Nome: Adeldivo Alves de Sousa Junior

Resenha: O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística

O artigo "O que é e como se constrói um corpus?", escrito por Sandra Maria Aluísio e Gladis Maria de Barcellos Almeida, explora o conceito de corpus no contexto da linguística, apresentando as definições, requisitos e procedimentos necessários para sua construção. O texto inicia discutindo a evolução histórica do uso de corpora em pesquisa linguísticas, mencionando exemplos desde o século XVIII, como o *Vocabulário Portuguez e Latino*, elaborado pelo Padre Rafael Bluteau . Destaca-se a diferença entre a concepção tradicional de corpus, focada em textos impressos, e a concepção moderna da Linguística de Corpus, que envolve o uso de corpora digitais, permitindo a análise empírica e automatizada, essencial para o processamento de grandes volumes de dados.

No contexto do Processamento de Linguagem Natural (PLN), os corpora desempenham um papel crucial, uma vez que a qualidade das ferramentas de PLN, como analisadores sintáticos[1], etiquetadores morfossintáticos[2] e tradutores automáticos, depende diretamente da representatividade e qualidade dos dados de treinamento. Os autores ressaltam que, para o sucesso de pesquisas e aplicações em PLN, o corpus deve ser autêntico, representativo e balanceado, ou seja, deve refletir adequadamente as características linguísticas da comunidade cuja língua está sob análise. Esses requisitos garantem que os modelos de PLN possam generalizar corretamente a partir dos dados e atuar de maneira eficiente em diferentes contextos linguísticos.

Um ponto relevante abordado no artigo é a importância do formato eletrônico dos corpora, uma vez que o processamento de grandes quantidades de dados linguísticos só é possível graças à informatização desses textos. Ferramentas computacionais modernas podem realizar análises complexas em segundos, como a extração de padrões sintáticos e semânticos, que são praticamente impossíveis de detectar manualmente. A criação de corpora eletrônicos facilita o desenvolvimento de sistemas de PLN que realizam tarefas como tradução automática, reconhecimento de fala, análise de sentimentos e resumo automático de textos.

Além de discutir a teoria por trás da construção de corpora, o artigo também destaca a contribuição prática para o PLN ao mencionar projetos como o *Lácio-Web* e o *BootCat*. O *Lácio-Web* é particularmente importante para o PLN no Brasil, pois oferece corpora em português brasileiro, que historicamente tem menos recursos disponíveis em comparação com o inglês. O acesso a esses corpora facilita o desenvolvimento de ferramentas de PLN mais precisas e eficientes para o português. O *BootCat*, por sua vez, automatiza a extração de corpora a partir da *web*, permitindo que ele seja atualizado e focado em temas específicos, sem a necessidade de coleta manual dos dados.

Em síntese, o artigo oferece uma visão abrangente sobre a importância dos corpora para a pesquisa linguística e, em especial, para o PLN. Ele destaca as ferramentas e métodos necessários para a construção e uso eficazes de corpora, com ênfase em sua aplicabilidade no contexto brasileiro, onde o desenvolvimento de ferramentas de PLN para o português ainda enfrenta desafios. O artigo também

[1] **Analisador sintático**: é um programa que, ao receber a descrição formal de uma gramática (associada a uma linguagem), gera como saída um código-fonte capaz de reconhecer cadeias (sentenças) válidas de acordo com essa gramática específica (Boss & Venske, 2008, p. 13-22).
[2] **Etiquetagem morfossintática**: representa um dos primeiros estágios na análise linguística. Situada entre a morfologia e a sintaxe, seu objetivo é identificar as classes gramaticais de cada palavra ou token (Silva, 2023).

reforça que o avanço dessas ferramentas depende da construção de corpora bem projetados, que possam ser reaproveitados em diversas pesquisas e aplicações práticas.

Referências Bibliográficas

Aluísio, Sandra Maria; Almeida, Gladis Maria de Barcellos. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística**. Calidoscópio Vol. 4. P. 156-178, 2006. Disponível em: https://edisciplinas.usp.br/pluginfile.php/391802/mod_resource/content/1/Corpus_o%20que%20%C3%A9.pdf

Boss, Silvio Luiz Bragatto; Venske, Sandra Mara Guse Scós. **Analisadores sintáticos: conflitos e ambiguidades**. Revista Científica da FAI, Vol. 8, P. 13-22, 2008. Disponível em: https://www.fai-mg.br/portal/download/revista_cientifica_2008/pub_dw_artigo_analisadores.pdf

Silva, Emanuel Huber da. **Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies**. 2023. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-04092023-145651/pt-br.php