



CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS

PPG EM ESTUDOS DE LINGUAGENS

Disciplina: Metodologia da Pesquisa

Docente: Dr^a Raquel Bambirra

Discente de mestrado: Flávio Ernani

O que é e como se constrói um *corpus*?

Lições aprendidas na compilação de vários *corpora*
para pesquisa linguística

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um corpus?
Lições aprendidas na compilação de vários corpora para pesquisa
linguística. **Calidoscópio**. v. 4. n. 3. p. 156-178. 2006

Concepção de *corpus*

Linguística

Galisson e Coste (1983)

Conjunto finito de enunciados tomados como objeto de análise ou, mais especificamente, é um conjunto finito de enunciados considerados característicos do tipo de língua a estudar, reunidos para servirem de base à descrição e, eventualmente, à elaboração de um modelo explicativo dessa língua.

Coleção de documentos orais, escritos ou orais e escritos, considerando o tipo de investigação. Pode ser um *Corpus exaustivo* quando compreende todos os enunciados. E é *selectivo* quando compreende apenas uma parte desses enunciados.

Dubois et al. (1993)

Conjunto de enunciados a partir do qual se estabelece a gramática descritiva de uma língua. Ele é uma amostra da língua. Deve ser representativo. Se o número de enunciados for indefinido, não existe a exaustividade. Grandes quantidades de dados inúteis podem complicar a pesquisa. Os autores sugerem que o linguista sempre deve desconfiar do método de pesquisa escolhido.

Linguística

Ducrot e Todorov (2001)

Corpus é um “conjunto, tão variado quanto possível, de enunciados efetivamente emitidos por usuários da referida língua em determinada época”.

Trask (2004)

Define como “um conjunto de textos escritos ou falados numa língua, disponível para análise”.

Linguística de Corpus

Sinclair (2005) explica que o *corpus* é uma coletânea de textos em certo idioma que esteja em formato eletrônico. Esses textos devem ser selecionados de acordos com critérios externos, ou seja, critérios que nascem a partir das necessidades da pesquisa na qual o *corpus* será usado e que sejam capazes de representar uma língua ou uma parcela de língua.

É uma “abordagem que se ocupa da coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados **criteriosamente** com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por computador” (Berber Sardinha, 2004).

- E os livros, revistas e textos impressos? Segundo a Linguística de *Corpus* eles não se constituem como *corpus*, já que não estão em um formato que possam ser processados por computador.
- Dois grandes pontos que diferem entre a Linguística e a Linguística de *Corpus* são: o formato computadorizado do *corpus* e a sua posterior disponibilização para outras pesquisas.
- McEnery e Wilson (1996) consideram quatros características fundamentais para a noção de *corpus*: amostragem e representatividade; tamanho finito; formato eletrônico; referência padrão.
- Aluísio e Almeida (2006) admitem que o formato computadorizado auxilia nos procedimentos de observação e descrição, que antes eram manuais. “Por meio de *corpus*, podem-se observar aspectos morfológicos, sintáticos, semânticos, discursivos (...). Em resumo, por meio de *corpus*, descreve-se a língua de forma objetiva.”

Requisitos e procedimentos para a elaboração

Corpus computadorizado - primeiro passo é verificar se o *corpus* de estudo serve ao propósito inicial da pesquisa. Ele será suficiente para responder à questão? Checada essa questão, então o pesquisador deverá se atentar para os seguintes aspectos:

- Autenticidade – os textos devem ser produzidos em linguagem natural e por quem é o falante a ser estudado.
- Representatividade – deve representar a língua ou a variedade da língua a ser pesquisada. Algumas questões para checar a representatividade: quais documentos? Quais tipos de textos? Quais gêneros textuais? De fato representa os usos linguísticos da comunidade pesquisada?
- Balanceamento – as escolhas devem ser adequadas à pesquisa que se pretende realizar, demonstrando que os textos foram escolhidos criteriosamente.

- Amostragem – deve ser uma amostragem representativa, não somente proporcional.
- Diversidade – no sentido de não tentar generalizar a língua, mas representá-la em sua diversidade de gêneros e tipos de textos.
- Tamanho – adequado ao tipo de pesquisa e metodologia a ser adotada

Etapas metodológicas para a compilação de um *corpus*

- Projeto de *corpus*: a seleção dos textos
- Compilação ou armazenamento, manipulação, nomeação dos arquivos e proteção da identidade dos participantes
- Anotação

Os *corpora*

- Arquivos da Folha
- Lácio-Web – disponibiliza vários *corpora*, em bancos de textos compilados, catalogados e codificados. Ainda disponibiliza ferramentas como contadores de frequência.
- Projeto COMET – disponibiliza um *corpus* eletrônico, servindo de suporte a pesquisas linguísticas, principalmente nas áreas de tradução, terminologia e ensino de línguas.
- Linguateca – tem como objetivo os estudos sobre processamento do português.

Ferramentas existentes e disponíveis

Processamento

WebCorp : permite extrair fatos sobre várias línguas como se a Web fosse um *corpus*. Tem versões *demo* disponibilizadas gratuitamente na Web. Pode ser usado para analisar como certas palavras e expressões são usadas, especialmente as palavras raras ou neologismos.

Unitex: é um conjunto de programas para processamento de *corpus* composto por interface gráfica. Possui dicionários e tabelas léxico-gramática. Suporte para mais de 14 idiomas.

← → ↻

Apps Sites Sugeridos

WebCorp Live

Concordance the web in real-time.

Search Wordlist Tool User Guide WebCorp LSE Publications Feedback

WebCorp Live lets you access the Web as a corpus - a large collection of texts from which examples of real language use can be extracted. [More...](#)

Search:

Case Insensitive: ☒ Span:

Search API: Language:

[Advanced Options](#)

Redefinir **Search**

WebCorp

Linguist's Search Engine

Have you tried WebCorp LSE?

Our large-scale search engine with more search options, part-of-speech tags and quantitative analyses.

[More details...](#)

Figura 1: Tela principal do *WebCorp* a partir da qual se podem escolher as opções do menu e acessar as opções avançadas de busca.

Ferramentas de geração e gerenciamento de corpora especializados

Corpógrafo - é um gestor de *corpus* para pesquisas terminológicas. Extração e organização em bases de dados.

BootCaT – propõe a construção interativa, a partir de textos obtidos na web. Ferramentas para extração de termos com mais de uma palavra, ou termos multipalavras.