

PROCESSAMENTO DE LINGUAGEM NATURAL

Corpus



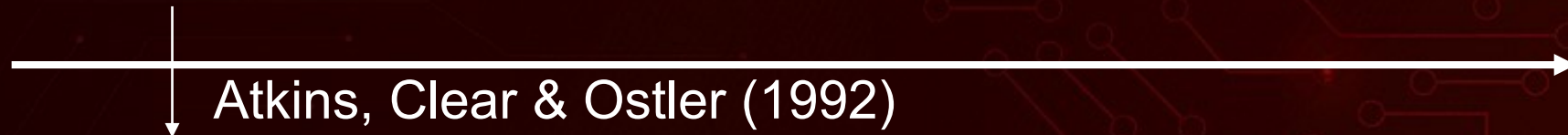
TÓPICOS

1. O que é um *corpus*?
2. História
3. Tipologia
4. *Corpora* atuais



O QUE É *CORPUS*?

Existem várias definições de *corpus* na literatura, algumas vezes divergentes:



Um subconjunto de uma biblioteca de texto eletrônico, construído de acordo com **critérios de projeto explícitos para uma finalidade específica**, ex., o corpus Cobuild, o corpus Longman/Lancaster

O QUE É CORPUS?

Existem várias definições de *corpus* na literatura, algumas vezes divergentes:

↓ Crystal, David (1992)

“*corpus, plural: corpora,*

A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language – for example, to determine how the usage of a particular sound, word, or syntactic construction varies. A computer corpus is a large body of machine-readable texts.”

O QUE É *CORPUS*?

Existem várias definições de *corpus* na literatura, algumas vezes divergentes:



Sanchez (1995)

“Um conjunto de **dados linguísticos** (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, **suficientemente extensos** em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que **possam ser processados por computador**, com a finalidade de propiciar resultados vários e úteis para a descrição e análise.”

O QUE É CORPUS?

Existem várias definições de *corpus* na literatura, algumas vezes divergentes:

↓ McEnery & Wilson (1996)

*In principle, **any collection of more than one text can be called a corpus**, (corpus being Latin for "body", hence a corpus is any body of text). But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition.*

*These may be considered under four main headings: **sampling and representativeness; finite size; machine-readable form; and a standard reference***

O QUE É *CORPUS*?

Existem várias definições de *corpus* na literatura, algumas vezes divergentes:

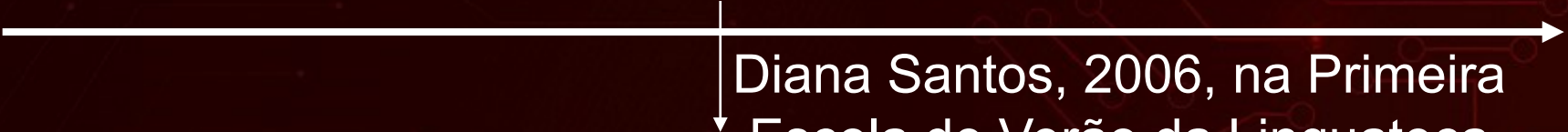


Sardinha (2004)

1. Os dados devem ser autênticos
2. A finalidade de ser um objeto de estudo linguístico
3. O conteúdo do *corpus* deve ser criteriosamente escolhido
4. Os dados devem ser legíveis por computador
5. O *corpus* deve ser representativo de uma língua ou variedade
6. O *corpus* deve ser vasto para ser representativo

O QUE É *CORPUS*?

Existem várias definições de *corpus* na literatura,
algumas vezes divergentes:



Diana Santos, 2006, na Primeira
Escola de Verão da Linguateca

“*Corpus* é uma coleção classificada de objetos linguísticos
para uso em Processamento de Linguagem
Natural/Linguística Computacional/Linguística”

A HISTÓRIA DOS *CORPORA*

- Década de 60 – *corpora* de 1 milhão de palavras!
- Brown – inglês americano (1964)
 - - textos publicados em 1961
 - 200 textos de 5.000 palavras cada
 - 15 categorias distintas
- LOB (Lancaster/Oslo/Bergen) – (1978) contrapartida em inglês britânico
- London-Lund: 500 mil, falado (1980)
- Frown + Flob (Freiburg) – 1990
- American English 2006 + British English 2006

A HISTÓRIA DOS *CORPORA*

- Década de 1960: advento do computador
- Década de 1970: consolidação, *corpora* diversos, línguas diversas; *corpora* maiores, *corpora* anotados (Escandinávia) – tudo digitalizado
- Década de 1980: invenção do *scanner*
- Década de 1990: disponibilidade de textos já digitalizados
- Novo milênio: textos da internet
- *Web* como *corpus*

A HISTÓRIA DOS *CORPORA*

- O **português** também avança
 - *Corpus* NILC e o fortalecimento do PLN no Brasil
 - *Corpus* Brasileiro, com 1 bilhão de palavras
 - Tycho-Brahe e o português histórico (autores nascidos entre 1380 e 1845)
 - C-Oral-Brasil, com fala espontânea
 - Floresta Sintá(c)tica, com anotação morfossintática e sintática
 - CSTNews, com diversas camadas de anotação
 - brWaC, com material da *web*
 - E muitos outros!
 - [AC/DC](#), projeto pioneiro da Linguateca

TIPOLOGIA

(Sardinha, 2000)

- **Modo**

- **Falado: composto por falas transcritas**
- **Escrito: composto por textos escritos, impressos ou não**

- **Tempo**

- **Sincrônico: compreende um período de tempo**
- **Diacrônico: compreende vários períodos de tempo**
- **Contemporâneo: representa o período de tempo corrente**
- **Histórico: representa um período de tempo passado**

TIPOLOGIA

(Sardinha, 2000)

- **Seleção**

- **De amostragem:** composto por porções de textos ou de variedades textuais, planejado para ser uma amostra finita da linguagem
- **Monitor:** a composição é reciclada para refletir o estado atual de uma língua, opondo-se a *corpus* de amostragem
- **Dinâmico ou orgânico:** o crescimento e diminuição são permitidos, qualifica o *corpus* monitor
- **Estático:** oposto de dinâmico
- **Equilibrado:** os componentes (gêneros, textos, etc.) são distribuídos em quantidades semelhantes

TIPOLOGIA

(Sardinha, 2000)

- **Conteúdo**

- Especializado: os textos são de tipos específicos (em geral gêneros ou registros definidos)
- Regional ou dialetal: os textos são provenientes de uma ou mais variedades sociolinguísticas específicas
- Multilíngue: inclui idiomas diferentes

- **Autoria**

- De aprendiz: os autores dos textos não são falantes nativos
- De língua nativa: os autores são falantes nativos

TIPOLOGIA

(Sardinha, 2000)

- **Disposição interna**

- Paralelo: os textos são comparáveis (por exemplo, original e tradução)
- Alinhado: as traduções aparecem abaixo de cada linha do original
 - Mas, como já comentado, esses conceitos já foram estendidos para outras situações

- **Finalidade**

- De estudo: o *corpus* que se pretende descrever
- De referência: usado para fins de contraste com o *cópus* de estudo
- De treinamento ou teste: construído para permitir o desenvolvimento de aplicações e ferramentas de análise
 - Muito comum em PLN

TIPOLOGIA

(Sardinha, 2000)

- Também há outros critérios possíveis
 - **Pluralidade de autoria:** os textos foram produzidos por um autor apenas ou mais?
 - **Integralidade:** os elementos do *corpus* são textos integrais ou fragmentos?
 - **Plurilinguismo:** o *corpus* possui só textos originais ou também as traduções desses textos para uma ou mais línguas?
 - Intercalação: as traduções dos textos são incorporadas a cada linha do texto original ou vêm em textos separados?

OS CORPORA ATUAIS



- **BNC – 1995 – 100 milhões de palavras – *online* 40 milhões**
 - 90% língua escrita
 - 10% língua falada
 - *corpus* fechado**BYU-BNC – completo on-line:**
<http://corpus.byu.edu/bnc/>
completo 100 milhões de palavras
- **COCA – *Corpus of Contemporary American English* –**
www.americancorpus.org/
 - 520 milhões de palavras
 - 1990 – 2015


OS *CORPORA* ATUAIS















English-Corpora.org

[corpora](#) [guides](#) [videos](#) [related resources](#) [users](#) [my account](#) **upgrade** [help](#)

These are the most **widely used** online corpora, and they are used for **many different purposes** by teachers and **researchers** at **universities** throughout the world. In addition, the corpus data (e.g. **full-text**, **word frequency**) has been used by a **wide range of companies** in many different fields, especially technology and **language learning**. Take a look at an overview (PDF or video):  

The links below are for the free online interface. You can also purchase and download  the corpora for use on your own computer.

Corpus	Overview  	Download	# words	Dialect	Time period	Genre(s)
News on the Web (NOW)			17.5 billion+	20 countries	2010-yesterday	Web: News
iWeb: The Intelligent Web-based Corpus			14 billion	6 countries	2017	Web
Global Web-Based English (GloWbE)			1.9 billion	20 countries	2012-13	Web (incl blogs)
Wikipedia Corpus			1.9 billion	(Various)	2014	Wikipedia
Coronavirus Corpus			1.5 billion	20 countries	Jan 2020-Dec 2022	Web: News
Corpus of Contemporary American English (COCA)			1.0 billion	American	1990-2019	Balanced
Corpus of Historical American English (COHA)			475 million	American	1820-2019	Balanced
The TV Corpus			325 million	6 countries	1950-2018	TV shows
The Movie Corpus			200 million	6 countries	1930-2018	Movies
Corpus of American Soap Operas			100 million	American	2001-2012	TV shows

OS CORPORA ATUAIS

- **ANC – American National Corpus** – <http://www.anc.org/>
 - Textos a partir de 1990
 - 15 milhões de palavras para *download* (“Open” *portion*)
- **MASC – Manually Annotated Sub-Corpus**
 - 19 *genres of American English*

OS CORPORA ATUAIS

- **Corpora de Português**
 - <https://comet.fflch.usp.br/corporaportugues>
- **Exemplo:**
 - www.linguateca.pt → Acesso a recursos → AC/DC Acesso a corpora /Disponibilização de corpora
 - 1 bilhão de palavras

PROJETO AC/DC

Acesso a corpos de português: Projeto AC/DC

[Linguateca](#)

[Information in English](#)

O projeto AC/DC (Acesso a corpos/Disponibilização de corpos), iniciado em 1999, surgiu da necessidade de juntar os poucos recursos disponíveis num único ponto na rede e dessa forma facilitar a comparação e a reutilização do material, permitindo ao mesmo tempo acesso a uma ferramenta poderosa de interrogação de corpos, o [Open CWB](#) (versão nova do *IMS corpus workbench*), para a qual desenvolvemos esta interface.

Desde 2000, a anotação dos corpos tem sido feita automaticamente pelo [PALAVRAS](#) de Eckhard Bick, e convertida para o "formato AC/DC", descrito pormenorizadamente na página de [Anotação](#).

Uma descrição quantitativa inicial dos corpos servidos presentemente pelo AC/DC encontra-se na tabela abaixo. **Clique num dos corpos para o interrogar**. Para cada corpo, pode pedir concordâncias, distribuição e frequências simples e complexas, veja [alguns exemplos](#). (Ao longo do tempo, fomos desenvolvendo uma [série de serviços especializados](#) para consultar os corpos de forma mais complexa.) Se é a primeira vez que visita o AC/DC e quer apenas experimentar, [procure no corpo Vercial](#). Veja também a nossa [PJR](#): lista de perguntas já respondidas.

Breve descrição dos corpos

Corpo	Tamanho (unidades)	Tamanho (palavras)	Tamanho (frases)	Variante(s)	Breve descrição
AmostRA-NILC	134.297	105.499	4.965	BR	AmostRA-NILC
ANCIB	1.672.505	1.243.068	80.775	BR	Correio electrónico correspondente ao tráfego na lista ANCIB
Avante!	7.666.370	6.506.813	193.111	PT	Semanário político Avante!, 1997-2002
Corpus Brasileiro	1.057.661.890	893.043.840	40.981.957	BR	Corpus Brasileiro, um bilhão (mil milhões) de palavras de português do Brasil de vários géneros
CD HAREM	290.001	225.766	12.558	PT BR	Colecção dourada do HAREM
CETEMPúblico	234.481.482	190.601.605	7.025.567	PT	Jornal PÚBLICO, dividido em extractos, 1991-1998
CHAVE	116.836.447	92.387.266	4.385.437	PT BR	Jornais PÚBLICO e Folha de São Paulo, 1994-1995
Ciência Viva	799.360	656.589	27.269	PT	Textos escritos sobre ciência em Portugal
Colonia	6.643.879	4.977.678	283.546	PT BR	Obras dos séculos XVI a XX
CONDIVport	7.132.225	5.558.299	301.047	PT BR	Jornais desportivos e revistas de moda e saúde
CONDIVport2	209.289	172.486	6.533	PT BR	Jornais diários
CoNE	911.431	671.756	31.571	PT BR	Mensagens de correio electrónico não-endereçadas
C-Oral-Brasil	435.507	263.937	30.632	BR	C-Oral-Brasil, português brasileiro oral informal

EX. DE PESQUISA

genero	Gênero	Fonte
ec	Literatura	Crônicas
ed	Jornalismo	Revistas
ee	Educação	Diversos
ef	Educação	Diversos
eg	Jornalismo	Jornais
eh	Literatura	Variados
ei	Acadêmico	Artigos
ej	Acadêmico	Teses e dissertações
et	Enciclopédia	Wikipédia
eu	Literatura	Biografias
ew	Literatura	Variados
ex	Legislação	Diversos
fa	Esporte	Narração de jogos de futebol
fc	Política	Pronunciamentos do presidente
fd	Política	Sessões do congresso
fe	Jornalismo	Entrevistas

Resultados da procura

15 de junho de 2023

Procura: "**Vasco**" "**da**" "**Gama**"

Distribuição de **genero**

Corpo: Corpus Brasileiro v. 6.4

1588 casos.

Distribuição

Houve **16** valores diferentes de **genero**.

<u>eg</u>	1145	250706964
<u>fa</u>	137	86130
<u>ej</u>	65	293646680
<u>ee</u>	55	78775381
<u>fd</u>	53	76659017
<u>et</u>	47	44651112
<u>ei</u>	44	256728689
<u>ef</u>	22	3010731
<u>ex</u>	7	8675387
<u>ew</u>	4	7216865
<u>fe</u>	3	3999318
<u>fc</u>	2	1803765
<u>ec</u>	1	162172
<u>ed</u>	1	493880
<u>eh</u>	1	1371644
<u>eu</u>	1	574933

O QUE VIMOS?

- O que é um *corpus*?
- História
- Tipologia
- *Corpora* atuais



PRÓXIMA VIDEOAULA

- **Similaridade textual**



REFERÊNCIAS

- **Curso de Processamento de Linguagem Natural**
 - **Profa. Helena Caseli (UFSCar)**
- **Curso de Processamento de Linguagem Natural**
 - **Prof. Thiago Pardo (ICMC-USP)**
- **Curso de Linguística Computacional**
 - **Prof. Thiago Castro Ferreira (UFMG)**
- **Curso de Linguística de Corpus**
 - **Profa. Stella Esther Ortweiler Tagnin (FFLCH-USP)**