Complexidade textual em narrativas orais produzidas por informantes de diferentes níveis de escolaridade

Juliano Desiderato Antonio[1]

[1]Departamento de Teorias Linguísticas e Literárias – Universidade Estadual de Maringá (UEM)
Maringá – PR – Brasil

jdantonio@uem.br

Abstract. *In this paper, a corpus formed by twenty oral narratives (ten produced by elementary school students and ten produced by undergraduate students) is analyzed with the aim of verifying whether the textual complexity of the narratives increases as the level of education increases. The tool used to analyze the textual complexity of the narratives in the corpus is the computational system NILC-Metrix, which employs two hundred metrics for this purpose. Eleven metrics were chosen which demonstrate that the narratives produced by the undergraduate students present higher textual complexity than the narratives produced by elementary school students.*

Resumo. *Neste trabalho, analisa-se um córpus formado por vinte narrativas orais (dez produzidas por estudantes do ensino fundamental e dez produzidas por estudantes de curso superior) com o objetivo de se verificar se a complexidade textual das narrativas aumenta conforme aumenta o nível de escolaridade. A ferramenta utilizada para analisar a complexidade textual das narrativas do córpus é o sistema computacional NILC-Metrix, que emprega duzentas métricas para esse fim. Escolheram-se onze métricas que demonstram que as narrativas do córpus produzidas pelos alunos de curso superior apresentam maior complexidade textual do que as narrativas produzidas pelos alunos de ensino fundamental.*

Considerações iniciais

Com as recentes evoluções no campo do Processamento de Linguagem Natural (PLN), cada vez mais tarefas linguisticamente complexas vêm sendo realizadas por ferramentas computacionais. Alguns exemplos são o reconhecimento da fala humana, a tradução automática, a sumarização automática, a análise de sentimentos, dentre outros. Uma outra possibilidade interessante que vem se desenvolvendo, segundo Santucci et al. (2020), é a possibilidade de se analisar automaticamente a complexidade de textos do ponto de vista linguístico. Branco et al. (2014a), por exemplo, apresentam um sistema

que classifica automaticamente textos produzidos em Língua Portuguesa com base no Quadro Europeu Comum de Referência para Línguas. A classificação tem cinco níveis de dificuldade: A1 (mais fácil), A2, B1, B2 e C1 (mais difícil) e utiliza os critérios do Instituto Camões para certificação de proficiência. Os critérios utilizados para classificação, segundo Branco et al. (2014b) são leiturabilidade, densidade lexical, quantidade de sílabas por palavra, quantidade de palavras por sentença. Ao reduzirem um grande número de propriedades textuais a um número menor de dimensões de complexidade textual [Goldman e Lee 2014], as ferramentas podem também auxiliar na seleção de textos adequados para diferentes níveis de aprendizagem [Sheehan, Flor e Napolitano 2013; McNamara et al. 2014]. Evers (2018) descreve padrões léxico-sintáticos de redações submetidas ao vestibular da Universidade Federal do Rio Grande do Sul. A pesquisadora utiliza recursos e ferramentas dos Estudos do Léxico, da Linguística Textual, da Linguística de Córpus e da Linguística Computacional para identificar padrões lexicais e sintáticos correspondentes a três faixas de desempenho. Uma das ferramentas utilizadas pela pesquisadora trata especificamente da complexidade textual.

Jensen (2009) alerta para o fato de que dificuldade e complexidade textual são dois conceitos distintos que não devem ser confundidos. Enquanto a dificuldade é subjetiva e pode variar de leitor para leitor, a complexidade é mais objetiva e pode ser calculada a partir de critérios factuais, como os índices de leiturabilidade, a frequência das palavras (quanto mais comum é uma palavra menos esforço cognitivo é dispensado no processamento daquela palavra), a não literalidade (metáforas, metonímias e expressões idiomáticas podem afetar a complexidade textual pelo fato de exigirem maior esforço cognitivo para serem processadas).

Neste trabalho, investiga-se a complexidade textual de narrativas orais produzidas por informantes com diferentes níveis de escolaridade com a finalidade de se verificar se a complexidade textual das narrativas aumenta conforme aumenta o nível de escolaridade. O córpus é composto por dez narrativas produzidas por alunos do sexto ano do ensino fundamental e por dez narrativas produzidas por alunos de curso superior (Comunicação Social). A descrição dessas possíveis diferenças pode fornecer subsídios para que se descrevam os recursos empregados nos textos mais complexos. E esses recursos poderão ser utilizados em sala de aula por professores para auxiliarem seus alunos a produzirem textos narrativos mais complexos.

Metodologia

Quando da coleta dos dados, adotaram-se alguns critérios para que se evitasse ao máximo o risco de diferenças nos resultados causados por discrepâncias no córpus. Para que os textos de todos os informantes fossem sobre um mesmo assunto e fossem semelhantes em aspectos como extensão, conteúdo, etc, decidiu-se que a coleta dos

dados seria feita a partir da exibição de um vídeo com uma história que seria recontada pelos sujeitos da pesquisa. A opção pela narrativa proveio do fato de que, para a produção desse tipo de texto, o filme serviria como um *script* a ser seguido pelos informantes, o que permitiria a obtenção de um córpus bastante homogêneo. Para se evitar que houvesse influência das falas do narrador ou de personagens sobre a maneira como os informantes formulariam linguisticamente a história, a solução foi procurar um filme mudo, cuja sequência de cenas fosse suficiente para a compreensão do enredo. O vídeo escolhido foi "O pavão misterioso", que se baseia em uma história do folclore nordestino de mesmo nome e que tem como personagens bonecos que representam seres humanos. Logo após assistirem ao filme, os informantes contaram a história oralmente, que foi gravada em fitas K-7.

A primeira parte do córpus foi coletada em 1996, com alunos do primeiro ano do curso de Comunicação Social de uma universidade situada no Norte do Paraná [Antonio 1998]. As demais narrativas foram coletadas em 2001, em uma escola estadual também situada em um município do Norte do Paraná. Os informantes eram alunos do sexto ano do ensino fundamental [Antonio 2004].

A ferramenta utilizada para analisar a complexidade textual das narrativas do córpus é o sistema computacional NILC-Metrix [Leal et al. 2022]. Esse sistema utiliza duzentas métricas propostas em estudos de Linguística Textual, Psicolinguística, Linguística Cognitiva e Linguística Computacional para investigar a complexidade textual no português brasileiro. Pode ser utilizado tanto com textos orais quanto com textos escritos.

As duzentas métricas são agrupadas em quatorze categorias [Leal et al. 2022]: índices descritivos, métricas de simplicidade textual, coesão referencial, coesão semântica, medidas psicolinguísticas, diversidade lexical, conectivos, léxico temporal, complexidade sintática, densidade do padrão sintático, informação morfossintática das palavras, frequência de palavras, fórmulas de leiturabilidade.

Por motivo de limitação de espaço, serão analisadas apenas onze métricas, as quais apresentaram maior possibilidade de caracterizar as diferenças de complexidade nas narrativas dos dois grupos de informantes. Para a obtenção dos resultados, calculou-se a média de cada métrica dos textos correspondentes a cada nível de escolaridade.

Resultados

As métricas de complexidade sintática se mostraram muito reveladoras no que diz respeito às diferenças de complexidade entre os textos produzidos pelos dois grupos de informantes. Os resultados de três dessas métricas são apresentados na tabela 1.

Tabela 1. Resultados de três métricas de complexidade sintática

	Ensino Fundamental	Ensino Superior

Distância na árvore de dependências	33,990112	52,351342
Proporção de orações subordinadas pela quantidade de orações do texto	0,231913	39,144816
Proporção de orações na voz passiva analítica em relação à quantidade de orações do texto	0,019933	0,031742

Como pode ser observado na tabela 1, as métricas das narrativas do córpus produzidas pelos alunos de ensino superior apresentam valores mais altos, indicando maior complexidade textual. Na métrica distância na árvore de dependências, existe uma relação entre a distância entre palavras e tempo de processamento. Quanto maior a distância, mais se exige da memória do destinatário do texto. Dessa forma, maior distância na árvore de dependências resulta em maior complexidade [Leal et al. 2021; Santucci et al. 2020]. No exemplo da figura 1, encontrado na narrativa ES9[1], pode-se observar uma grande distância entre a raiz da árvore (verbo "sair") e a oração "parece ser alguma coisa assim".

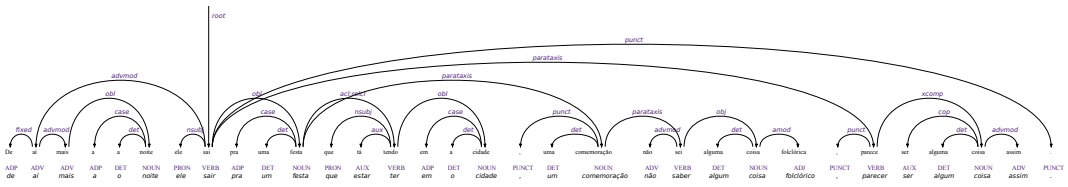

Figura 1. Distância na árvore de dependências

Pelo fato de as orações subordinadas serem estruturas mais complexas e que exigem maior esforço de processamento, a maior proporção desse tipo de construção indica maior complexidade textual [Leal et al. 2021; McNamara et al. 2014; Santucci et al. 2020]. No exemplo (1), retirado da narrativa ES1, encontram-se duas orações adjetivas ("que pode ser muito observada" e "que é Viva") e uma oração completiva ("que ele acabou de presenciar"). No exemplo (2), retirado da narrativa ES3, encontram-se duas orações adverbiais (temporal: "quando já mais nem esperava encontrá-la"; causal: "porque nem mesmo sabia onde procurá-la") e uma oração completiva ("onde procurá-la").

(1) Uma antítese *que pode ser muito observada* seria o contraste entre a morte *que ele acabou de presenciar* e o nome do hotel, *que é Viva*.

(2) Então, *quando já mais nem esperava encontrá-la*, *porque nem mesmo sabia onde*

[1] ES: ensino superior. EF: ensino fundamental.

procurá-la, (…) encontrou a linda moça dos cachos dourados.

Uma maior proporção de orações na voz passiva analítica também indica maior complexidade textual, pois se trata de uma estrutura que as crianças adquirem mais tardiamente [Leal et al. 2021; McNamara et al. 2014]. No exemplo (3) a seguir, retirado da narrativa ES1, observam-se quatro construções passivas analíticas.

(3)
Ele é impedido de se aproximar dela por um senhor encarado e carrancudo,
e *ela é retirada* dele,
impedida de chegar perto dele por muitas pessoas
é trancafiada no quarto.

As métricas de diversidade lexical também apontam para uma maior complexidade das narrativas produzidas pelos alunos de curso superior. Na tabela 2, apresentam-se os resultados de duas dessas métricas.

Tabela 2. Resultados de duas métricas de diversidade lexical

	Ensino Fundamental	Ensino Superior
Proporção de *types* de substantivos em relação à quantidade de *tokens* de substantivos no texto	0,382287	0,56517
Proporção de *types* de verbos em relação à quantidade de *tokens* de verbos no texto	2,366978	250,618799

No caso das duas métricas, quanto maior a proporção de *types* em relação à quantidade de *tokens*, maior complexidade, ou seja, o produtor do texto demonstra conhecer uma maior diversidade de itens lexicais dessas duas classes de palavras. No caso dos substantivos, uma menor proporção de *types* indicaria menor complexidade textual pelo fato de a repetição de substantivos ser uma das formas mais simples de se construir uma cadeia de referência [Leal et al. 2021], como pode ser observado no exemplo (4), retirado da narrativa EF1, em que o produtor do texto, um aluno do ensino fundamental, repete os substantivos “homem”, “mulher” e “menina” para retomar os referentes anaforicamente.

(4)
A história começa assim: é de *um homem* que chega numa cidade. (...) Chega lá daí de caravela. Chegando perto do hotel, vê um velório de *um homem*. Daí ele se apaixona pela *mulher do homem que morreu*. Aí ele chega perto da *menina*, tinha uma festa lá. Aí chega perto da *menina*, *o pai da menina* chega junto com os segurança dele, daí ele não queria deixar ele ficar perto da *menina*.

Algumas métricas de informações morfossintáticas e de informações semânticas de palavras também demonstram maior complexidade nas narrativas dos informantes de

curso superior, como pode ser observado na tabela 2.

Tabela 3. Resultados de duas métricas de informações morfossintáticas e de duas métricas de informações semânticas de palavras

		Ensino Fundamental	Ensino Superior
Informações morfossintáticas	Proporção de pronomes relativos em relação à quantidade de pronomes do texto	0,243944	0,891063
	Proporção de verbos em relação à quantidade de palavras do texto	0,696373	1,439126
Informações semânticas	Proporção de nomes próprios em relação à quantidade de palavras do texto	0,021081	1,194962
	Proporção de substantivos abstratos em relação à quantidade de palavras do texto	0,196411	1,469733

Em se tratando das informações morfossintáticas, segundo Leal et al. (2021) e McNamara et al. (2014), a frequência mais alta de pronomes relativos indica maior complexidade pelo fato de esses pronomes introduzirem orações adjetivas, que elaboram o conteúdo de um sintagma nominal. No exemplo (5), encontrado na narrativa ES9, produzida por aluno de curso superior, o pronome relativo "que" retoma e elabora, na primeira ocorrência, o sintagma nominal "um marinheiro"; na segunda ocorrência, o sintagma nominal "uma fotografia no jornal"; na terceira ocorrência, o sintagma nominal "uma pessoa".

(5) A história é de um não sei parece ser um marinheiro *que* chega de navio numa cidade e ele se hospeda num hotel, mas antes ele vê uma fotografia no jornal *que* parece ser de uma mulher, parece ser de uma pessoa *que* ele tá procurando.

Na métrica seguinte, a maior proporção de verbos plenos (não se consideram os verbos auxiliares na contagem) também indica maior complexidade pelo fato de os verbos constituírem orações.

No que diz respeito às informações semânticas, nomear entidades demanda mais

memória, motivo pelo qual uma maior proporção de substantivos próprios indica maior complexidade [Feng et al. 2010]. No exemplo (4), aqui retomado para facilitar a visualização, o produtor do texto, um aluno do ensino fundamental, utiliza substantivos comuns para nomear os personagens como "homem", "mulher", "menina", "pai". Já no exemplo (6), encontrado na narrativa ES10, o informante, um aluno de curso superior, utilizou nomes próprios "João" e "Maria" para designar os referentes.

(4) A história começa assim: é de *um homem* que chega numa cidade. (...) Chega lá daí de caravela. Chegando perto do hotel, vê um velório de *um homem*. Daí ele se apaixona pela *mulher do homem que morreu*. Aí ele chega perto da *menina*, tinha uma festa lá. Aí chega perto da *menina*, *o pai da menina* chega junto com os segurança dele, daí ele não queria deixar ele ficar perto da *menina*.

(6) *João* voltou para sua cidade natal, num pequeno vilarejo nordestino. (…) Quando decidiu voltar para o hotel, *João* encontrou uma moça muito bonita chamada *Maria* e por ela ele se apaixonou.

Em relação à outra métrica, o processamento de substantivos abstratos é mais trabalhoso do que o de substantivos concretos, motivo pelo qual uma maior proporção de substantivos abstratos indica maior complexidade [Leal et al. 2021].

Por fim, todos os índices de leiturabilidade demonstraram a maior complexidade das narrativas dos informantes de ensino superior. Ponomarenko e Evers (2022, p. 42) definem leiturabilidade como "potencial facilidade ou dificuldade de leitura de um texto", levando em conta não apenas fatores linguísticos mas também o perfil do leitor pretendido pelo texto. De acordo com Yasseri, Kornai e Kertész (2012), a leiturabilidade é um dos principais temas relacionados à complexidade linguística. Apresentam-se, na tabela 4, os resultados do índice Gunning Fog [Gunning 1952] e do índice Flesch [Flesch 1979]. Esses dois índices são amplamente utilizados. Segundo Yasseri, Kornai e Kertész (2012), o índice Gunning Fox é uma das métricas mais confiáveis de leiturabilidade, e Branco et al. (2014b) afirmam que o índice Flesch é uma das métricas mais aceitas no que diz respeito à leiturabilidade.

Tabela 4. Resultados de dois índices de leiturabilidade

	Ensino Fundamental	Ensino Superior
Índice Gunning Fog	6,64537833333333	8,062712
Índice Flesch	71,5664383333333	60,57434

Segundo Štajner et al. (2012) e Leal et al. (2021), o índice Gunning Fog soma a quantidade média de palavras por sentença ao percentual de palavras com mais de duas sílabas (palavras difíceis) no texto e multiplica o resultado por 0,4. A fórmula é *0,4 x (comprimento médio das sentenças + palavras difíceis)*. Quanto maior o resultado, mais complexo o texto. Como se pode observar na tabela 4, as narrativas produzidas pelos alunos de ensino superior apresentam complexidade mais alta no que diz respeito ao

índice Gunning Fog.

Ainda segundo Štajner et al. (2012) e Leal et al. (2021), o índice Flesch relaciona o comprimento médio das sentenças e o número médio de sílabas por palavra. A fórmula é *248,835 – (1,015 x comprimento médio das sentenças) – (84.6 x número médio de sílabas por palavra)*. Ao contrário do índice Gunning Fox, no índice Flesch, maior resultado da métrica indica menor complexidade textual. Dessa forma, como as narrativas produzidas pelos alunos de curso superior apresentam métrica mais baixa, elas são mais complexas do que as narrativas dos alunos de ensino fundamental de acordo com o índice Flesch.

Considerações finais

Neste trabalho, analisou-se um córpus formado por vinte narrativas orais (dez produzidas por estudantes do ensino fundamental e dez produzidas por estudantes do curso de Comunicação Social) com o objetivo de se verificar se a complexidade textual das narrativas aumenta conforme aumenta o nível de escolaridade. As onze métricas selecionadas demonstraram que as narrativas do córpus produzidas pelos alunos de curso superior apresentam maior complexidade textual do que as narrativas produzidas pelos alunos de ensino fundamental.

Nas métricas de complexidade sintática, foram encontradas, nas narrativas de curso superior, árvores de dependência com maior distância entre palavras relacionadas, uma proporção muito mais alta de orações subordinadas e uma maior proporção de construções na voz passiva analítica. Nas métricas de diversidade lexical, a proporção de *types* de substantivos e de verbos (em relação aos *tokens* dessas respectivas classes) também foi mais alta nas narrativas de curso superior.

Nas métricas de informação morfossintática das palavras, encontraram-se, nas narrativas dos alunos de ensino superior, uma maior proporção de pronomes relativos e uma maior proporção de verbos plenos. Em relação às métricas de informação semântica de palavras, foram encontradas, nas narrativas dos alunos de curso superior, uma maior proporção de nomes próprios bem como de substantivos abstratos.

Os índices de leiturabilidade, que levam em conta o comprimento das sentenças e a proporção de palavras difíceis, também indicaram maior complexidade nas narrativas dos alunos de curso superior.

Espera-se que este trabalho possa auxiliar no trabalho docente indicando algumas características que são esperadas de acordo com o grau de escolaridade. Conforme o aluno vai avançado nos níveis de ensino, presume-se que seus textos apresentem maior diversidade lexical (palavras diferentes e mais difíceis), maior complexidade sintática (uso de orações subordinadas e de pronomes relativos, orações mais longas, uso de construções passivas).

References

Antonio, J. D. (1998). “Narrativas orais e narrativas escritas: a estrutura argumental preferida, e outras preferências”. In Faculdade de Ciências e Letras: Doutorado. Universidade Estadual Paulista Júlio de Mesquita Filho.

Antonio, J. D. (2004). “Estrutura retórica e articulação de orações em narrativas orais e em narrativas escritas do português”. In Faculdade de Ciências e Letras: Doutorado. Universidade Estadual Paulista Júlio de Mesquita Filho.

Evers, A. (2018). “A redação engaiolada: padrões lexicais e ensino de redação em cursos pré-vestibulares populares”. In Instituto de Letras: Doutorado. Universidade Federal do Rio Grande do Sul.

Branco, A., Rodrigues, J., Costa, F., Silva, J. and Vaz, R. (2014a). Rolling out Text Categorization for Language Learning Assessment Supported by Language Technology. In *Computational Processing of the Portuguese Language*: 11th International Conference, PROPOR 2014, Sao Carlos/SP, Brazil, October 6-8, 2014, Proceedings (Vol. 8775, p. 256). Springer.

Branco, A., Rodrigues, J., Costa, F., Silva, J. and Vaz, R. (2014b). Assessing automatic text classification for interactive language learning. In *International Conference on Information Society* (i-Society 2014) (pp. 70-78). IEEE.

Goldman, S. R. and Lee, C. D. (2014). Text complexity: State of the art and the conundrums it raises. *The Elementary School Journal*, 115(2), 290-300.

Feng, L., Jansche, M., Huenerfauth, M. and Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics*: Posters (p. 276-284).

Flesch, Rudolf (1979). “How to write in plain English: A book for lawyers and consumers”, New York, Harper.

Gunning, R. (1952). “The technique of clear writing”, McGraw-Hill, New York.

Leal, S. E., Scarton, C. E., Cunha, A., Hartmann, N. S., Duran, M. S. and Aluísio, S. M. (2021) *NILC-Metrix Doc*. NILC-Metrix. Acesso em 19 mai 2023. Disponível em <http://fw.nilc.icmc.usp.br:23380/metrixdoc>.

Leal, S. E., Duran, M. S., Scarton, C. E., Hartmann, N. S. and Aluísio, S. M. (2022). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *arXiv preprint arXiv*:2201.03445.

McNamara, D. S., Graesser, A. C., McCarthy, P. M. and Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix, Cambridge, Cambridge University Press.

Ponomarenko, G. L. and Evers, A. (2022). “Leiturabilidade e ensino: autores-base e seus trabalhos”, In Acessibilidade textual e terminológica, Edited by Maria José B. Finatto & Liana Braga Paraguassu, Uberlândia, Edufu, p. 41-71.

Santucci, V., Santarelli, F., Forti, L. and Spina, S. (2020). Automatic classification of text complexity. *Applied Sciences*, 10(20), 7285.

Sheehan, K. M., Flor, M. and Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility* (pp. 49-58).

Štajner, S., Evans, R., Orasan, C. and Mitkov, R. (2012). What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility* (p. 14-22).

Yasseri, T., Kornai, A. and Kertész, J. (2012). A practical approach to language complexity: a Wikipedia case study. *PloS one*, 7(11), e48386.