ENTRE NÚMEROS E PALAVRAS: UMA ABORDAGEM PARA PESQUISAS EM LINGUÍSTICA DE *CORPUS*

BETWEEN NUMBERS AND WORDS: AN APPROACH TO RESEARCH IN CORPUS LINGUISTICS

Maria Eduarda Faraco Ávila e Silva[1]
Universidade Federal de Alfenas

Jackson Wilke da Cruz Souza[2]
Universidade Federal da Bahia

Flaviane Faria Carvalho[3]
Universidade Federal de Alfenas

Resumo
Neste artigo objetiva-se apresentar o software AntConc como uma alternativa para pesquisadores que desejam facilitar o processo de análise, evitando que a fase manual seja extensa. Para tanto, o presente trabalho utilizou a pesquisa "Análise de Campos Semânticos em Textos Orais de Divulgação Científica" como base para exemplificar os conceitos que compõem um corpus linguístico, que, grosso modo, pode ser definido como um conjunto de dados linguísticos analisáveis por computador. De acordo a literatura da área, a representatividade e o tamanho do corpus a ser analisado são elementos essenciais para se alcançar o objetivo de uma pesquisa, uma vez que, quanto mais representativo e, consequentemente, maior, esse corpus for, maior a chance de ele contemplar o cerne do trabalho. Sendo assim, propôs-se que o software criado por Anthony Lawrence sirva não só como ferramenta facilitadora de análise, mas também como uma abordagem alternativa de teoria linguística, uma vez que é possível analisar um único corpus de diversas maneiras e partir de vários níveis de análise linguística, como a Semântica, Sintaxe, Morfologia e até mesmo Análise do Discurso.

Palavras-Chave*:* Linguística de *corpus*. Antconc. Representatidade.

Abstract
We aim to present the software AntConc as an alternative for researchers that wish to make the analysis process easier by avoiding the extensiveness of manual phase. Therefore, we took the research "Analysis of Semantic Fields in Oral Texts of Scientific Dissemination" as basis to exemplify the concepts that compose a linguistics *corpus*, which, roughly speaking, can be defined as a set of computer-analyzable linguistic data. According to the literature of the area, the *corpus* representativeness and size are essential elements to achieve the goal of the research, because the more representative the *corpus*, the greater the chance of contemplating the core of a research. Therefore, we proposed that the software created by Anthony Lawrence works not only as a facilitating tool for analysis, but also as an alternative approach of linguistics theory, once it is possible to analyze one *corpus* in different ways and from several distinct levels of linguistic analysis, such as Semantics, Syntax, Morphology, and even, Discourse Analysis.
Keywords: *Corpus* linguistics. Antconc. Representativeness.

[1]Graduanda em Letras pela Universidade Federal de Alfenas (UNIFAL-MG).
E-mail: maria.avila@sou.unifal-mg.edu.br
ORCID: https://orcid.org/0000-0002-3332-4680

[2]Doutor em Linguística pela Universidade Federal de São Carlos (UFSCar) e docente no Instituto de Ciência, Tecnologia e Inovação da Universidade Federal da Bahia (UFBA).
E-mail: jackcruzsouza@gmail.com
ORCID: https://orcid.org/0000-0003-1881-6780

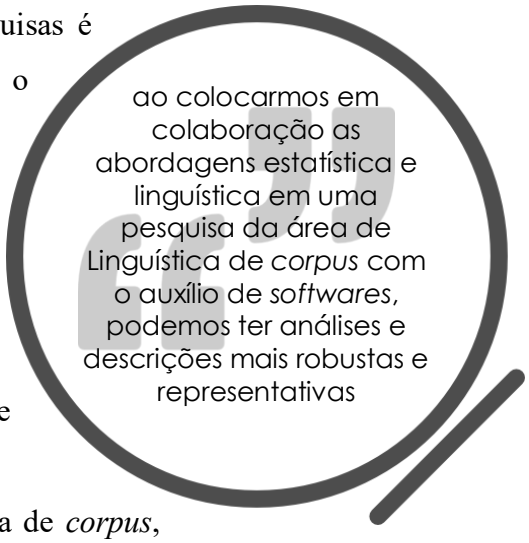
[3] Doutora em Linguística Aplicada pela Universidade de Lisboa e docente no Instituto de Ciências Humanas e Letras (ICHL) da UNIFAL-MG.
E-mail: flaviane.carvalho@unifal-mg.edu.br.
ORCID: https://orcid.org/0000-0002-0663-670X

O QUE HÁ POR TRÁS DO PROCESSAMENTO?

Você já se perguntou por que em determinados contextos de fala algumas pessoas trocam o "l" pelo "r", como em "ploblema"? As razões para isso acontecer são inúmeras, e podemos ter respostas a essa questão de diferentes áreas do conhecimento, como a Fonologia, Neurociência e a Linguística. Esta última, em especial, apresenta outras tantas hipóteses de aspectos sociais, fonético-fonológicos e contextuais para lidar com essa questão. Entretanto, não citamos aqui esse exemplo para tentar entender tais hipóteses ou construir juízo de valor para cada uma delas; o interessante deste exemplo é pensar as formas como se corroboram ou se refutam essas hipóteses.

No caso da Linguística, será necessário olhar para um conjunto de dados, coletados com o objetivo de observar se nosso exemplo ocorre de maneira isolada ou disseminada entre os brasileiros, por exemplo, ou ainda entender quais as condições que levam um falante a pronunciar ou grafar determinadas palavras dessa ou daquela maneira. Assim, pode ser definida a Linguística de *corpus* (LC), subárea que trata da coleta de dados linguísticos textuais selecionados especificamente para uma pesquisa linguística (SARDINHA, 2000). Mais especificamente, o *corpus*, enquanto conjunto de dados sobre a linguagem, só pode ser considerado como tal se os procedimentos de extração, manipulação e armazenamento forem realizados por meio de um computador.

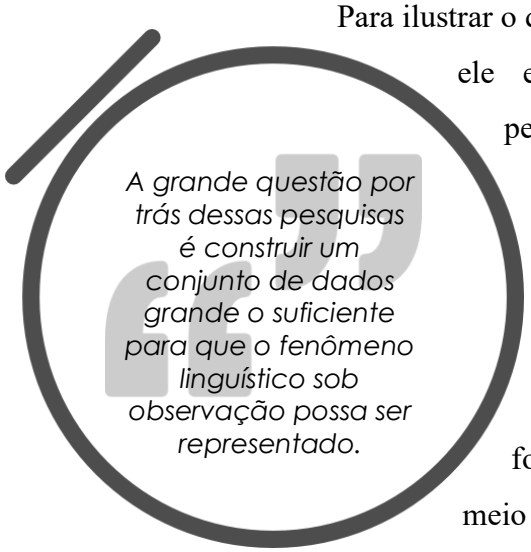
A grande questão por trás dessas pesquisas é construir um conjunto de dados grande o suficiente para que o fenômeno linguístico sob observação possa ser representado. Nesse sentido, apontar para a representatividade é condicioná-la ao tamanho, significando, assim, que quanto maior/mais variado for o *corpus*, maior será a probabilidade de ele representar aquilo que se pretende pesquisar.



É importante ressaltar que a Linguística de *corpus*, portanto, destaca-se tanto como área quanto metodologia para *pesquisas quantitativas* (ou seja, importa a quantificação para os resultados) ou *qualitativas* (ou seja, para pesquisas que se importam com a explicação ou a descrição de fenômenos e usos linguísticos). Voltando ao nosso exemplo inicial: se quisermos observar o fenômeno citado a nível

nacional para chegarmos à confirmação de uma hipótese será necessário aumentar expressivamente a quantidade de dados para, então, poder tecer descrições ou explicações sobre a troca consonantal.

Para uma análise mais rica e detalhada sobre o fenômeno exemplificado frente ao aumento do conjunto de dados será necessária a utilização de ferramentas e sistemas computacionais. É exatamente aí que as abordagens computacionais entram em cena: com o objetivo de poder auxiliar quantitativa e qualitativamente os pesquisadores a poderem ampliar suas percepções sobre a linguagem, seus usos e seus falantes, como apontam Rodrigues, Souza e Santos (2022).



Para ilustrar o quão importante é a organização do *corpus* para que ele evidencie determinados fenômenos, citamos a pesquisa "Análise de Campos Semânticos em Textos Orais de Divulgação Científica" (SILVA; CARVALHO; SOUZA, 2022). Ela foi desenvolvida com o objetivo de analisar os campos semânticos do *corpus* oral selecionado. Para tanto, as palavras que circulam em volta de uma ideia foram analisadas quantitativa e qualitativamente por meio do *software* AntConc (ANTHONY, 2005). Esses grupos de ideias podem ser definidos, de maneira geral, como campos semânticos, que podem ter proximidade a partir da formação da palavra (como em "saúde" e "saudável"), mas sobretudo com a ideia, (como em "vacina" e "Covid-19").

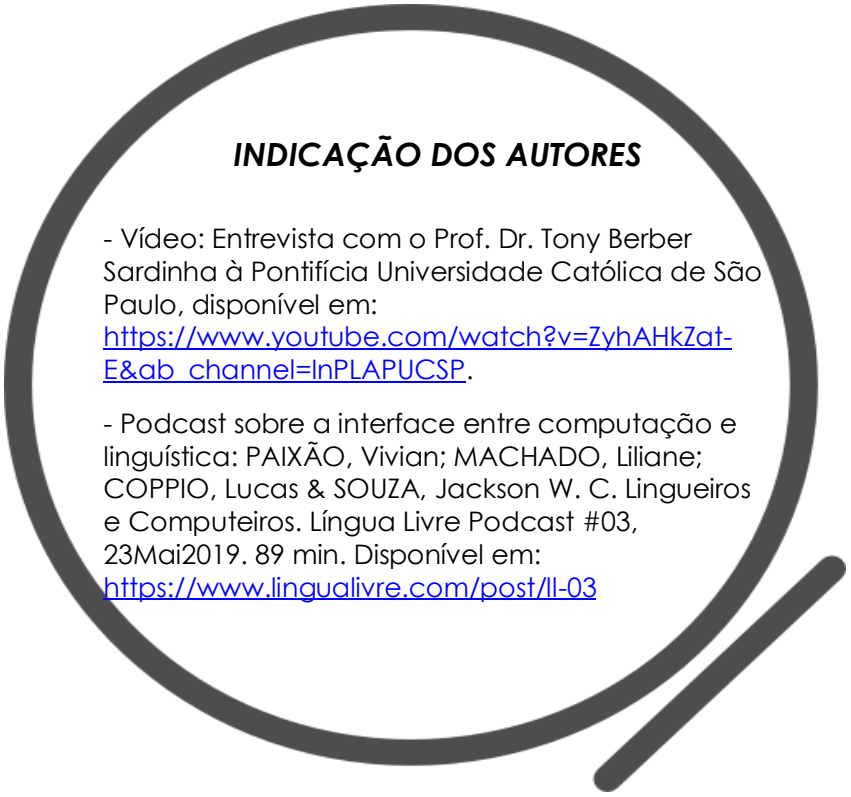
No referido *software*, seleciona-se o *corpus* que será analisado e, em seguida, indicam-se quais combinações podem ser feitas. É possível, por exemplo, investigar a ocorrência de uma ou duas palavras juntas, ou ainda as combinações e contextos frasais com determinadas palavras. Após isso, numa fase manual, as palavras são agrupadas e classificadas em seus respectivos campos semânticos.

Para garantir a representatividade, além do tamanho, foram considerados cinco temas diferentes, a saber: i) Covid-19, ii) Assuntos Diversos, iii) Saúde Mental, iv) Uso Racional de Medicamentos, v) A Comunidade LGBTQIAP+. Se observarmos com atenção, os temas podem ser abordados a partir de diferentes áreas do conhecimento, como Ciências da Saúde, Ciências Humanas, Ciências Biológicas e Ciências Sociais, por exemplo. Assim, os campos semânticos levantados representam não apenas a tentativa de

divulgar determinados conhecimentos à sociedade não especializada em determinados assuntos, mas também a interseção entre termos e expressões com esse objetivo. Silva, Carvalho e Souza (2022), então, demonstram como alguns conceitos e palavras podem se agrupar para auxiliar na circulação desses conhecimentos entre os ouvintes de uma programação de rádio e/ou podcast.

Obviamente, esses resultados poderiam ser alcançados de maneira puramente manual e qualitativa. Entretanto, o volume de dados analisados e o tempo de execução da pesquisa jamais poderiam ser os mesmos. Assim, ao colocarmos em colaboração as abordagens estatística e linguística em uma pesquisa da área de Linguística de *corpus* com o auxílio de *softwares*, podemos ter análises e descrições mais robustas e representativas sobre usos linguísticos. Além disso, os resultados são obtidos com mais rapidez e segurança, uma vez que se diminui a chance de erros descritivos, próprios de análises puramente manuais.

Finalizamos aqui destacando que é necessário fazer uma abordagem adicional à computacional. Não descartamos a possibilidade de se utilizar uma teoria linguística que dialogue com a metodologia aqui rapidamente rascunhada. Na verdade, entendemos elas estabelecem uma relação de mutualismo e complementaridade, uma vez que a pesquisa se torna mais rica e completa quando as abordagens e áreas são agregadas.



REFERÊNCIAS BIBLIOGRÁFICAS

ANTHONY, L. AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. *In* IPCC 2005. **Proceedings**. International Professional Communication Conference, 2005. IEEE, 2005. p. 729-737.

RODRIGUES, R.; SOUZA, J.W.C.; SANTOS, R.L.S. Descrição linguística e aprendizado de máquina: análise de verbos locativos do espanhol. **Cadernos de Estudos Linguísticos**, Campinas, SP, v. 64, p. 1-15, 2022.

SARDINHA, T. B. Linguística de *Corpus*: histórico e problemática. DELTA, São Paulo, v. 16, n. 2, p. 323-367, 2000.

SILVA, M.E.F.A.; CARVALHO, F.F.; SOUZA, J.W.C. Análise de aspectos sintáticos e semânticos de textos multimodais de divulgação científica. *In* **Anais do VIII Simpósio Integrado da UNIFAL-MG**, 2022, Alfenas/MG. ISSN 2763-9282. Disponível em: https://sistemas.unifal-mg.edu.br/app/caex/comum/paginas/baixaArquivo.php?caminho=/server/arquivos/caex/inscricoes/submissoes/S000022126.pdf. Acesso em: 14/03/2023.