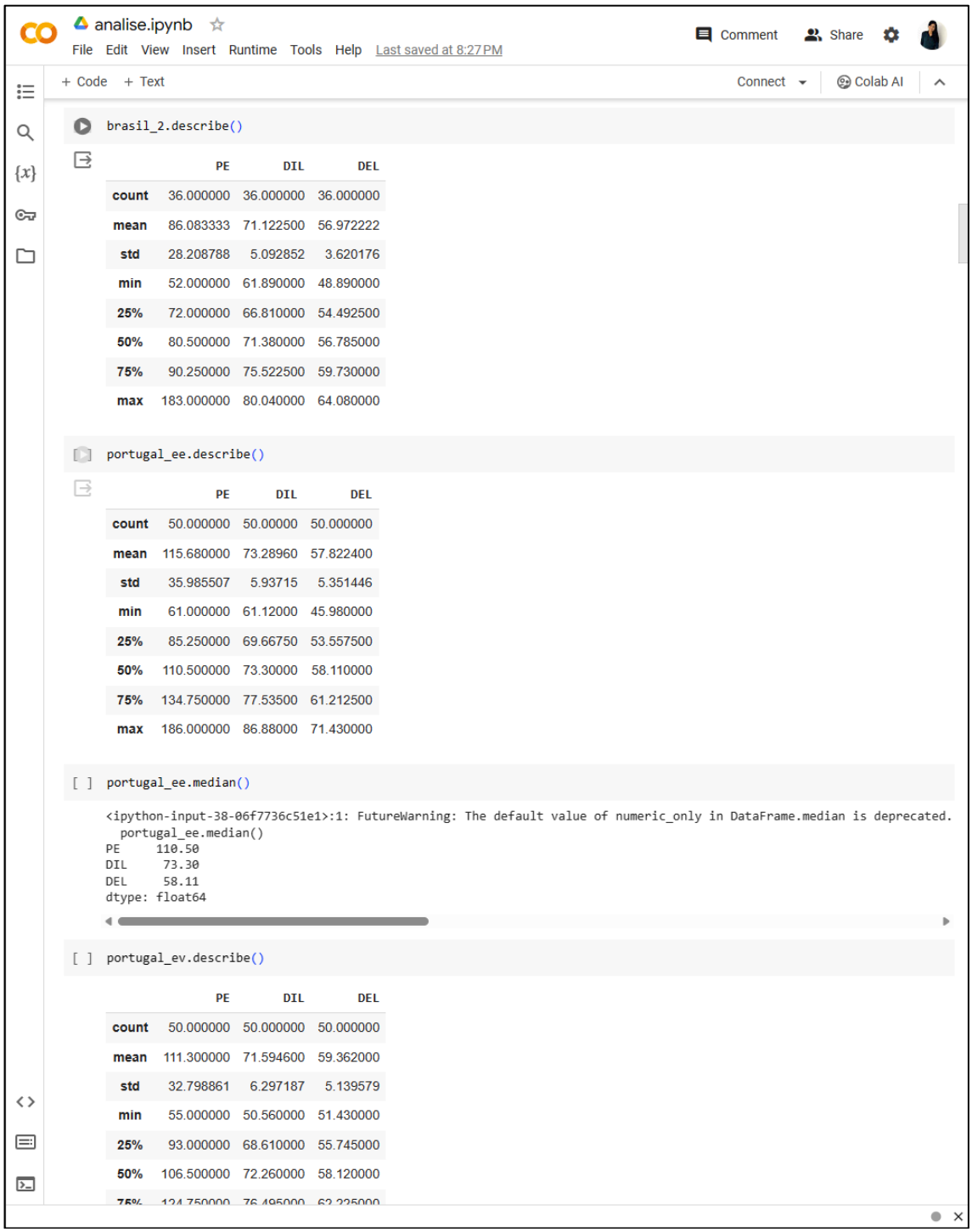
```
todos.plot.hexbin(x='DIL', y='DEL', gridsize=30, figsize=(20,12), sharex=False)
plt.show()
```



Mostrando estatística descritiva

```
[ ] brasil_1.describe()
```

	PE	DIL	DEL
count	50.000000	50.000000	50.000000
mean	107.080000	72.069400	56.149400
std	36.640471	4.866539	5.261735
min	51.000000	61.770000	47.620000
25%	78.750000	69.000000	52.115000
50%	107.500000	72.405000	55.610000
75%	128.750000	74.340000	58.642500
max	211.000000	83.970000	73.370000



	PE	DIL	DEL
count	36.000000	36.000000	36.000000
mean	86.083333	71.122500	56.972222
std	28.208788	5.092852	3.620176
min	52.000000	61.890000	48.890000
25%	72.000000	66.810000	54.492500
50%	80.500000	71.380000	56.785000
75%	90.250000	75.522500	59.730000
max	183.000000	80.040000	64.080000

	PE	DIL	DEL
count	50.000000	50.00000	50.000000
mean	115.680000	73.28960	57.822400
std	35.985507	5.93715	5.351446
min	61.000000	61.12000	45.980000
25%	85.250000	69.66750	53.557500
50%	110.500000	73.30000	58.110000
75%	134.750000	77.53500	61.212500
max	186.000000	86.88000	71.430000

	PE	DIL	DEL
count	50.000000	50.000000	50.000000
mean	111.300000	71.594600	59.362000
std	32.798861	6.297187	5.139579
min	55.000000	50.560000	51.430000
25%	93.000000	68.610000	55.745000
50%	106.500000	72.260000	58.120000
75%	124.750000	76.495000	62.225000

analise.ipynb

File Edit View Insert Runtime Tools Help Last saved at 8:27 PM

Comment Share

+ Code + Text Connect Colab AI

```
portugal_ev.describe()
```

	PE	DIL	DEL
count	50.000000	50.000000	50.000000
mean	111.300000	71.594600	59.362000
std	32.798861	6.297187	5.139579
min	55.000000	50.560000	51.430000
25%	93.000000	68.610000	55.745000
50%	106.500000	72.260000	58.120000
75%	124.750000	76.495000	62.225000
max	252.000000	82.260000	73.020000

```
[ ] todos.describe()
```

	index	PE	DIL	DEL
count	186.000000	186.000000	186.000000	186.000000
mean	25.645161	106.462366	72.086505	57.621989
std	15.527235	35.274080	5.626447	5.088370
min	0.000000	51.000000	50.560000	45.980000
25%	13.000000	79.250000	68.435000	54.000000
50%	25.000000	102.000000	72.285000	57.340000
75%	38.000000	126.750000	76.700000	60.962500
max	58.000000	252.000000	86.880000	73.370000

```
[ ] plt.figure(figsize=(20,12))
    ax1 = plt.subplot(2,2,1)
    brasil_1.boxplot()
    ax1.set_ylim((40, 260))
    ax1.title.set_text('BR1')
    ax2 = plt.subplot(2,2,2)
    brasil_2.boxplot()
    ax2.set_ylim((40, 260))
    ax2.title.set_text('BR2')
    ax3 = plt.subplot(2,2,3)
    portugal_ee.boxplot()
    ax3.set_ylim((40, 260))
    ax3.title.set_text('PT2A')
    ax4 = plt.subplot(2,2,4)
    portugal_ev.boxplot()
    ax4.set_ylim((40, 260))
    ax4.title.set_text('PT2B')
    plt.savefig("boxplots.png")
    plt.show()
```

analise.ipynb

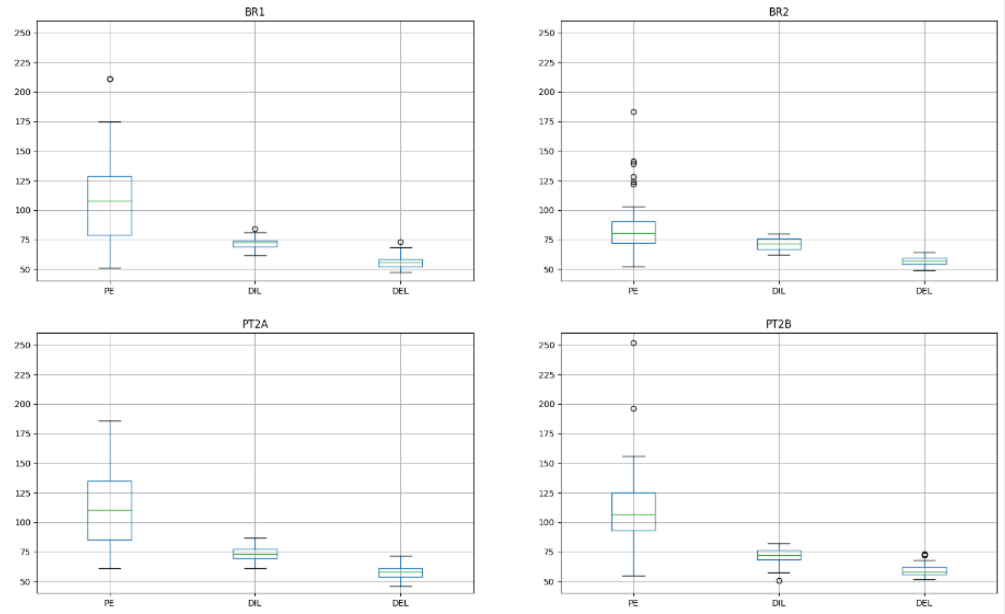
File Edit View Insert Runtime Tools Help Saving...

Comment Share

+ Code + Text

Connect Colab AI

```
plt.savefig('boxplots.png')
plt.show()
```



Aplicando labels (rótulos)

Uma forma de analisar todos os dados como um só, mas mantendo a informação de qual dado pertence a qual escola/ano é através da rotulação.

```
[ ] label = ['BR1', 'BR2', 'PT2A', 'PT2B']
    true_labels = np.concatenate([np.repeat(0, 50), np.repeat(1, 36), np.repeat(2, 50), np.repeat(3, 50)])
    true_labels_names = [label[i] for i in true_labels]
    todos = todos.assign(label=true_labels)
    todos = todos.assign(Textos=true_labels_names)
    print(todos)
```

```
     index Producao   PE    DIL    DEL  label Textos
0       22   005_D2   72  72.55  52.78      0    BR1
1        2   001_D4  151  64.93  64.36      0    BR1
2       49   010_D4   66  66.35  56.06      0    BR1
3       26   005_D6   78  74.38  53.85      0    BR1
4       33   007_D2   64  67.39  57.81      0    BR1
..     ...      ...  ...    ...    ...    ...    ...
181     45   005_D7  156  74.45  61.54      3   PT2B
182      7   001_D8  121  67.80  57.02      3   PT2B
183     56   006_D8  100  67.21  54.00      3   PT2B
184      1   001_D2  121  74.09  66.12      3   PT2B
185     16   002_D8   75  70.60  52.00      3   PT2B

[186 rows x 7 columns]
```

analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

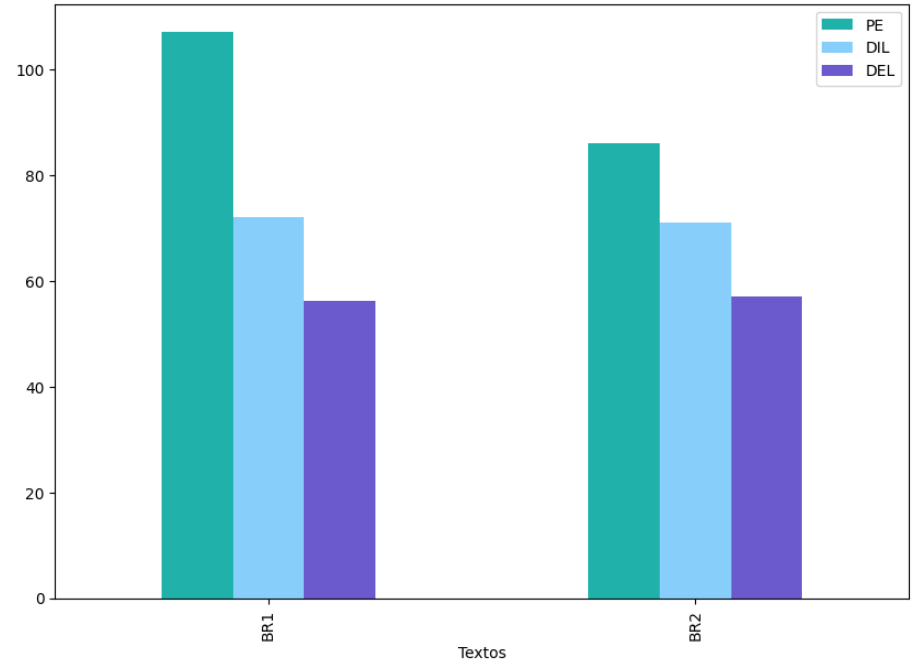
Comment Share

+ Code + Text Connect Colab AI

Comparação das médias das produções brasileiras

```
centrais = todos.groupby('Textos', as_index=False).mean()
brasil = centrais.loc[np.logical_or(centrais['Textos']=='BR1',centrais['Textos']=='BR2')]
brasil.plot.bar(x='Textos', y=['PE','DIL','DEL'], color=['lightseagreen', 'lightskyblue', 'slateblue'], figsize=(10,7
plt.savefig("bar_mean_br.png")
plt.show()
```

```
<ipython-input-43-a9297343eb35>:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.mean is deprec
  centrais = todos.groupby('Textos', as_index=False).mean()
```



analise.ipynb

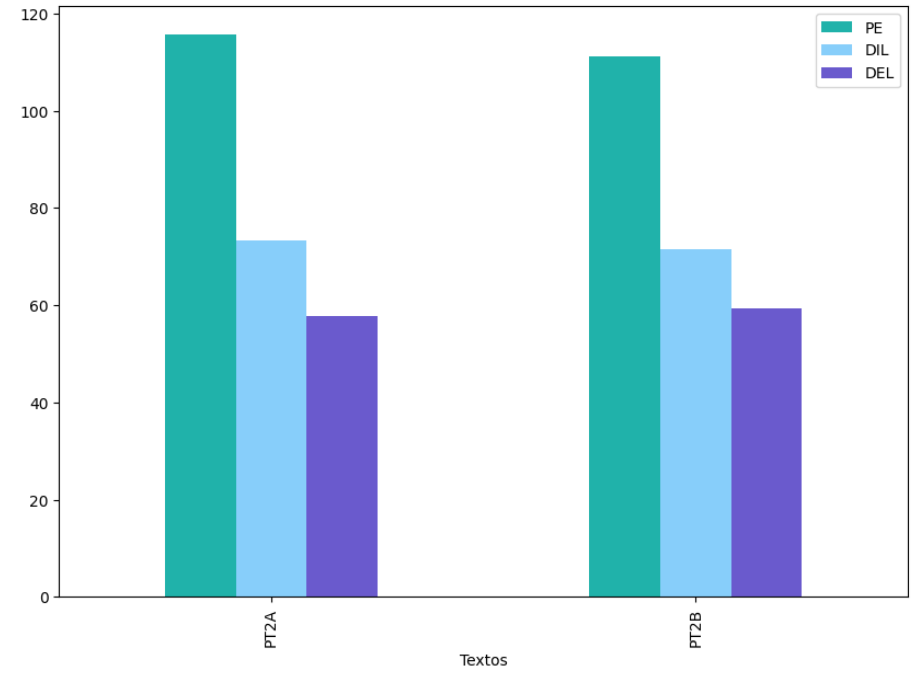
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

Comparação das médias das produções portuguesas

```
portugal = centrais.loc[np.logical_or(centrais['Textos']=='PT2A',centrais['Textos']=='PT2B')]
portugal.plot.bar(x='Textos', y=['PE','DIL','DEL'], color=['lightseagreen', 'lightskyblue', 'slateblue'], figsize=(10
plt.savefig("bar_mean_pt.png")
plt.show()
```



analise.ipynb

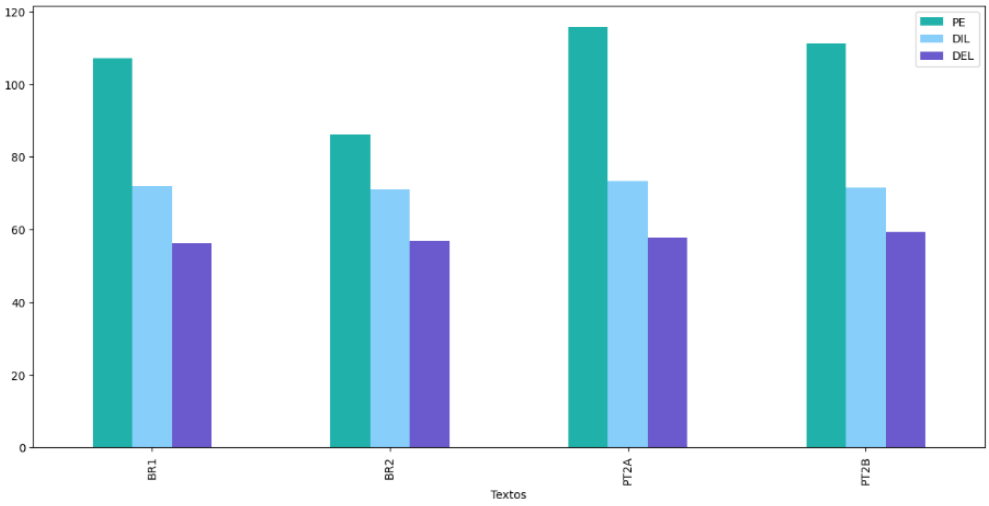
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

Por fim, a comparação entre Brasil e Portugal

```
centrais.plot.bar(x='Textos', y=['PE','DIL','DEL'], color=['lightseagreen', 'lightskyblue', 'slateblue'], figsize=(15
plt.savefig("bar_mean_geral.png")
plt.show()
```



Abaixo está uma plot de dispersão mostrando que as distribuições de DIL e DEL nos eixos e a concentração das produções. Nota-se que não há linha divisória que consiga categorizar os dados em classes separadas.

analise.ipynb

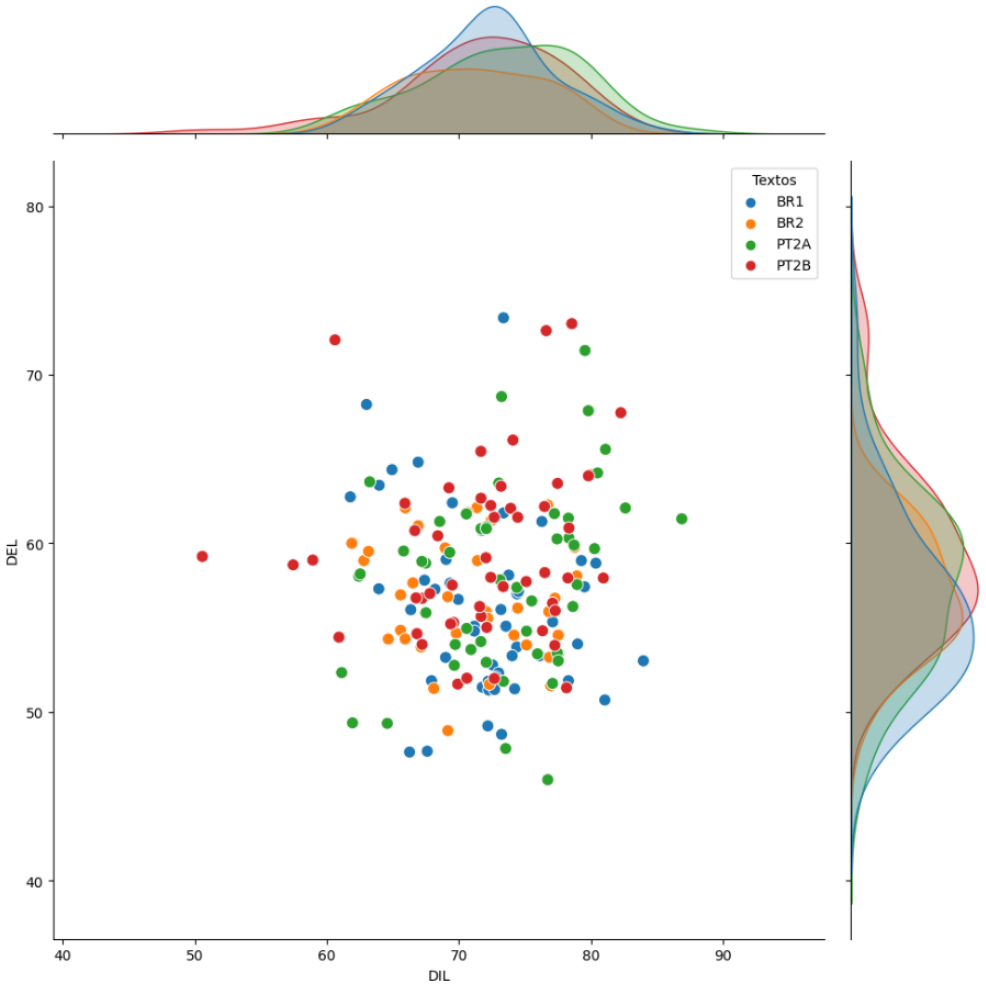
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

Abaixo está uma plot de dispersão mostrando que as distribuições de DIL e DEL nos eixos e a concentração das produções. Nota-se que não há linha divisória que consiga categorizar os dados em classes separadas.

```
sns.jointplot(data=todos, x="DIL", y="DEL", hue="Textos", height=10, s=75)
plt.savefig("dispersao.png")
plt.show()
```



analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

Trasformação de DIL e DEL

Para trazer o conceito de ordem entre os textos, decidimos aplicar a norma das variáveis DIL e DEL.

$$RL = \sqrt{DIL^2 + DEL^2}$$

```
[ ] todos['RL'] = np.sqrt(todos['DIL']**2 + todos['DEL']**2)
    todos.describe()
```

	index	PE	DIL	DEL	label	RL
count	186.000000	186.000000	186.000000	186.000000	186.000000	186.000000
mean	25.645161	106.462366	72.086505	57.621989	1.537634	92.431135
std	15.527235	35.274080	5.626447	5.088370	1.153693	5.535972
min	0.000000	51.000000	50.560000	45.980000	0.000000	77.867336
25%	13.000000	79.250000	68.435000	54.000000	0.000000	88.862075
50%	25.000000	102.000000	72.285000	57.340000	2.000000	92.344154
75%	38.000000	126.750000	76.700000	60.962500	3.000000	95.473583
max	58.000000	252.000000	86.880000	73.370000	3.000000	107.247484

```
[ ] todos
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
0	22	005_D2	72	72.55	52.78	0	BR1	89.717506
1	2	001_D4	151	64.93	64.36	0	BR1	91.422724
2	49	010_D4	66	66.35	56.06	0	BR1	86.862225
3	26	005_D6	78	74.38	53.85	0	BR1	91.827049
4	33	007_D2	64	67.39	57.81	0	BR1	88.788559
...	...	...	...	...	...	...	...	...
181	45	005_D7	156	74.45	61.54	3	PT2B	96.591791
182	7	001_D8	121	67.80	57.02	3	PT2B	88.589618
183	56	006_D8	100	67.21	54.00	3	PT2B	86.215916
184	1	001_D2	121	74.09	66.12	3	PT2B	99.303487
185	16	002_D8	75	70.60	52.00	3	PT2B	87.683294

186 rows × 8 columns

analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

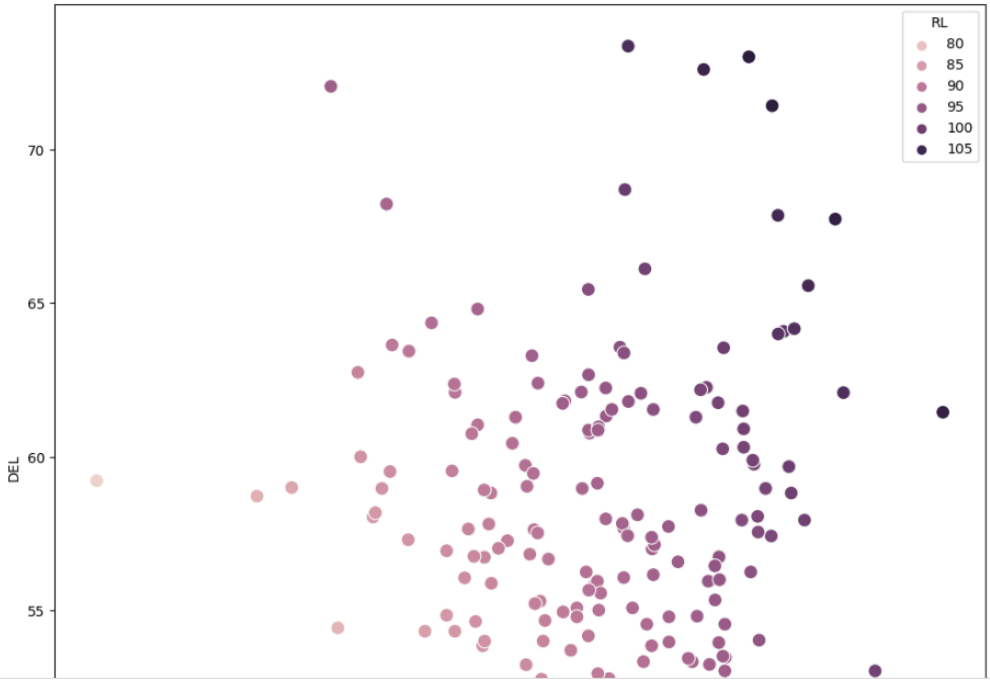
+ Code + Text Connect Colab AI

```
todos.loc[todos['Textos']=='PT2B'].describe()
```

	index	PE	DIL	DEL	label	RL
count	50.000000	50.000000	50.000000	50.000000	50.0	50.000000
mean	31.420000	111.300000	71.594600	59.362000	3.0	93.153569
std	16.592462	32.798861	6.297187	5.139579	0.0	6.127463
min	1.000000	55.000000	50.560000	51.430000	3.0	77.867336
25%	19.250000	93.000000	68.610000	55.745000	3.0	89.003810
50%	31.500000	106.500000	72.260000	58.120000	3.0	93.680940
75%	45.750000	124.750000	76.495000	62.225000	3.0	96.574803
max	58.000000	252.000000	82.260000	73.020000	3.0	107.247484

O gráfico a seguir prova que a medida adotada é proporcional ao desempenho das duas métricas, quando as duas são altas, a RL é alta, quando as duas são baixas, a RL é baixa.

```
[ ] plt.figure(figsize=(12,12))
    sns.scatterplot(data=todos, x="DIL", y="DEL", hue="RL", s=100)
    plt.savefig("dispersao_rl.png")
    plt.show()
```



Para checar se essa proporcionalidade é real de forma quantitativa, podemos usar o coeficiente de correlação de Pearson.

```
np.corrcoef(todos['DIL'], todos['RL'])
```

```
array([[1.        , 0.81262262],
       [0.81262262, 1.        ]])
```

```
[ ] np.corrcoef(todos['DEL'], todos['RL'])
```

```
array([[1.        , 0.62465994],
       [0.62465994, 1.        ]])
```

Estamos interessados nos valores fora da diagonal das matrizes, então a correlação entre DIL e RL foi 0.81, já entre DEL e RL, 0.62. Como ambas correlações foram maiores que 0.6, temos uma correlação positiva.

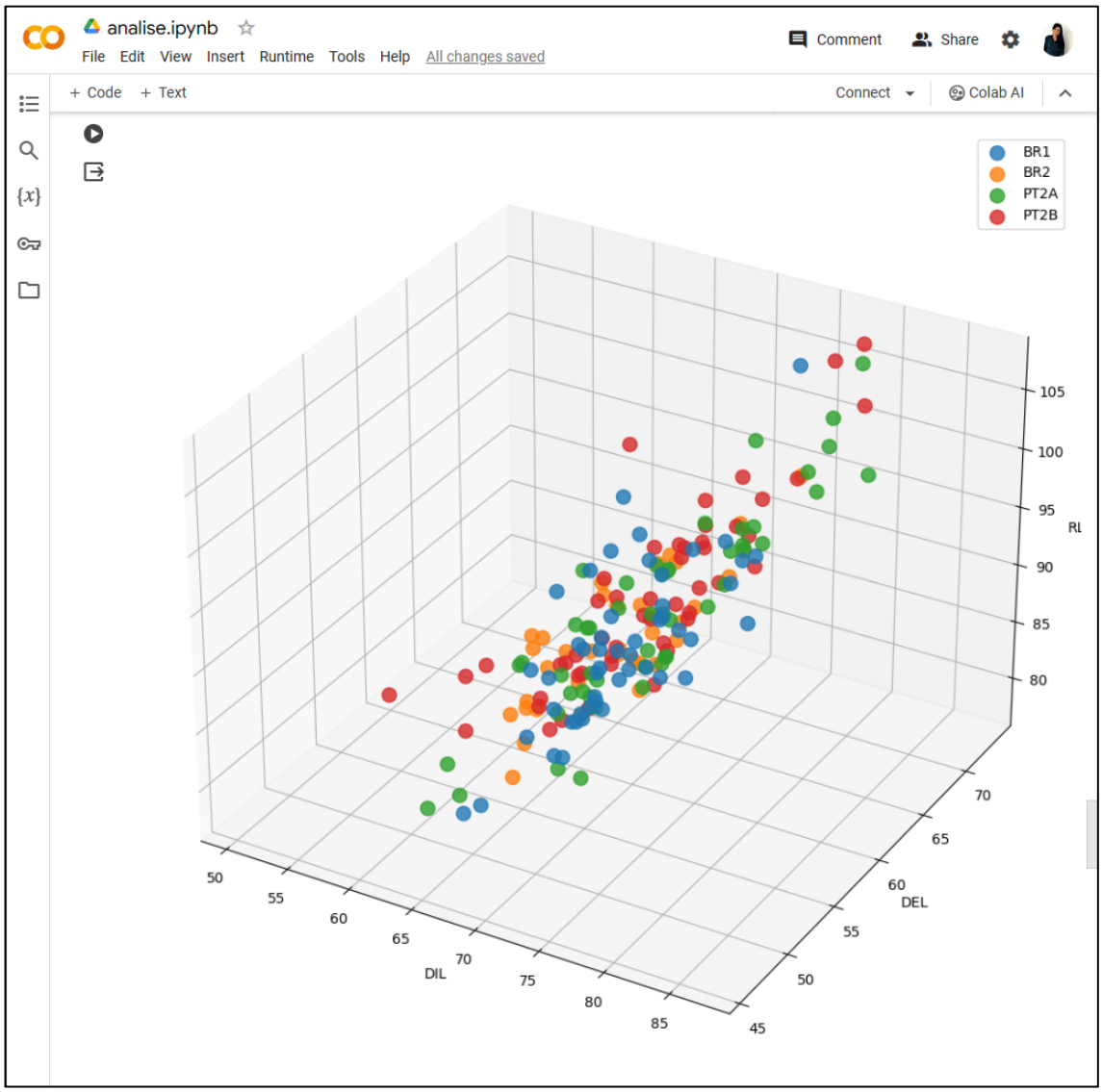
Vamos agora observar se adicionando RL à análise de DIL e DEL conseguimos traçar uma linha (ou um plano, no caso) divisor das categorias.

```
[ ] from mpl_toolkits.mplot3d import Axes3D
    from matplotlib.colors import ListedColormap
    fig = plt.figure(figsize=(12,12))
    ax = fig.add_subplot(projection='3d')

    br1 = todos.loc[todos['label']==0]
    br2 = todos.loc[todos['label']==1]
    ptee = todos.loc[todos['label']==2]
    ptev = todos.loc[todos['label']==3]

    sc = ax.scatter(br1['DIL'], br1['DEL'], br1['RL'], s=100, c='tab:blue', marker='o', alpha=0.8)
    sc = ax.scatter(br2['DIL'], br2['DEL'], br2['RL'], s=100, c='tab:orange', marker='o', alpha=0.8)
    sc = ax.scatter(ptee['DIL'], ptee['DEL'], ptee['RL'], s=100, c='tab:green', marker='o', alpha=0.8)
    sc = ax.scatter(ptev['DIL'], ptev['DEL'], ptev['RL'], s=100, c='tab:red', marker='o', alpha=0.8)
    ax.set_xlabel('DIL')
    ax.set_ylabel('DEL')
    ax.set_zlabel('RL')

    ax.legend(['BR1', 'BR2', 'PT2A', 'PT2B'])
    plt.savefig("dispersao_3d.png")
    plt.show()
```


analise.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Connect
Colab AI
BR1
BR2
PT2A
PT2B
105
100
95
90
85
80
RL
50
55
60
65
70
75
80
85
DIL
45
50
55
60
65
70
DEL

analise.ipynb

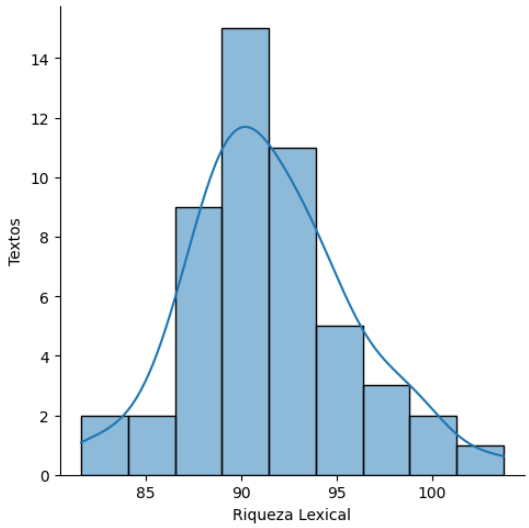
File Edit View Insert Runtime Tools Help All changes saved

Comment Share Connect Colab AI

+ Code + Text

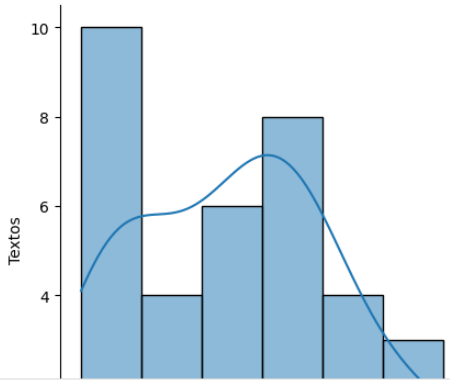
Podemos analisar o comportamento da distribuição da riqueza lexical no Brasil e em Portugal. Primeiro com BR1

```
[ ] g1 = sns.displot(data=todos.loc[todos['Textos']=='BR1'], x="RL", kde=True)
    g1.set_axis_labels("Riqueza Lexical", "Textos")
    plt.savefig("hist_br1.png")
    plt.show()
```



Agora com BR2

```
[ ] g1 = sns.displot(data=todos.loc[todos['Textos']=='BR2'], x="RL", kde=True)
    g1.set_axis_labels("Riqueza Lexical", "Textos")
    plt.savefig("hist_br2.png")
    plt.show()
```



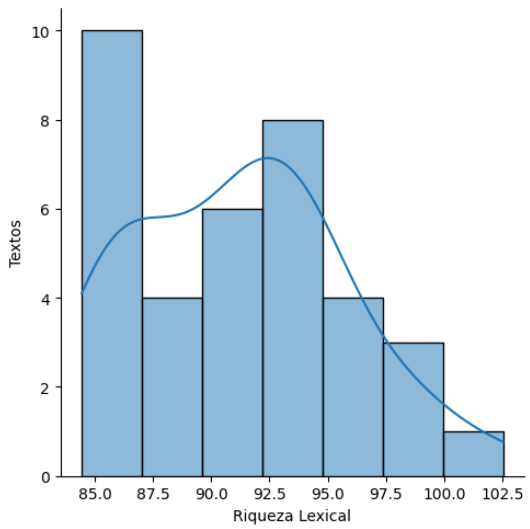
analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share Connect Colab AI

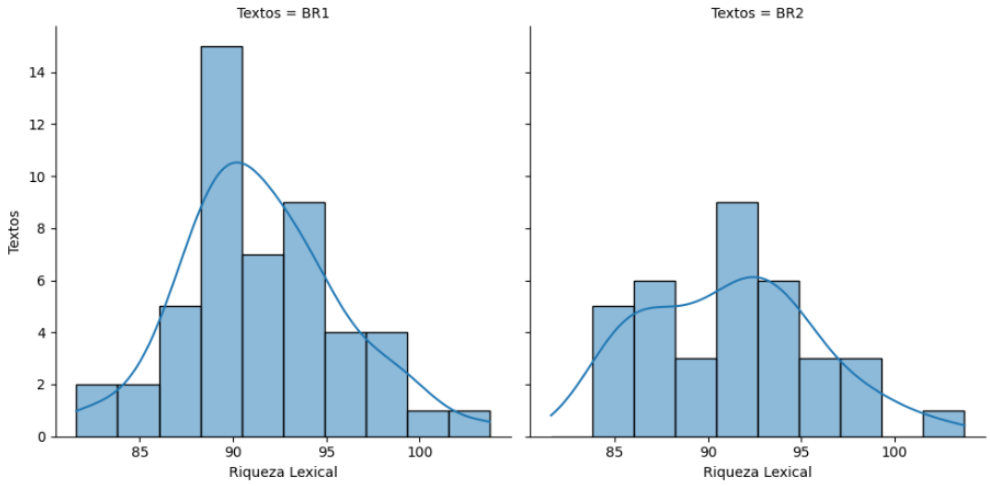
+ Code + Text

```
plt.show()
```



Agora entre as produções brasileiras

```
[ ] g1 = sns.displot(data=todos.loc[np.logical_or(todos['Textos']=='BR1', todos['Textos']=='BR2')], x="RL", col="Textos",
    g1.set_axis_labels("Riqueza Lexical", "Textos")
    plt.savefig("hist_br1_br2.png")
    plt.show()
```



analise.ipynb

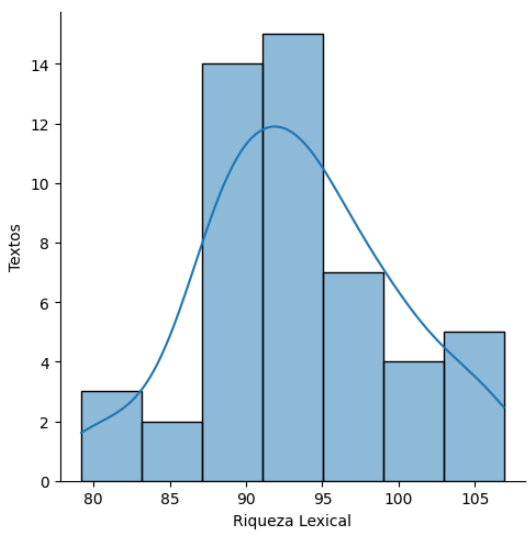
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

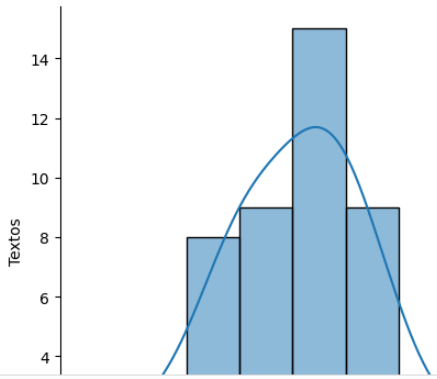
Agora com PT2A

```
g1 = sns.displot(data=todos.loc[todos['Textos']=='PT2A'], x="RL", kde=True)
g1.set_axis_labels("Riqueza Lexical", "Textos")
plt.savefig("hist_ptee.png")
plt.show()
```



Agora com PT2B

```
g1 = sns.displot(data=todos.loc[todos['Textos']=='PT2B'], x="RL", kde=True)
g1.set_axis_labels("Riqueza Lexical", "Textos")
plt.savefig("hist_ptev.png")
plt.show()
```



analise.ipynb

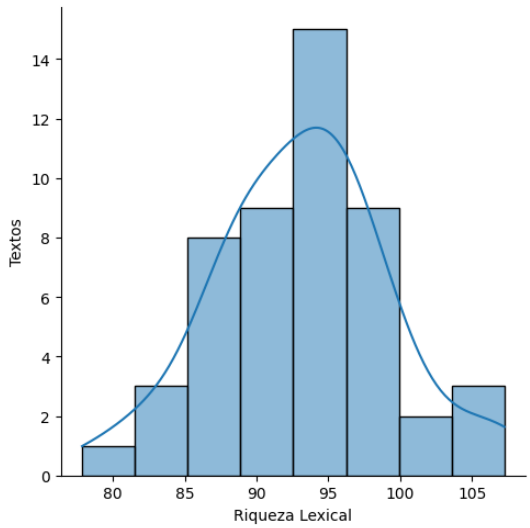
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

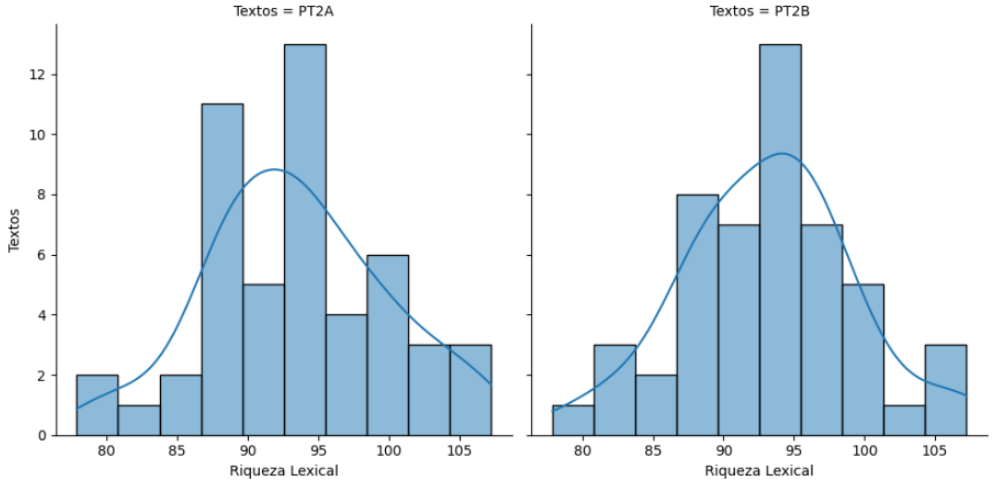
Connect Colab AI

```
g1.set_axis_labels("Riqueza Lexical", "Textos")
plt.savefig("hist_ptev.png")
plt.show()
```



Agora entre os textos portugueses

```
[ ] g1 = sns.displot(data=todos.loc[np.logical_or(todos['Textos']=='PT2A', todos['Textos']=='PT2B')], x="RL", col="Textos
    g1.set_axis_labels("Riqueza Lexical", "Textos")
    plt.savefig("hist_ptee_ptev.png")
    plt.show()
```



analise.ipynb

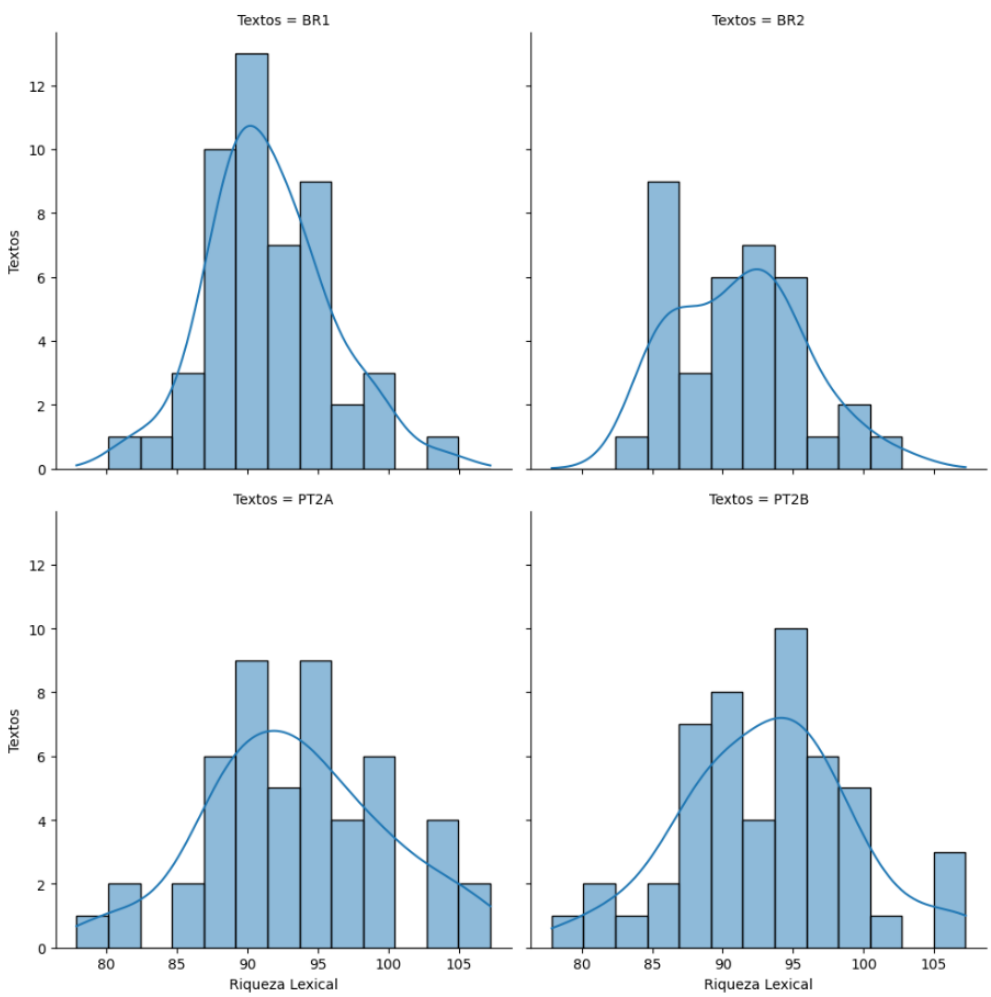
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

Por fim, a comparação entre todos, separado por escola/textos

```
g4 = sns.displot(data=todos, x="RL", col="Textos", kde=True, col_wrap=2)
g4.set_axis_labels("Riqueza Lexical", "Textos")
plt.savefig("hist_geral.png")
plt.show()
```



Com isso, podemos estabelecer um ranking de produções com maior valor de RL

analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text Connect Colab AI

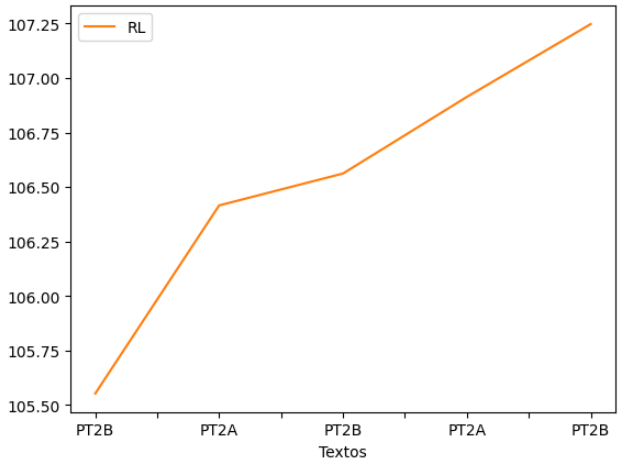
Com isso, podemos estabelecer um ranking de produções com maior valor de RL

```
todos_ordenado = todos.sort_values(by=['RL'])
todos_ordenado.tail()
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
173	28	003_D10	80	76.61	72.61	3	PT2B	105.552377
124	20	003_D2	166	86.88	61.45	2	PT2A	106.415398
155	30	004_D2	93	82.26	67.74	3	PT2B	106.561791
96	10	002_D1	119	79.55	71.43	2	PT2A	106.913270
176	18	002_D10	126	78.55	73.02	3	PT2B	107.247484

```
[ ] todos_ordenado.tail().plot.line(x='Textos', y='RL', color='tab:orange')
```

```
<Axes: xlabel='Textos'>
```



Além das produções com menor valor de RL

```
[ ] todos_ordenado.head()
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
171	14	002_D6	103	50.56	59.22	3	PT2B	77.867336
121	38	005_D4	154	61.94	49.35	2	PT2A	79.195872
123	53	006_D10	86	61.12	52.33	2	PT2A	80.461688
135	9	001_D10	73	64.57	49.32	2	PT2A	81.251137
31	28	006_D2	84	66.26	47.62	0	BR1	81.596887

analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

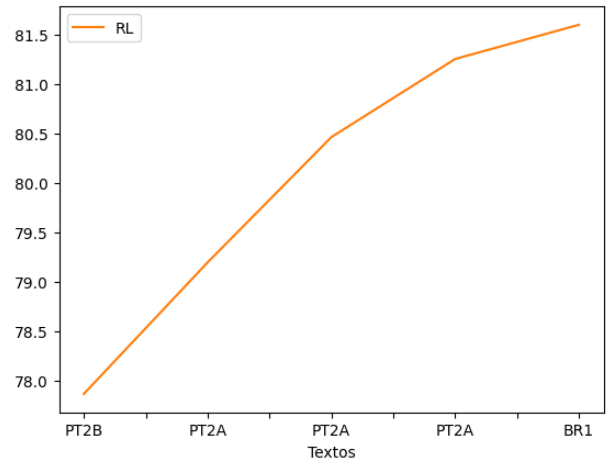
+ Code + Text Connect Colab AI

```
todos_ordenado.head()
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
171	14	002_D6	103	50.56	59.22	3	PT2B	77.867336
121	38	005_D4	154	61.94	49.35	2	PT2A	79.195872
123	53	006_D10	86	61.12	52.33	2	PT2A	80.461688
135	9	001_D10	73	64.57	49.32	2	PT2A	81.251137
31	28	006_D2	84	66.26	47.62	0	BR1	81.596887

```
[ ] todos_ordenado.head().plot.line(x='Textos', y='RL', color='tab:orange')
```

```
<Axes: xlabel='Textos'>
```



Podemos montar um gráfico de barras com os valores crescentes de RL

```
[ ] todos_ordenado
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
171	14	002_D6	103	50.56	59.22	3	PT2B	77.867336
121	38	005_D4	154	61.94	49.35	2	PT2A	79.195872
123	53	006_D10	86	61.12	52.33	2	PT2A	80.461688
135	9	001_D10	73	64.57	49.32	2	PT2A	81.251137
31	28	006_D2	84	66.26	47.62	0	BR1	81.596887
...	...	...	...	...	...	...	...	...
173	28	003_D10	80	76.61	72.61	3	PT2B	105.552377
124	20	003_D2	166	86.88	61.45	2	PT2A	106.415398

analise.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

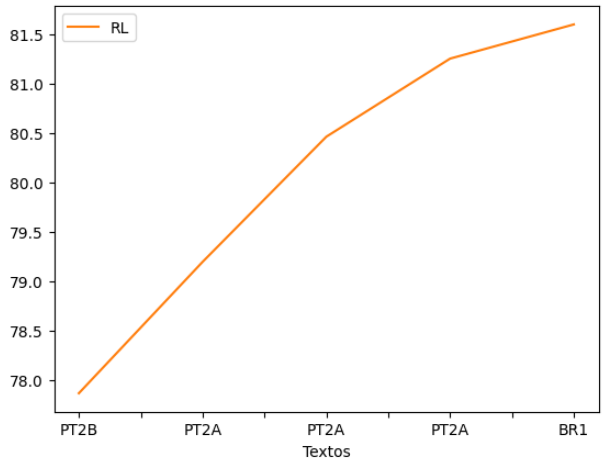
+ Code + Text Connect Colab AI

```
todos_ordenado.head()
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
171	14	002_D6	103	50.56	59.22	3	PT2B	77.867336
121	38	005_D4	154	61.94	49.35	2	PT2A	79.195872
123	53	006_D10	86	61.12	52.33	2	PT2A	80.461688
135	9	001_D10	73	64.57	49.32	2	PT2A	81.251137
31	28	006_D2	84	66.26	47.62	0	BR1	81.596887

```
[ ] todos_ordenado.head().plot.line(x='Textos', y='RL', color='tab:orange')
```

```
<Axes: xlabel='Textos'>
```



Podemos montar um gráfico de barras com os valores crescentes de RL

```
[ ] todos_ordenado
```

	index	Producao	PE	DIL	DEL	label	Textos	RL
171	14	002_D6	103	50.56	59.22	3	PT2B	77.867336
121	38	005_D4	154	61.94	49.35	2	PT2A	79.195872
123	53	006_D10	86	61.12	52.33	2	PT2A	80.461688
135	9	001_D10	73	64.57	49.32	2	PT2A	81.251137
31	28	006_D2	84	66.26	47.62	0	BR1	81.596887
...	...	...	...	...	...	...	...	...
173	28	003_D10	80	76.61	72.61	3	PT2B	105.552377
124	20	003_D2	166	86.88	61.45	2	PT2A	106.415398

155	30	004_D2	93	82.26	67.74	3	PT2B	106.561791
96	10	002_D1	119	79.55	71.43	2	PT2A	106.913270
176	18	002_D10	126	78.55	73.02	3	PT2B	107.247484

186 rows × 8 columns

Estatística inferencial com Shapiro

O teste de Shapiro-Wilk avalia a hipótese nula de que a distribuição de uma determinada variável aleatória é gaussiana (normal).

```
[ ] print(stats.shapiro(br1['RL']))
    print(stats.shapiro(br2['RL']))
    print(stats.shapiro(ptee['RL']))
    print(stats.shapiro(ptev['RL']))
```

```
ShapiroResult(statistic=0.9785518646240234, pvalue=0.4926373064517975)
ShapiroResult(statistic=0.9596760272979736, pvalue=0.2100713849067688)
ShapiroResult(statistic=0.9773673415184021, pvalue=0.44705623388290405)
ShapiroResult(statistic=0.9860053658485413, pvalue=0.8138304352760315)
```

Como todas as amostras apresentaram pvalue maior que 0.05, a hipótese nula não é rejeitada, ou seja, elas seguem o comportamento de uma distribuição normal.

Estatística inferencial com ANOVA

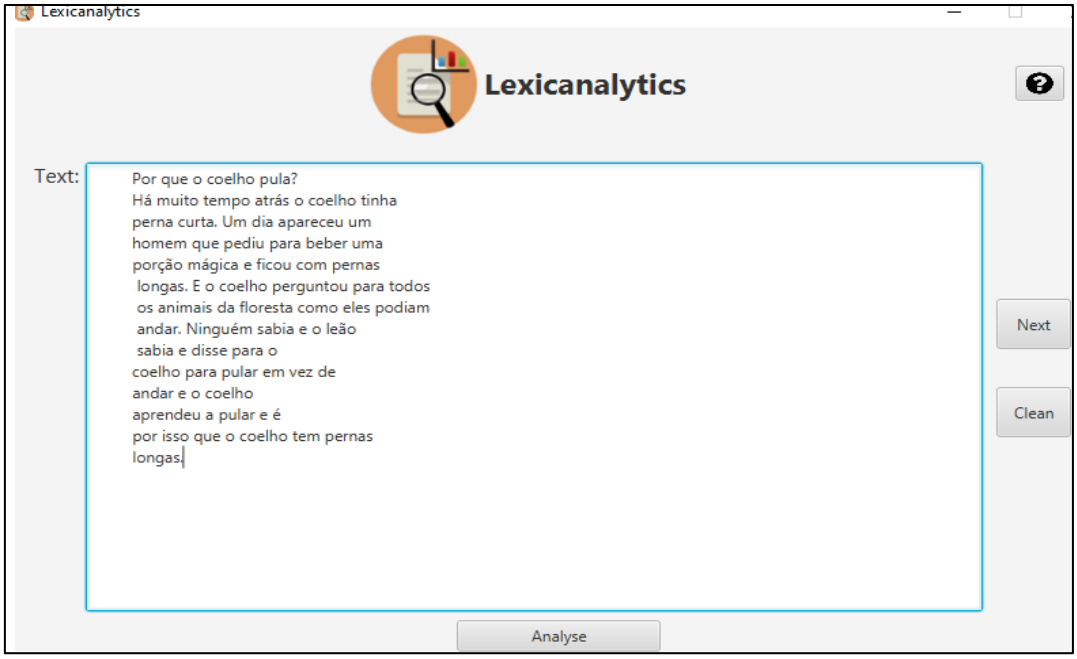
Realizando teste paramétrica ANOVA para distribuição gaussiana. Mais especificamente estamos usando a one-way anova para amostragem múltipla.

```
[ ] print('BR1, BR2, PT2A e PT2B:', stats.f_oneway(br1['RL'], br2['RL'], ptee['RL'], ptev['RL']))
    print('BR1 e BR2:', stats.f_oneway(br1['RL'], br2['RL']))
    print('PT2A e PT2B:', stats.f_oneway(ptee['RL'], ptev['RL']))
```

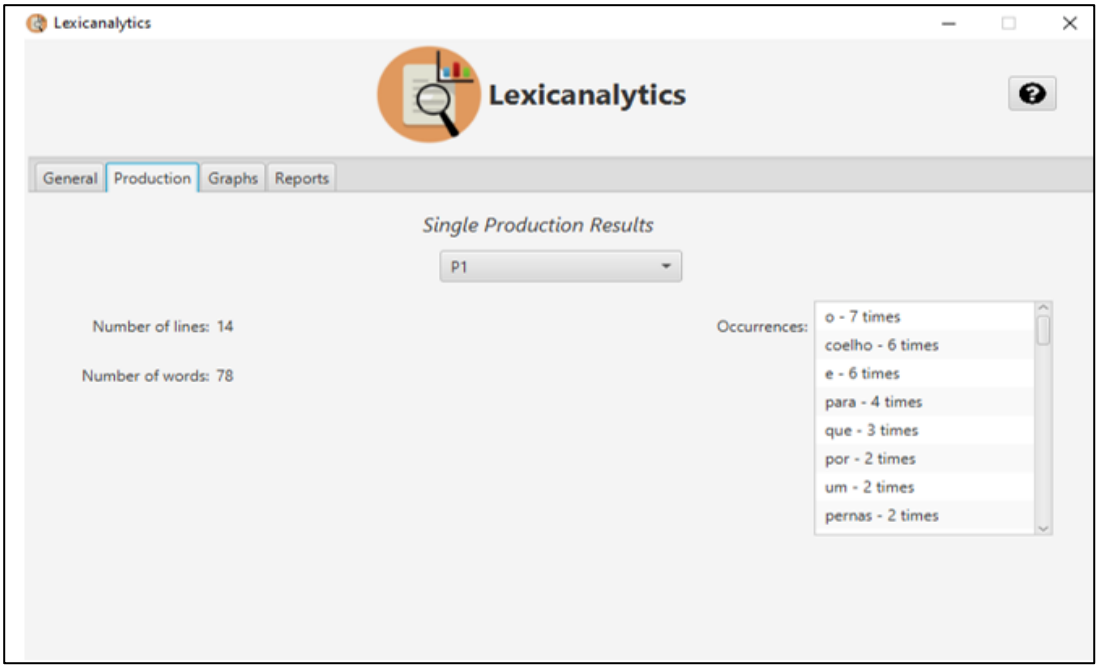
```
BR1, BR2, PT2A e PT2B: F_onewayResult(statistic=1.905143682957977, pvalue=0.13033655430399124)
BR1 e BR2: F_onewayResult(statistic=0.10341527277717798, pvalue=0.7485678912262412)
PT2A e PT2B: F_onewayResult(statistic=0.06427034229436682, pvalue=0.8004006176496209)
```

ANEXO 3 - Primeira versão do *Lexicanalytics Web* (2016)

Tela inicial do *Lexicanalytics*

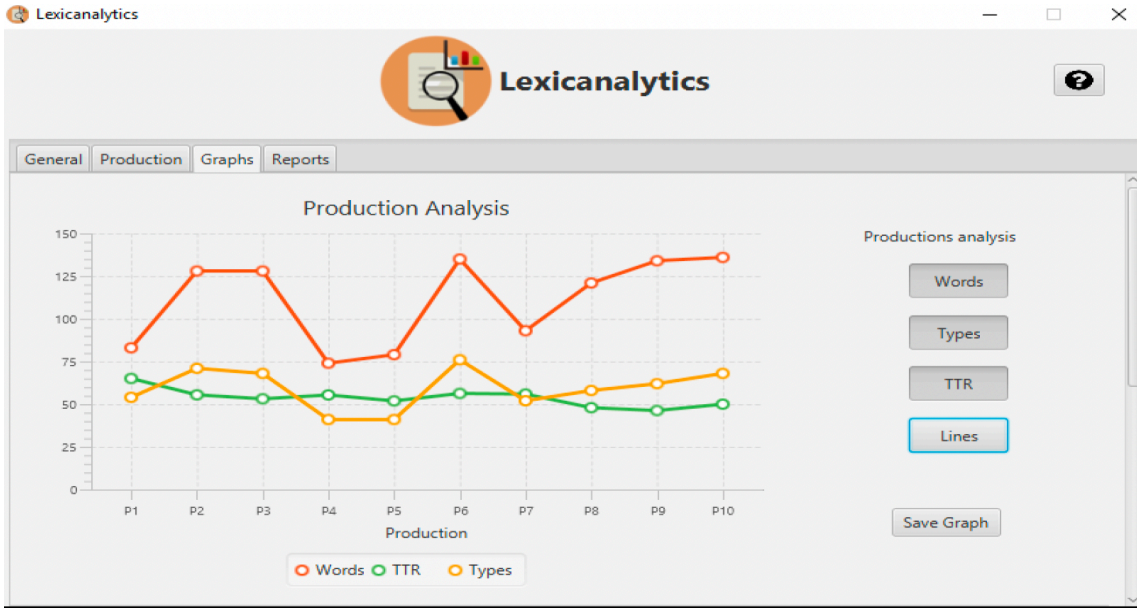


Tela de resultados 1



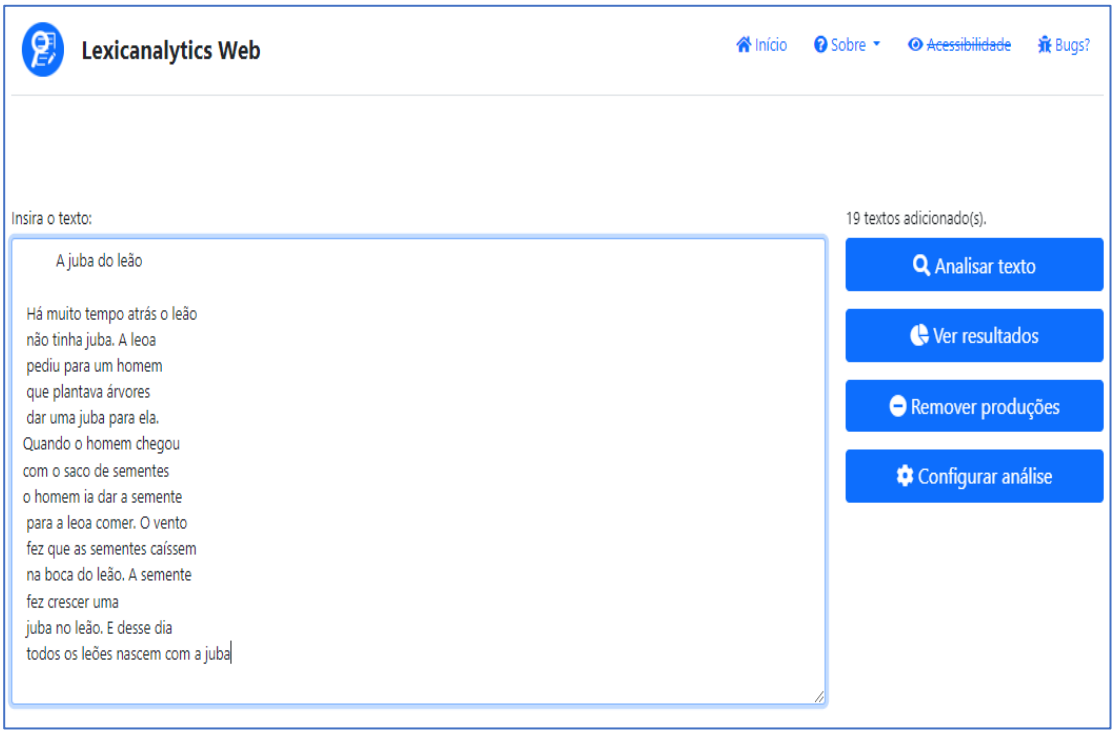
Tela de resultados 2

Tela da representação gráfica dos resultados



ANEXO 4 - Segunda versão do Lexicanalytics Web (2021)

Tela inicial do *Lexicanalytics web*

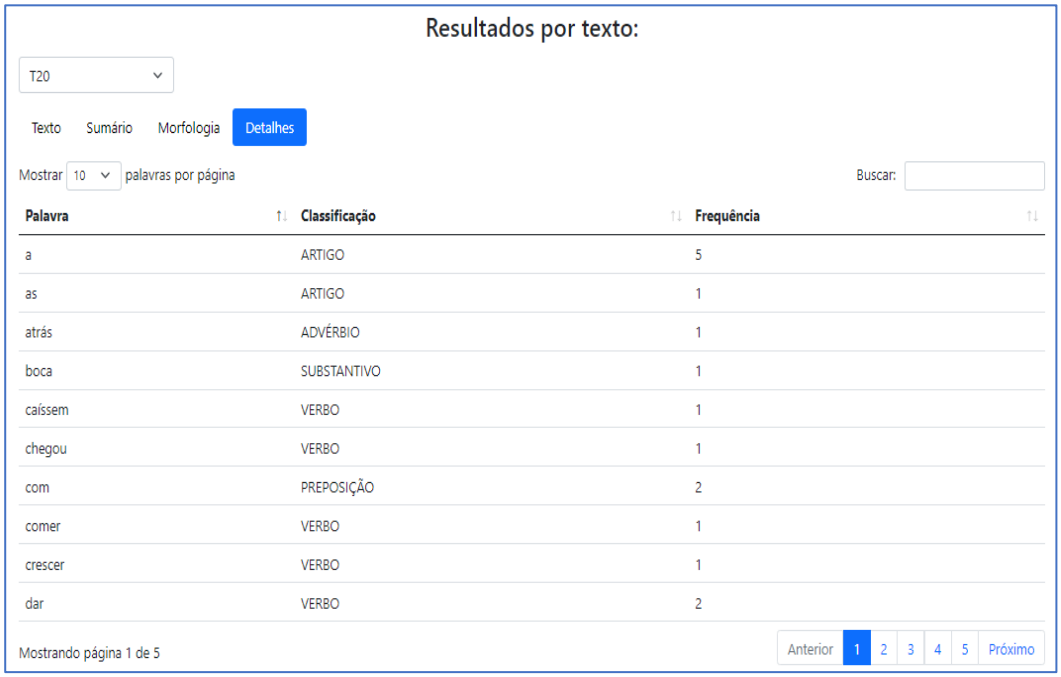


Tela de Resultado Geral

Métrica	Média	Mediana	Moda	Desvio Padrão	Menor	Maior
Número de Palavras	84.63	84.63	55	25.40	53 (T12)	141 (T19)
Vocabulário	50.58	50.58	40	12.65	35 (T18)	81 (T19)
Diversidade Lexical	72.42	72.42	77.52	4.86	65.58 (T18)	78.93 (T3)
Densidade Lexical	56.61	56.61	55.95	3.50	51.39 (T8)	62.26 (T12)
Substantivos	20.74	20.74	17	5.87	12 (T1)	36 (T19)
Verbos	18.89	18.89	16	6.69	10 (T11)	36 (T5)
Adjetivos	2.63	2.63	0	2.21	0 (T4)	7 (T13)
Advérbios	5.00	5.00	3	3.21	1 (T7)	13 (T5)
Pronomes	9.16	9.16	8	5.27	1 (T11)	20 (T2)
Artigos	11.58	11.58	8	3.86	7 (T10)	19 (T19)

Tela dos Resultados individuais por texto

Tela da classificação morfológica e ocorrência de palavras



Tela com mensagem de erro