Congresso Brasileiro sobre Alfabetização, Linguagens e Letramentos

ISBN: 978-65-86901-91-7

LEXICANALYTICS WEB: UMA FERRAMENTA DE ANÁLISE LEXICAL PARA AUXILIAR NO PROCESSO DE AVALIAÇÃO DA ESCRITA ESCOLAR

ELIAN SANTOS[1]
GEOVANE LEITE DE CARVALHO FILHO[2]
GLAUBER RODRIGUES LEITE[3]
EDUARDO CALIL[4]

RESUMO

Este trabalho apresenta o ***Lexicanalytics web***, uma ferramenta de análise lexical que pode auxiliar no acompanhamento e diagnóstico dos conhecimentos lexicais dos alunos recém-alfabetizados. Adotando técnicas de Processamento de Linguagem Natural e de estatística, este ***software*** extrai medidas como Densidade Lexical e Diversidade Lexical, além oferecer um levantamento do vocabulário e das classes gramaticais utilizadas pelos escreventes. Caracterizada como uma pesquisa de estudo de caso exploratório e seguindo uma abordagem qualitativa para avaliação dos resultados, nosso percurso metodológico é constituído pelas seguintes etapas: i) identificação dos ***softwares*** disponíveis que fazem a mensuração direta da DeL e DiL; ii) Etapa de desenvolvimento do ***Lexicabalytics,*** que compreende o levantamento dos requisitos, a seleção das métricas e detalhamento de seus componentes; iii) Testes e validações do ***software*** através de um ***corpus***

1 Doutoranda em Educação na Universidade Federal de Alagoas – UFAL, elian.santos@cedu.ufal.br

2 Graduando em Ciência da Computaçã na Universidade Federal de Alagoas – UFAL, geovane.filho@arapiraca.ufal.br

3 Doutorando em Engenharia Elétrica e da Computação na Universidade Federal do Rio Grande do Norte -UFRN, glauber.leite.090@ufrn.edu.br

4 Professor orientador: Doutor em linguística e Professor do Centro de Educação da Universidade Federal de Alagoas (CEDU/UFAL), calil@cedu.ufal.br.

ISBN: 978-65-86901-91-7

composto por textos de alunos recém-alfabetizados. A partir dos resultados das etapas de desenvolvimento, testes e validações, foi possível verificar que o *Lexicanalytics web* pode oferecer aos professores e pesquisadores um maior acesso aos conhecimentos lexicais dos alunos, assim como novos caminhos para auxiliar na tomada de decisões pedagógicas para o ensino -aprendizagem da escrita textual.

Palavras-chave: Densidade Lexical; Diversidade Lexical; Produção Textual; Escrita Infantil.

ISBN: 978-65-86901-91-7

INTRODUÇÃO

O uso dos recursos computacionais para a análise de dados textuais tem sido um grande aliado no processo de extração e visualização dos aspectos lexicais da escrita dos alunos. Seja através de técnicas básicas como a quantificação do número de palavras do texto ou de técnicas mais complexas como a classificação morfológicas, usando como aporte a Inteligência Artificial (IA), a tecnologia vem oferecendo ao campo educacional subsídios para auxiliar na tomada de decisão de metodologias de ensino, voltadas para a produção textual em sala de aula.

Dentro deste cenário, encontra-se o Processamento de Linguagem Natural (PLN)[5] um ramo da IA, que utiliza técnicas computacionais para aprender, entender e produzir conteúdo na linguagem humana a fim de interpretá-la e automatizá-la (FILATRO, 2021). No que concerne à avaliação de dados textuais, o PLN atua em dois principais aspectos: i) Estrutura, composta pela análise da morfologia e sintaxe; ii) Significado, envolvendo a semântica e pragmática.

Na estrutura é analisada a morfologia, voltada para o reconhecimento e classificação das palavras, e a sintaxe que se concentra na definição da estrutura da frase, tomando como base a forma como as palavras se relacionam. Já o estudo do significado pelo PLN se concentra na semântica, através da análise de associação do significado das palavras a sua estrutura e, por fim, a pragmática, voltada para a verificação do significado associando-o a uma estrutura sintática, considerando o contexto mais apropriado para o uso de determinada palavra (CRUZ, 2015). Vale destacar que a natureza interdisciplinar do PLN possibilita a "integração de áreas como a linguística, ciência da computação, psicologia e educação" (FILATRO, 2021, p 129).

Partindo dessa perspectiva, podemos compreender que o aporte tecnológico nos permite explorar áreas que antes não eram acessadas, principalmente em contexto de análise em larga escala, visto que esse tipo ação, executada de forma manual, seria demorada e suscetível a erros. Diante deste contexto, o presente trabalho tem como objetivo apresentar o ***Lexicanalytics web,*** um ***software*** voltado para a avaliação

5 Do inglês ***Natural Language Processing*** - NLP

ISBN: 978-65-86901-91-7

e acompanhamento do desenvolvimento lexical dos alunos a partir da Densidade Lexical (DeL) e Diversidade Lexical (DiL) dos seus textos. Ao longo deste artigo apresentaremos o referencial teórico que fundamenta as métricas adotadas, assim como o detalhamento das técnicas implementadas no desenvolvimento do ***Lexicanalytics web***. Por fim, demonstraremos o passo a passo de sua utilização na avaliação de um ***corpus*** composto por produções textuais de alunos recém-alfabetizados, matriculados no 2° ano do Ensino Fundamental.

METODOLOGIA

Caracterizada como uma pesquisa de estudo de caso exploratório e seguindo uma abordagem qualitativa para avaliação dos resultados, o percurso metodológico seguiu as seguintes etapas:

i. Identificação dos ***softwares*** disponíveis que fazem a mensuração direta da DeL e DiL.
ii. Etapa de desenvolvimento do ***Lexicabalytics,*** que compreende o levantamento dos requisitos, a seleção das métricas e detalhamento de seus componetes.
iii. Testes e validações do ***software*** através de um ***corpus*** composto por textos de alunos recém-alfabetizados.

Para a etapa de testes e validações, foi adotado um conjunto de 20 narrativas ficcionais, produzidas por alunos recém-alfabetizados matriculados no 2º ano do Ensino Fundamental, de uma escola da rede particular de Maceió/AL. Os textos foram produzidos a partir de uma metodologia de escrita a dois, isto é, de uma escrita colaborativa (CALIL, 2019). Vale ressaltar que o ***corpus*** em questão faz parte do acervo do nosso laboratório de pesquisa, LAME, e sua coleta foi realizada no ano de 2012.

REFERENCIAL TEÓRICO

A avaliação da escrita dos alunos através de medidas quantitativas é uma prática recorrente na literatura, sendo possível identificar seu uso em diferentes contextos, seja para determinar a taxa de repetitividade das palavras de um texto oral ou escrito (JHOMSON, 1915; TEMPLIM, 1957), para estimar o vocabulário (CARROLL, 1938) ou

ISBN: 978-65-86901-91-7

mensurar a taxa de densidade informacional na escrita de crianças e adolescentes (URE, 1971; HALLIDAY, 1985).

Uma das medidas consolidadas na literatura para a avaliação do conhecimento lexical é a Riqueza Lexical, definida como "uma característica multidimensional da escrita, mensurada através de alguns indicadores linguísticos como a Densidade Lexical e Diversidade Lexical" (READ, 2000, p. 200). A Diversidade Lexical (DiL) representa a variedade ou diversidade de palavras usadas pelo escrevente na produção de um texto, indicando seu alcance vocabular (KIM, 2014) e a qualidade de sua escrita (MCCARTHY e JARVIS, 2007). Sua mensuração é historicamente fundamentada na classificação das palavras em ***types*** (compreendida como palavras ortograficamente diferentes, considerando apenas sua primeira ocorrência) e ***tokens*** (todas as palavras escritas no texto, considerando todas as suas ocorrências).

Uma métrica tradicional para essa mensuração é a ***Type Token Ratio*** (TTR), representada pela equação

$$TTR = \frac{Typens}{Tokens}$$

Por outro lado, estudos sobre a confiabilidade da TTR (JARVIS e MCCARTHY; 2005, 2010) concluíram que seus resultados tendem a demonstrar sensibilidade a variações do comprimento do texto, provocando uma relação inversa entre comprimento e valor da DiL, ou seja, textos extensos apresentavam baixa DiL, enquanto os textos menores apresentavam altas taxas.

Como alternativa para corrigir esse problema, estudiosos como MacCarthy e Jarvis (2010) desenvolveram a medida ***Hypergeometric Distribution D-index***[6] (HD-D), que extrai (sem reposição) uma amostra aleatória de 42 palavras do texto e calcula a probabilidade de todos os ***types*** (MACCARTHY e JARVIS, 2010, p. 383). Após esse levantamento, as probabilidades são somadas e o resultado dessa soma indica o valor da DiL do texto. Vale salientar que por ser uma técnica complexa, para medir a DiL em um conjunto textos ou em larga escala é necessário um aporte computacional.

6 Tradução: Índice D para distribuição Hipergeométrica.

ISBN: 978-65-86901-91-7

Como mencionado anteriormente, outro componente da Riqueza Lexical é a Densidade Lexical (DeL) (READ, 200), representada pela proporção de itens ou palavras com valores lexicais (substantivos, adjetivos, verbos e advérbios terminados com o sufixo 'mente') do texto (URE, 1971; HALLIDAY, 1985). Assim como a DiL, a avaliação da DeL também faz parte da classificação das palavras do texto, neste caso, classificando-as em itens lexicais ou gramaticais.

Segundo Halliday (1985), os itens lexicais são considerados como aquelas palavras que concentram a informação no texto, isto é, durante uma produção textual o escrevente precisa discorrer sobre uma determinada temática, seja para construir uma narrativa ficcional, um texto dissertativo, descritivo entre outros gêneros textuais, mas, para isso acontecer, torna-se necessário recorrer aos itens lexicais. Por outro lado, os itens gramaticais definidos como artigos, pronomes, numerais, preposições, conjunções e interjeições, são palavras que exercem apenas a função de conectar um item lexical ao outro ou uma sentença a outra.

A aferição da DeL pode ser feita através do método de Halliday (1985) que consiste na razão dos itens lexicais (substantivos, adjetivos, verbos e advérbios modais) pelo total de palavras. O resultado obtido dessa operação pode assumir valores entre 0 e 1 ou em porcentagem de 0% a 100%. De modo geral, seu percurso metodológico é composto pelas seguintes etapas: i) quantificação do total de palavras do texto; ii) classificação dos itens lexicais; iii) quantificação dos itens lexicais, considerando todas as suas ocorrências; iv) cálculo da razão de itens lexicais pelo total de palavras.

Sobre a aplicação da DiL e da DeL, na literatura é possível constatar sua utilização em estudos que avaliam a escrita dos estudantes em diferentes contextos educacionais, seja para analisar as contribuições que a DiL traz para a qualidade de composições textuais de estudantes universitário (GONZÁLES, 2017), seja na avaliação da progressão escolar de crianças e adolescentes (JOHANSSON, 2009; MARTINS, 2016), ou para analisar como os diferentes gêneros textuais podem impactar na DiL dos alunos (SADEGHI e DILMAGHANI, 2013).

Contudo, apesar da potencialidade desses indicadores para avaliação da riqueza lexical do texto, ainda é observado um número reduzido de programas destinados a sua aferição, principalmente adotando métricas atualizadas para o cálculo da DiL. Dos programas encontrados

ISBN: 978-65-86901-91-7

para este fim, podemos destacar o ***Text Inspector***[7], o ***Computerized Language ANalysis*** (CLAN)[8], ***softwares*** que fornecem apenas a medição da diversidade lexical de forma gratuita ou parcialmente gratuita (para textos de até 250 palavras). Foi a partir desse cenário que surgiu a motivação para o desenvolvimento do ***Lexicanalytics,*** cuja principal finalidade é oferecer a comunidade escolar um ***software*** para auxiliar na avaliação e acompanhamento dos conhecimentos lexicais dos alunos.

RESULTADOS E DISCUSSÃO

DESENVOLVIMENTO DO LEXICANALYTICS WEB

Como ponto de partida para o desenvolvimento do ***Lexicanalytics***, foi realizado uma busca por ferramentas voltadas para a análise da Língua Portuguesa e que possuíssem em suas funcionalidades a extração direta da DiL e DeL. A partir disso, foi iniciado o processo de especificação de requisitos para descrever suas funcionalidades a partir das necessidades estabelecidas, neste caso, a avaliação lexical de textos. Posteriormente, foi estruturada a arquitetura do programa, detalhando seus componentes e tecnologias que seriam usadas, como mostra a Figura 1:

7 Disponível em https://textinspector.com/help/lexical-diversity/

8 Disponível gratuitamente em https://dali.talkbank.org/clan/

ISBN: 978-65-86901-91-7

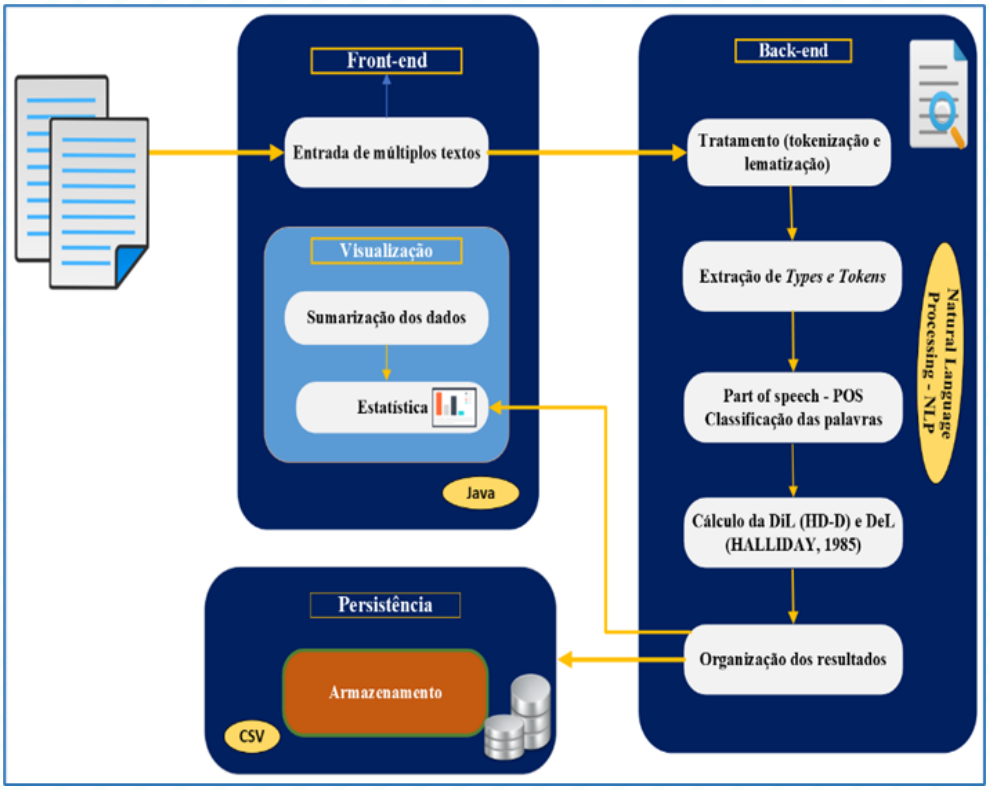

Figura 1- Arquitetura do *Lexicanalytics*

O back-end é o componente do sistema que irá tratar do pré-processamento (antes do cálculo da DeL e DiL), o processamento do texto e gerenciamento de armazenamento (através de um serviço de persistência). Na etapa de pré-processamento, acontece a tokenização e lematização dos textos. Na tokenização, as palavras são separadas e analisadas individualmente, além disso também são removidos caracteres como pontos, vírgulas, aspas, parênteses, chaves, colchetes, dentre outros. Já na lematização, as palavras são analisadas e agrupadas de acordo com suas formas flexionadas. Vale ressaltar que a lematização é essencial para o sistema diferenciar e classificar as palavras de acordo com sua classe gramatical. Por fim, para finalizar o pré-processamento, são extraídos do texto seu número de ***types*** e ***tokens***.

Na etapa de processamento do texto, o ***back-end*** realiza todas as ações para extração dos dois indicadores linguísticos. A mensuração da DeL foi realizada através de uma classificação morfológica usando um

ISBN: 978-65-86901-91-7

POS-tagger[9], um classificador de palavras para língua portuguesa treinado usando PLN. Em seguida, o sistema aplica o método de Halliday (1985), isto é, calcula a razão de itens lexicais (palavras que produzem maior densidade informacional no texto, como os substantivos, verbos, adjetivos e advérbios terminados com o sufixo 'mente') pelo total de palavras do texto.

Para medição da DiL, o sistema avalia parâmetros em uma distribuição Hipergeométrica, esses parâmetros são usados como índices representando a DiL através do algoritmo HD-D. Por fim, após essa etapa de extração concluída, o ***back-end*** organiza os dados e encaminha para o ***front-end***, onde os dados brutos serão calculados estatisticamente e representados através de tabelas e gráficos. Todas as informações geradas no ***Lexicanalytics web*** são encaminhadas e armazenadas na persistência.

LEXICANALYTIC WEB: O PASSO DA AVALIAÇÃO TEXTUAL

Neste tópico exibiremos a interface do programa, assim como a demonstração do passo a passo de sua execução em uma análise de 20 textos. Outro ponto importante a destacar é que os resultados aqui reportados são provenientes da versão funcional do programa, executada em um navegador que atualmente está acessível apenas localmente para os seus desenvolvedores. Dito isto, para a análise de um texto é necessário seguir os seguintes passos:

9 Disponível em: https://github.com/inoueMashuu/POS-tagger-portuguese-nltk

ISBN: 978-65-86901-91-7

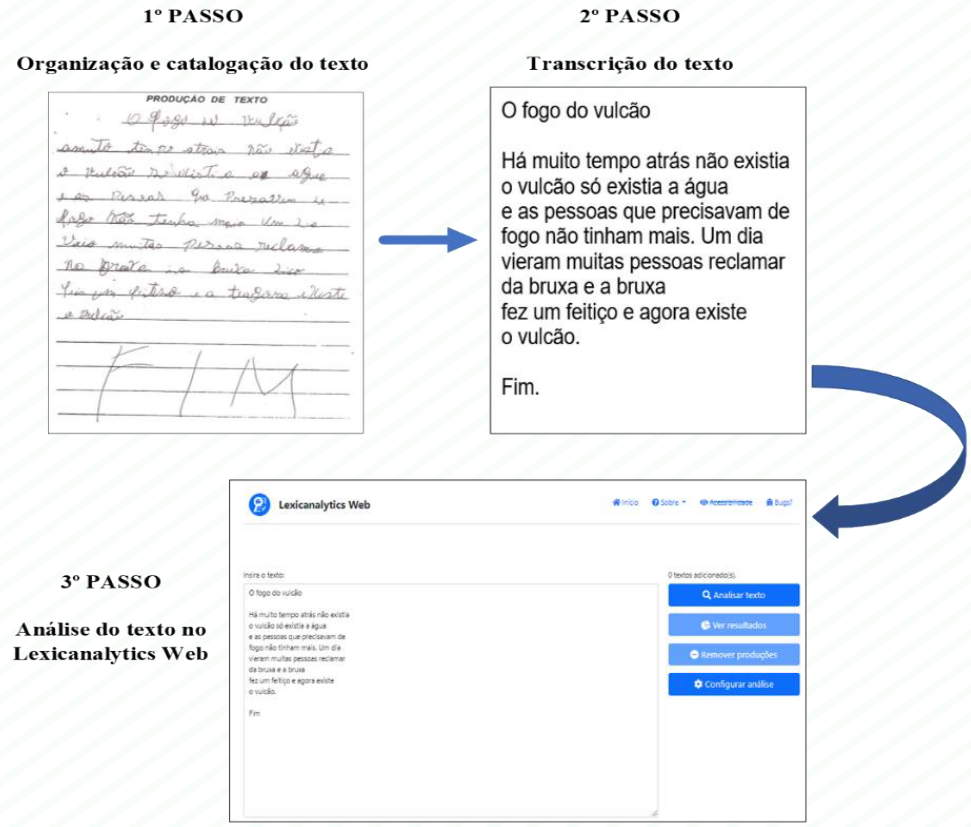

Figura 2 – Passos da preparação do texto para inserção e avaliação no *Lexicanalytics*

A etapa de catalogação ajuda na organização dos resultados, principalmente quando se trabalha com um número considerável de textos para a análise, enquanto a transcrição é essencial para análise lexical, pois nela são ajustadas as separações de palavras, marcas de rasuras e erros de ortografia. Seguindo esse contexto, destacamos também que para a extração dos aspectos lexicais o programa não considera as pontuações, ou seja, o foco é exclusivamente nos aspectos das palavras, o que envolve sua posição no texto, sua função e significado.

No contexto da presente pesquisa, na etapa de transcrição dos textos adotamos o *Microsoft Word*, mas pode ser usado qualquer outro recurso de transcrição digital. Por fim, é necessário observar que para uma análise mais segura e real dos textos produzidos pelos alunos não são permitidas neste processo de transcrição a adição ou exclusão de palavras. Dito isto, seguindo os passos da avaliação no *Lexicanalytics,*

ISBN: 978-65-86901-91-7

após a transcrição do texto, basta copiá-lo e colá-lo na janela inicial do programa (Figura 3):

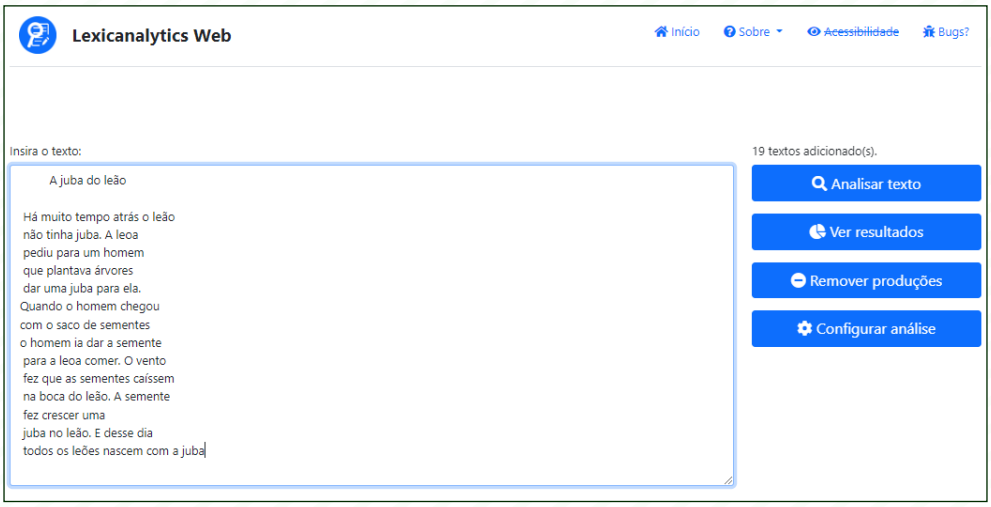

Figura 3 - Tela inicial do *Lexicanalytics web.*

Após adicionar as produções, seleciona-se no lado direito da tela a opção "Ver resultados". Mas, para analisar um conjunto de textos, é necessário inserir um por vez, isto é, insere-se o primeiro texto na aba inicial do programa e clica em "Analisar texto". Esse processo deve ser repetido até que seja adicionado o número desejado de textos. Caso, algum texto seja adicionado erroneamente, é possível removê-lo clicando na opção "Remover Produções". Por fim, para iniciar a análise, basta apenas clicar em "Ver resultados" que uma nova aba aparecerá com as seguintes informações:

ISBN: 978-65-86901-91-7

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)

Figura 4 - Tela de Resultado Geral

Como pode ser visto na figura 4, a tela "Resultado Geral" exibe as principais informações uma extraídas dos textos dos alunos, representadas estaticamente através de medidas como média, moda, mediana e desvio-padrão do número de palavras, do vocabulário, da Diversidade Lexical, Densidade Lexical e das classes gramáticas. Esse recurso oferece aos usuários uma maior visualização do desempenho textual dos alunos, contribuindo para um maior acompanhamento da progressão do vocabulário usado e do comportamento relativo à riqueza lexical. Além de proporcionar uma visão panorâmica dos aspectos lexicais de um conjunto de textos, o programa também gera uma aba com os resultados individuais para cada produção textual. A partir destas informações, o usuário terá um maior acompanhamento do desempenho e progresso de cada aluno. Outro ponto a destacar é a interface interativa do ***lexicanalytics*** ao oferecer a exibição gráfica dos resultados, como ilustra a Figura 5:

ISBN: 978-65-86901-91-7

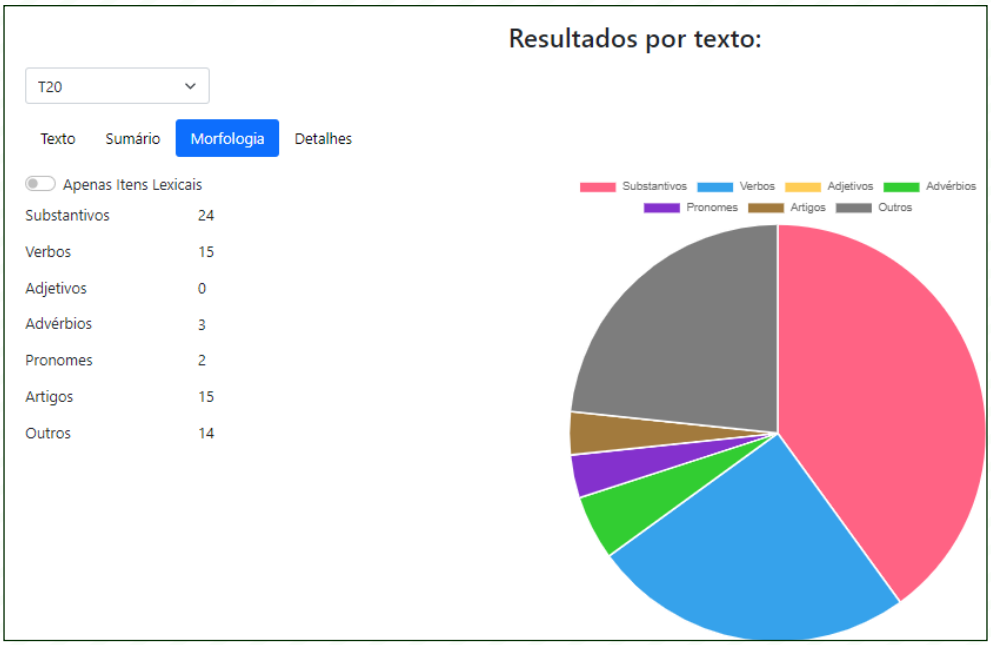

Figura 5 - Resultados individuais por texto

Neste gráfico de setores, os usuários terão acesso ao resumo quantitativo dos itens lexicais, o que poderá contribuir para determinar as classes com maior e menor ocorrência de palavras escrita pelos alunos. Já o resumo das outras classes gramaticais como preposição, numeral, conjunção e interjeição aparecem no quantitativo de "Outros". Como podemos observar no gráfico 5, as palavras que pertencem à classe dos substantivos foram as mais recorrentes no texto de número 20, logo, essas informações geradas pelo programa podem facilitar o acompanhamento dos alunos e auxiliar os professores nos aspectos lexicais que precisam ser mais reforçados no ensino em sala de aula.

Por fim, na aba "Detalhes", encontra-se a classificação morfológica de cada palavra, seguido de sua frequência. Com esse recurso será possível diagnosticar com precisão a palavra mais recorrente em cada produção textual, como exemplo o número de verbos escritos do texto 20 (T20), representado no gráfico 5, que é 15, mas para entender de forma detalhada quais são esses verbos, bastar acessar esta aba "Detalhes". Outra informação importante é a seleção do número de palavras por páginas, no caso da figura abaixo são apresentadas 10 palavras por página, dentre elas os verbos: caíssem, chegou, comer e dar.

ISBN: 978-65-86901-91-7

Figura 6 – Página 1 da classificação morfológica e ocorrência de palavras

Para concluir, o ***lexicanalytics*** também conta com a opção de busca, tornando possível filtrar e analisar com maior agilidade o comportamento de uma palavra específica. Este recurso foi pensado para que os usuários possam identificar a ocorrência das palavras no texto, assim como avaliar se o programa as classificou corretamente.

CONSIDERAÇÕES FINAIS

O conhecimento lexical dos alunos é um fator essencial para sua prática de produção textual, pois é através das palavras que eles imprimem suas visões de mundo e seu domínio de língua portuguesa. A partir dessa prepositiva, e considerando os avanços dos aportes tecnológicos voltados para o uso no contexto escolar, o ***lexicanalytics web*** surge como uma proposta educacional e tecnológica que adota técnicas de Processamento de Linguagem Natural para auxiliar os professores e pesquisadores na extração de informações lexicais dos alunos.

Nesse sentido, esperamos com esta pesquisa possa oferecer uma ferramenta que contribua para o trabalho em sala de aula, proporcionando informações aprofundadas da escrita dos alunos para auxiliar na tomada de decisões pedagógicas no processo de ensino-aprendizagem da escrita textual.

ISBN: 978-65-86901-91-7

Para trabalhos futuros, buscaremos disponibilizar o ***Lexicanalytics Web*** para o público em um servidor persistente, além de implementar técnicas de processamento de imagens para substituir a transcrição manual do texto. E, por conseguinte, pretende-se implementar tecnologias de aprendizagem contínua para correção do sistema, como por exemplo, realização periódica de novos treinamentos no algoritmo de classificação a partir de correção de situações de classificação morfológica de palavras de forma equivocada durante a utilização do sistema.

REFERÊNCIAS

CALIL, E. Sistema Ramos: método para captura multimodal de processos de escritura a dois no tempo e no espaço real da sala de aula. Revista ALFA, n. 63, vol.1, 2019.

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

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

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

HALLIDAY, M. A. K. Spoken and written modes of meaning. In Comprehending oral and writ-ten language, (Eds, Horowitz, R. & Samuels, S.J.) Academic Press, Orlando, 1985.

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

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

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

ISBN: 978-65-86901-91-7

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

SADEGHI, K.; DILMAGHANI, S. The Relationship between Lexical Diversity and Genre in Iranian EFL Learners Writings. Journal of Language Teaching and Research, vol. 4, n. 2, p. 328-334, 2013.

TEMPLIN M. C. Certain language skills in children: their development and interrelations. Westport, CT: Greenwood, 1957.

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.