Análise de Trabalhos de Conclusão de Curso utilizando Técnicas de Processamento de Linguagem Natural

Ricardo Radaelli Meira, PPGIE/UFRGS, ricardoradaelli@gmail.com, 0000-0002-4922-2267
Eliseo Reategui, PPGIE/UFRGS, eliseoreategui@gmail.com, 0000-0002-5025-9710
Regina Motz, UDELAR, rmotz@fing.edu.uy, 0000-0002-1426-562X

Resumo: Este trabalho apresenta um estudo sobre a análise da escrita de trabalhos de conclusão de curso em diferentes áreas, considerando métricas da linguística computacional. Utilizou-se para isso o sistema NILC-Metrix, uma adaptação do sistema Coh-Metrix que utiliza recursos linguísticos para avaliar aspectos como o nível de complexidade e de leiturabilidade de um texto, compondo 200 métricas divididas em 14 categorias. Este estudo utilizou o NILC-Metrix para avaliar 2152 Trabalhos de Conclusão de Curso (TCCs) de instituições federais brasileiras, analisando métricas como sofisticação lexical, complexidade sintática e dispositivos de coesão. Os resultados revelaram diferenças significativas entre áreas, como por exemplo o fato de que os textos de Tecnologia da Informação (TI) apresentam menor complexidade em relação a trabalhos na área de Ciências Humanas e Sociais. Também observou-se que a variação da complexidade textual tende a acompanhar o tamanho do texto, sendo que textos com menos palavras por sentença e menor número de sentenças por parágrafo têm menor nível de complexidade textual, como nas áreas de TI e Engenharia.

Palavras-chave: Analítica de aprendizagem, Ensino da escrita; Analítica da escrita

Analysis of Undergraduate Theses using Natural Language Processing

Abstract: This work presents a study on the analysis of writing in undergraduate theses across different fields, considering metrics from computational linguistics. The NILC-Metrix system, an adaptation of the Coh-Metrix system, was used for this purpose. It employs linguistic resources to evaluate aspects such as the level of complexity and readability of a text, comprising 200 metrics divided into 14 categories. This study utilized NILC-Metrix to assess 2152 undergraduate theses from Brazilian federal institutions, analyzing metrics such as lexical sophistication, syntactic complexity, and cohesion devices. The results revealed significant differences between fields, such as the fact that theses in Information Technology (IT) exhibit lower complexity compared to those in the Humanities and Social Sciences. It was also observed that the variation in textual complexity tends to correlate with text length, with texts containing fewer words per sentence and a lower number of sentences per paragraph showing lower levels of textual complexity, as seen in IT and Engineering fields.

Keywords: Learning analytics, Writing instruction; Writing analytics

1. Introdução

Num mundo cada vez mais permeado pela tecnologia, a educação se encontra em constante processo de transformação. Nesse contexto, surge um campo de estudo específico chamado analítica da escrita, do inglês *writing analytics*, concentrado na mensuração e análise de textos para compreender os processos e produtos de escrita em contextos educacionais (Lang et al., 2019). A analítica da escrita pode ser empregada para (a) descrever as características de diferentes tipos de textos, comparando, por exemplo, a produção textual em diferentes áreas do conhecimento, estilos ou níveis de proficiência; (b) prever o desempenho de um aluno em uma tarefa de escrita futura, identificando padrões em seus textos anteriores e delineando um panorama de suas habilidades e desafios; (c) definir intervenções personalizadas, fornecendo feedback individualizado e orientando o aluno para o aprimoramento de suas habilidades de escrita.

Nesta pesquisa, exploramos o potencial descritivo da analítica da escrita, comparando métricas de escrita extraídas de trabalhos de conclusão de curso em diferentes áreas do conhecimento. Por meio dessa análise, buscamos identificar padrões e características que possam contribuir para uma melhor compreensão das disparidades na produção textual em diversas disciplinas. O restante deste artigo está estruturado da seguinte forma: a seção 2 apresenta o tema analítica da escrita, detalhando técnicas e ferramentas; a seção 3 descreve aspectos metodológicos do estudo desenvolvido na análise dos TCCs; a seção 4 apresenta Resultados e a última seção traz as conclusões do estudo e direcionamentos para trabalhos futuros.

2. A Analítica da Escrita

A Analítica da Escrita é uma subárea da Analítica da Aprendizagem que se concentra na coleta e análise de dados relacionados à produção textual (Lang et al., 2019). Seu objetivo é fornecer uma visão detalhada da habilidade de escrita dos estudantes e auxiliar no aprimoramento dessa habilidade. Utiliza métodos analíticos para avaliar diversos aspectos da escrita, como estrutura, coerência, vocabulário e gramática. Essa avaliação fornece feedback preciso aos alunos e educadores, permitindo a identificação de deficiências e padrões na aprendizagem da escrita. Também permite aos educadores monitorar o progresso de seus alunos ao longo do tempo, fornecendo informações relevantes para o desenvolvimento de intervenções pedagógicas personalizadas.

Já o Processamento de Linguagem Natural (PLN) é um campo que reúne linguística computacional, inteligência artificial e ciência da computação para possibilitar às máquinas compreenderem, interpretarem e manipularem a linguagem humana (Jurafsky e Martin, 2018). Desde os sistemas de recomendação que sugerem filmes e músicas, até os chatbots que conversam de forma natural, o PLN está no centro de muitas inovações tecnológicas (Hovy e Lavid, 2010). Trabalhos como os de Xia et. al. (2016) e Pires et. al. (2017) demonstram como o PLN pode ser utilizado para avaliar a legibilidade e a complexidade dos textos, direcionando-os para o nível de ensino e faixa etária adequados.

O Coh-Metrix é uma ferramenta para análise da escrita que é capaz de avaliar um texto em várias dimensões, incluindo coesão, sintaxe, complexidade, nível de leiturabilidade e outras características linguísticas (Crossley, et. al., 2011). Também pode ser usado para avaliar o nível de legibilidade e a compreensibilidade do texto. A partir de 2008, o Núcleo Interinstitucional de Linguística Computacional (USP) trabalhou no desenvolvimento de uma versão em português para funções disponíveis no Coh-Metrix. Assim, foi desenvolvido o NILC-Metrix, o qual fornece 200 métricas sobre o texto avaliado, divididas em 14 categorias: Medidas Descritivas, Simplicidade Textual, Coesão Referencial, Coesão Semântica, Medidas Psicolinguísticas, Diversidade Lexical, Conectivos, Léxico Temporal, Complexidade Sintática, Densidade de Padrões Sintáticos, Informações Morfossintáticas de Palavras, Informações Semânticas de Palavras, Frequência de Palavras e Índices de Leiturabilidade (Leal et. al., 2021). Este foi o sistema utilizado nas análises textuais apresentadas neste artigo.

A maioria dos estudos sobre o emprego do PLN para a análise e avaliação de textos se concentra principalmente no idioma inglês, o que requer adaptações para o contexto da língua portuguesa. Dentre as várias aplicações dessas tecnologias, a avaliação do nível de legibilidade e complexidade dos textos é especialmente relevante, com o objetivo de direcioná-los para o nível escolar ou de compreensão apropriado (Pires et al., 2017). A simplificação do conteúdo textual de acordo com a etapa escolar dos estudantes também é uma prática comum, buscando garantir que o material seja acessível e interessante para os leitores (Crossley et al., 2011; Aluísio et al., 2010). No entanto, há uma escassez de estudos que se concentram no uso de métricas textuais para avaliar a qualidade textual, especialmente

no contexto da língua portuguesa. Meira et al. (2023) faz um apanhado sobre vários trabalhos de pesquisa da última década sobre o emprego da linguística computacional para análise de textos na língua portuguesa, apresentando em seguida um estudo sobre a relação entre diversas métricas de análise textual, obtidas através do processamento de linguagem natural, e as avaliações de projetos de pesquisa de graduação em língua portuguesa. Os resultados da pesquisa apontaram para uma tendência de atribuição de notas mais altas para trabalhos que apresentam maior sofisticação lexical e complexidade sintática. São achados que reiteram resultados de pesquisas anteriores realizadas em língua inglesa (McNamara & Graesser, 2012), mas que são importantes no contexto da pesquisa em língua portuguesa.

Nos últimos anos, muitos estudos têm explorado as diferenças nos estilos de escrita acadêmica entre disciplinas. Dong (2023) descobriu que os estilos de escrita estão se tornando mais complexos e informativos, com variações significativas em características linguísticas entre disciplinas. Alluqmani (2018) utilizou ciência de dados para identificar estilos de escrita distintos em física, astrofísica, matemática e ciência da computação, com mudanças ao longo do tempo. Kostrova (2015) destacou a influência da cultura e tradição na escrita acadêmica, com diferenças nos objetivos comunicativos, volume de material e responsabilidade do autor.

No trabalho aqui apresentado busca-se avançar nesta mesma perspectiva, investigando como os estudantes de diferentes áreas compõe de modo distinto seus projetos de final de curso (TCCs), no que diz respeito a um dado conjunto de métricas disponibilizado pelo NILC-Metrix.

3. Metodologia

O corpus de textos utilizados neste estudo foi composto por trabalhos de conclusão de curso (TCCs) de instituições federais de ensino brasileiras, instituições que mantêm acesso público aos trabalhos já defendidos, concluídos e aprovados.

Foram obtidos cerca de 9000 trabalhos de conclusão de curso, divididos entre 8 instituições, sendo estas, universidades e institutos federais, de cursos de diversas áreas, como Ciências Biológicas e da Saúde (CBS), Ciências Humanas e Sociais (CHS), Ciências Exatas e da Terra (CET), Engenharias (ENG), Gestão e Negócios (GN) e Tecnologia da Informação (TI). Deste universo de trabalhos, aplicou-se um filtro para exclusão de trabalhos anteriores ao ano de 2015, restando 5801. A tabela 1 mostra o número de trabalhos encontrados por ano.

Tabela 1 - Número de trabalhos encontrados por ano de registro

Ano	Nº de trabalhos
2023	675
2022	996
2021	694
2020	260
2019	697
2018	869
2017	517
2016	587
2015	506

Após os acessos para download dos arquivos, foram obtidos 5098 arquivos, a diferença entre o total de arquivos obtidos e o total de arquivos listados nas bases de dados é resultado de que nem todos os arquivos listados nas bases das instituições encontram-se disponíveis para visualização ou download. Após a obtenção dos arquivos, iniciou-se o processo de identificação e categorização, com isto, foram removidos 2243 arquivos que correspondiam a trabalhos de conclusão de curso de nível técnico ou pós-graduação, permanecendo 2855 trabalhos a serem limpos e processados.

Como critério final para elencar os trabalhos foram excluídos os trabalhos escritos em outras línguas que não o português brasileiro, e trabalhos com extração textual menor que 70% do total de palavras existentes no arquivo finalizando a construção do corpus com 2152 arquivos. O percentual representativo para cada área ficou em 10% (CBS), 30% (CHS), 12% (CET), 10% (ENG), 20% (GN) e 18% (TI). Apesar de não haver homogeneidade na distribuição de trabalhos por área, as áreas com um menor número de trabalhos ainda permanecem com um número significativo.

A extração textual foi efetuada em linguagem de programação Python e biblioteca *python-docx*, sendo que nesta extração não foram considerados pontos como tabelas, já excluídas, referências bibliográficas, apêndices, anexos e demais componentes extras ao texto de um trabalho de conclusão de curso.

Para o conjunto de trabalhos elencados, utilizou-se o NILC-Metrix para calcular o um conjunto de métricas de acordo com categorias estabelecidas na literatura da área, como nos trabalhos de Crossley e McNamara (2011) e McNamara e Graesser (2012), sendo eles: Sofisticação Léxica (Diversidade léxica, Frequência de palavras, Nível de concretude, Polissemia das palavras); Complexidade Sintática (Palavras antes do verbo principal, Modificadores por sintagma nominal); Dispositivos de Coesão (Conectivos lógicos, Sobreposição de palavras de conteúdo); Tamanho e Extensão (Palavras por sentença, Sentenças por parágrafo). Antes do envio do texto para processamento, removeram-se as páginas indesejadas, como capa, folha de rosto, sumário, lista de tabelas e figuras, referencial teórico, apêndices e anexos. A seção de resultados, a seguir, detalha os valores obtidos nesses cálculos.

4. Resultados

O corpus utilizado neste estudo foi composto apenas por trabalhos já defendidos e aprovados em instituições federais de ensino. Assim, os resultados aqui apresentados giram em torno do objetivo de construção de parâmetros e balizas para as áreas de conhecimento analisadas. Os resultados estão dispostos na tabela 2.

Tabela 2 - Quadro de resultados para o corpus TCCs

	Sofisticação Léxica				Complexidade Sintática		Dispositivos de Coesão		Tamanho e Extensão	
	1	2	3	4	1	2	1	2	1	2
MG	0.719	475163	3.968	4.156	5.852	6.226	0.041	1.453	21,33	2.018
DP	0.037	158513	0.186	0.865	0.746	0.974	0.012	0.723	5.154	0.548
CBS	0.746	492061	4.094	3.989	6.168	6.396	0.038	1.341	21.70	2.338
DP	0.021	164819	0.136	0.606	0.735	0.960	0.009	0.628	5.531	0.585

CHS	0.736	495444	3.963	4.312	6.083	6.595	0.044	1.305	23.94	2.377
DP	0.020	142145	0.111	0.406	0.732	0.875	0.007	0.504	4.030	0.619
CET	0.716	471879	3.988	4.365	5.654	5.918	0.044	1.274	20.32	1.867
DP	0.036	135498	0.128	0.403	0.611	0.793	0.007	0.425	4.784	0.385
ENG	0.704	466730	4.041	4.223	5.976	5.704	0.042	1.331	18.56	1.650
DP	0.030	158770	0.139	0.340	0.717	0.865	0.006	0.483	4.519	0.341
GN	0.717	507464	3.904	4.446	5.995	6.323	0.045	1.543	23.00	1.768
DP	0.036	190418	0.206	0.451	0.622	0.776	0.007	0.502	3.847	0.333
TI	0.693	394872	4.009	3.373	5.218	5.873	0.030	1.928	15.74	1.891
DP	0.052	126399	0.284	1.672	0.646	1.202	0.019	1.225	3.454	0.382

Fonte: do autores

Na tabela 2, entende-se MG como a Média Geral, ou seja, a média para todos os trabalhos, DP como desvio padrão e as áreas como CBS para Ciências Biológicas e da Saúde, CHS para Ciências Humanas e Sociais, CET para Ciências Exatas e da Terra, ENG para Engenharias, GN para Gestão e Negócios e TI para Tecnologia da Informação.

Para a leitura da tabela 2, faz-se necessário compreender suas linhas e colunas. Cada coluna corresponde a uma métrica de avaliação da qualidade textual, divididas nas 4 categorias já apresentadas na tabela 2. Cada dupla de linha, por sua vez, representa o valor representante da média dos trabalhos da referida área, acompanhando sucessivamente abaixo pelo valor do desvio padrão da área em questão. Exemplificando a leitura, a coluna 3 da linha 3 apresenta o valor médio obtido por trabalhos da área de Ciências Biológicas e da Saúde (CBS) para a métrica Nível de Concretude, sendo este de 4.094, ou seja, a média do valor de concretude das palavras de trabalhos desta área é de aproximadamente 4,09 pontos de concretude.

A célula ligeiramente abaixo, coluna 3, linha 4, apresenta o valor do desvio padrão dos trabalhos da área CBS para esta mesma métrica de Nível de Concretude. O valor apresentado de 0.136 significa que a cada 0,13 pontos, acima ou abaixo de 4,09, tem-se um desvio padrão da média da área. Assim entende-se que um trabalho que obtenha, por exemplo, pontuação de 5 pontos de Nível de Concretude, estará cerca de 8 desvios padrão acima da média, o que representa uma divergência de alerta para o professor.

Para uma melhor visualização das diferenças foi construído a figura 1 que contém o gráfico demonstrativo da média de valores z-scores obtidos para cada área quando comparados a média geral dos trabalhos.

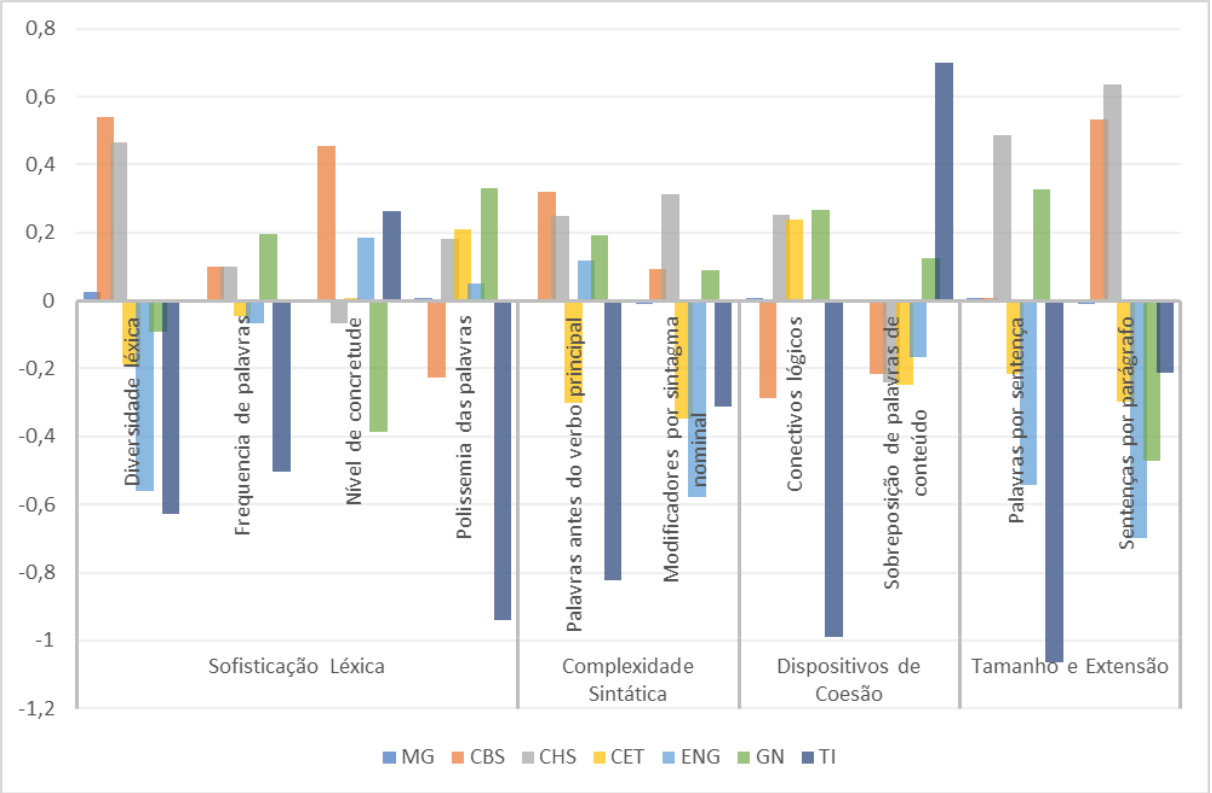

Figura 1 - Gráfico de disposição de z-scores por área
Fonte: do autores

A leitura da figura 1 apresenta o posicionamento de cada área de conhecimento que compõem o corpus de TCCs normalizadas via z-score da média geral dos trabalhos. O primeiro ponto a salientar consiste na escala do gráfico, a qual varia de -1,2 desvios padrão à 0,8 desvios padrão da média geral. Este fator por si só já demonstra certa homogeneidade do corpus, mesmo sendo composto por TCCs de seis áreas do conhecimento, divididos em cerca de 34 cursos diferentes, a variação do valor z-score não ultrapassa 1,2 desvios padrão, para mais ou para menos.

Por outro lado, ao analisar os pormenores da figura 1, pode-se notar as diferenças individuais de cada área. A maior diferença notada está na categoria Tamanho e Extensão de trabalhos da área de Tecnologia da Informação, os quais obtiveram média de -1,06 desvios padrão abaixo da média do corpus de TCCs, ou seja, isto significa que os trabalhos desta área tendem a possuir cerca de 5.15 palavras por sentença a menos que a média geral do corpus. Esta variação pode ser lida na tabela 5, linha 2 (DP), coluna 9 (Tamanho e Extensão 1).

Considerando que as métricas refletem a complexidade textual que o texto analisado demanda do leitor, é interessante notar como as métricas classificam esta complexidade quanto às áreas dos trabalhos. Nota-se na figura 1 que a área de TI é a que possui maior número de métricas com valores z-score negativos, ou seja, uma quantidade de desvios padrão abaixo da média, isto diz que, teoricamente, os textos contidos em trabalhos desta área têm menor nível de complexidade.

Já as áreas que apresentam maior complexidade de leitura textual são as áreas de CHS e CBS, com métricas apontando maiores quantidades de desvios padrão acima da média dos trabalhos do corpus. Também é possível relacionar através da figura 1, que a variação da complexidade textual tende a acompanhar o tamanho do texto, sendo que textos com menos palavras por sentença e menor número de sentenças por parágrafo foram considerados menos

complexos, como nas áreas de TI e ENG. O inverso, para as áreas de CHS e CBS também se mostra coerente.

Corroborando com as afirmações de Alluqmani (2018) e Dong (2023), onde são apresentadas diferenças significativas nas características da escrita entre disciplinas, os resultados apresentados neste trabalho apontam para a importância da divisão de comparativos também por área de conhecimento, ficando ainda mais evidente que comparar um trabalho da área de Ciências Humanas e Sociais (CHS) com as métricas de trabalhos de TI causa uma distorção nos resultados obtidos.

Para a área de TI, nota-se o predomínio negativo, com relação à média, mas métricas dos trabalhos, sendo possível notar que as métricas de Diversidade léxica (-0,6), Frequência de palavras (-0,5), Polissemia das palavras (-0,9), Palavras antes do verbo principal (-0,8), Conectivos lógicos (-0,9) e Palavras por sentença (-1,0) representam uma diferença significativa quando comparados aos valores obtidos por trabalhos da área CHS, sendo que os respectivos valores apresentados pela figura 1 são 0.4, 0.1, 0.1, 0.2, 0.2 e 0.4. Esta diferença entre valores das duas áreas, caso não existissem, poderia gerar parametrizações equivocadas, resultando em feedbacks indevidos ou incorretos.

Como resultado, este corpus de TCCs permitiu a categorização dos feedbacks gerados através da área do curso do aluno. Assim, para cada caso de avaliação, é possível considerar as particularidades de cada área, considerando o respectivo valor de desvio padrão. Outra utilidade consiste na possibilidade de utilização de um sistema para trabalhos de TCCs, ou seja, para que orientadores façam uso do sistema como ferramenta de auxílio na avaliação dos trabalhos de seus orientandos.

A construção dos corpus e a parametrização das Métricas de Complexidade Textual permitiram a construção de um gráfico comparativo entre eles. A figura 2 abaixo apresenta este gráfico.

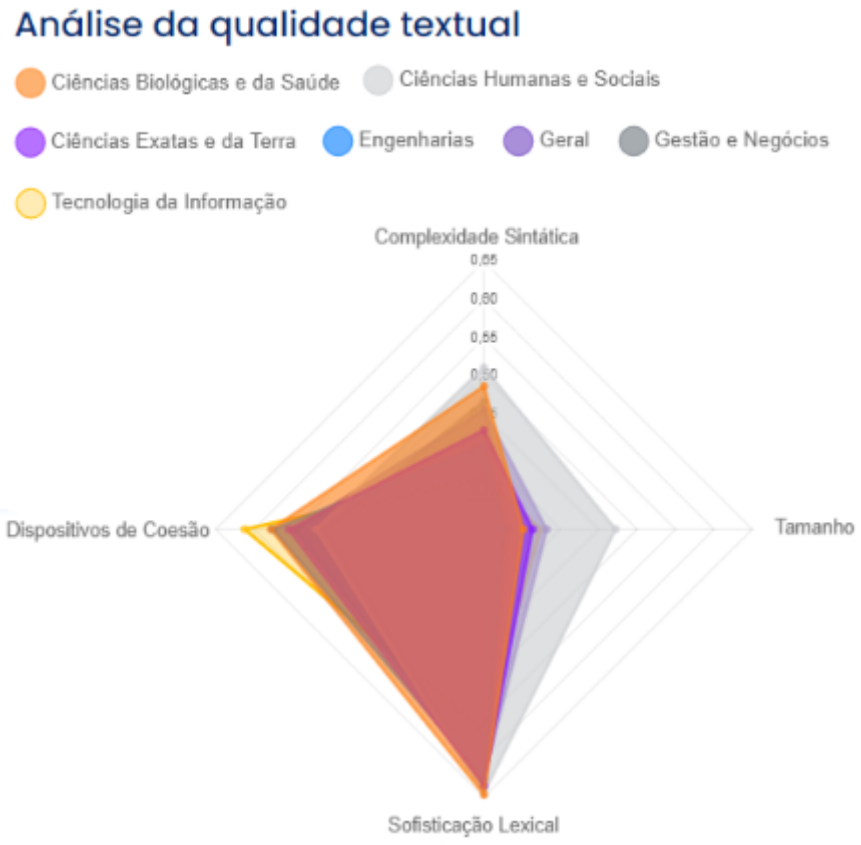

Figura 2 - Gráfico de complexidade textual

Fonte: Autores

Para a figura 2, foram plotados os comparativos das áreas de conhecimento do corpus de TCCs. A escala deste gráfico varia entre 0 e 1.

Para o corpus de TCCs, nota-se que os trabalhos das diversas áreas possuem proximidade no valor de Sofisticação Lexical, sendo o mais alto no valor de 0,647 pontos, obtido pela área de Ciências Biológicas e da Saúde, a média dos cursos ficou em 0,625 pontos e a pontuação mais baixa foi de 0,564 para a área de Tecnologia da Informação.

Já para Dispositivos de Coesão, ve-se que trabalhos da área de Tecnologia da informação obtiveram valores maiores, tendo a área obtido média de 0,612 pontos, seguidos por Ciências Biológicas e da Saúde juntamente com trabalhos da área de Engenharias a área que obteve menor pontuação na escala de Dispositivos de Coesão foi Gestão e Negócios com 0,509 pontos. A média de todas as áreas ficou em 0,547 pontos.

O gráfico da figura 2 apresenta, para a métrica de Complexidade Sintática, valores de 0,512 a 0,369, sendo a área de Ciências Sociais e Humanas a que obteve maior pontuação e Tecnologia da Informação a área com menor pontuação. A média entre as áreas ficou em 0,456 pontos.

No quesito Tamanho, o gráfico mostra a área de Ciências Sociais e Humanas como contendo os trabalhos mais extensos, obtendo 0,472 pontos, a média geral das áreas de conhecimento obteve 0,382 pontos e a área de Engenharias obteve a menor pontuação em extensão dos trabalhos, 0,327 pontos.

Este gráfico e sua pontuação foram obtidos ao se normalizar e parametrizar as métricas que compõem cada ponta do gráfico, por exemplo, o valor de Dispositivos de Coesão foi obtido com a normalização de escala e parametrização dos valores obtidos para as métricas de Conectivos Lógicos e Sobreposição de Palavras de Conteúdo, já para a valoração de Sofisticação Lexical, foram utilizados os valores normalizados das métricas de Diversidade Léxica, Frequência de Palavras, Nível de Concretude das Palavras e Polissemia das Palavras. Para a pontuação de Complexidade Sintática, foram utilizadas as métricas de Palavras Antes do Verbo Principal e Modificadores por Sintagma Nominal, por fim, a pontuação de Tamanho se deu com a utilização de valores de palavras do texto, sentenças por palavras e sentenças por frase. Esta segmentação está detalhada na tabela 2.

A análise dos dados deste corpus se fez importante no sentido de demonstrar a variação das métricas para cada tipo de área de ensino, permitindo assim identificar peculiaridades na forma de escrita e composição de trabalhos de conclusão de curso em diferentes áreas.

5. Conclusão

Este artigo apresentou um estudo sobre a análise da escrita de trabalhos de conclusão de curso, abordando diversas áreas por meio do uso de métricas da linguística computacional. Ao empregar o sistema NILC-Metrix, foram avaliados 2152 TCCs de instituições federais brasileiras. Os resultados destacaram algumas diferenças entre as áreas, evidenciando, por exemplo, a menor complexidade dos textos de Tecnologia da Informação em comparação com os trabalhos das Ciências Humanas e Sociais. Além disso, foi observada uma relação direta entre a complexidade textual e o tamanho do texto, onde textos com menor número de palavras por sentença e de sentenças por parágrafo foram associados a uma menor complexidade, especialmente nas áreas de TI e Engenharia. Essas descobertas contribuem para o entendimento da produção textual em diferentes disciplinas e preparam o cenário para o desenvolvimento de modelos de avaliação textual que sejam específicos para cada uma.

Como sugestão para pesquisas futuras, propõe-se o desenvolvimento de ferramentas de visualização de dados que possam ampliar o uso das métricas de análise textual com

objetivos avaliativos. Outra sugestão para trabalho futuro seria investigar como as métricas da linguística computacional podem ser combinadas com abordagens de aprendizado de máquina para desenvolver sistemas automatizados de avaliação de escrita mais precisos e eficazes. Essa pesquisa poderia explorar o treinamento de algoritmos de aprendizado de máquina usando dados de escrita previamente avaliados para identificar padrões e fornecer feedback detalhado aos alunos e educadores.

Agradecimentos

A pesquisa apresentada nesta publicação foi realizada com o apoio da CAPES/PROEX, Brasil, e também da Agência Nacional de Pesquisa e Inovação (ANII) no Uruguai, sob o código FSED_2_2021_1_169701.

Referências

ALLUQMANI, Amnah ; SHAMIR, Lior. Writing styles in different scientific disciplines: a data science approach. **Scientometrics**, v. 115, n. 2, p. 1071–1085, 2018.

ALUISIO, Sandra et al. Readability Assessment for Text Simplification. In: NAACL HLT, 2010, Los Angeles, CA. **Fifth Workshop on Innovative Use of NLP for Building Educational Applications.** Los Angeles, CA: Association for Computational Linguistics, 2010. p. 1–9. Disponível em: https://aclanthology.org/W10-1001.pdf. Acesso em: 29 out. 2023.

CROSSLEY, Scott A.; ALLEN, David; MCNAMARA, Danielle S. **Text simplification and comprehensible input:** A case for an intuitive approach. Language Teaching Research, [s. l.], v. 16, n. 1, p. 89–108, 2011.

CROSSLEY, Scott A.; MCNAMARA, Danielle S. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. **International Journal of Continuing Engineering Education and Life-Long Learning,** [s. l.], v. 21, n. 2/3, p. 170, 2011. Disponível em: https://doi.org/10.1504/IJCEELL.2011.040197. Acesso em: 29 out. 2023.

DONG, Shuyi; MAO, Jin ; PEI, Lei. Comparing the Writing Styles of Multiple Disciplines: A Large‐Scale Quantitative Analysis. **Proceedings of the Association for Information Science and Technology,** v. 60, n. 1, p. 941–943, 2023.

HOVY, Eduard ; LAVID, Julia. Towards a “Science” of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. **INTERNATIONAL JOURNAL OF TRANSLATION,** v. 22, n. 1, 2010. Acesso em 20 de abril 2024. Disponível em: https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf.

JURAFSKY, Daniel ; MARTIN, James. **Speech and Language Processing:** An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Third Edition draft. [s.l.: s.n.], 2020. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf. Acesso em 20 de abril 2024.

KOSTROVA, Olga ; KULINICH, Marina. Text Genre Academic Writing: Intercultural View. **Procedia - Social and Behavioral Sciences,** v. 206, p. 85–89, 2015.

LANG, Susan; AULL, Laura; MARCELLINO, William. A Taxonomy for Writing Analytics. The **Journal of Writing Analytics,** [s. l.], v. 3, n. 1, p. 13–37, 2019. https://doi.org/10.37514/JWA-J.2019.3.1.03. Acesso em: 29 out. 2023.

LEAL, Sidney Evaldo et al. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. **Language Resources and Evaluation,** [s. l.], v. 58, p. 73–110, 2023. https://doi.org/10.1007/s10579-023-09693-w https://doi.org/10.1007/s10579-023-09693-w.

MCNAMARA, Danielle S; GRAESSER, Arthur C. Coh-Metrix An Automated Tool for Theoretical and Applied Natural Language . In: MCCARTHY, P; BOONTHUM-DENECKE, C (org.). **Applied Natural Language Processing: Identification, Investigation and Resolution.** IGI Global eBooks. [S. l.: s. n.], 2012. p. 188–205. E-book. Disponível em: https://doi.org/10.4018/978-1-60960-741-8.ch011.

MEIRA, Ricardo R.; WEIAND, Augusto.; REATEGUI, Eliseo.; BIGOLIN, Márcio.; MOTZ, Regina. A Analítica da Escrita para Identificação de Indicadores de Qualidade Textual. Revista Novas Tecnologias na Educação, Porto Alegre, v. 21, n. 2, p. 342–351, 2023. Disponível em: https://doi.org/10.22456/1679-1916.137756.

PIRES, Carla; CAVACO, Afonso; VIGÁRIO, Marina. Towards the Definition of Linguistic Metrics for Evaluating Text Readability. **Journal of Quantitative Linguistics,** [s. l.], v. 24, n. 4, p. 319–349, 2017. Disponível em: https://doi.org/10.1080/09296174.2017.1311448.

XIA, Menglin; EKATERINA KOCHMAR; BRISCOE, Ted. Text Readability Assessment for Second Language Learners. In: BEA@ACL, 2016, San Diego, CA. **11th Workshop on Innovative Use of NLP for Building Educational Applications.** [S. l.]: Association for Computational Linguistics, 2016. https://doi.org/10.48550/arXiv.1906.07580

Data de submissão: 29/04/2024 | Data de revisão: 25/06/2024 | Data de aceite: 28/06/2024 | Data de publicação: 31/07/2024