NORMALIZAÇÃO DE TEXTOS

CEFET-MG

PALAVRAS

Quantas palavras temos na seguinte frase?

```
Extremamente adaptável, pode ocorrer em
ambientes altamente alterados pelo ser
humano.
```

PALAVRAS

Quantas palavras temos na seguinte frase?

```
Extremamente adaptável, pode ocorrer em
ambientes altamente alterados pelo ser
humano.
```

11 palavras.

13 palavras se considerarmos também os sinais de pontuação.

PALAVRAS

Na frase:

```
Muito longo para os que lamentam,
muito curto para os que festejam
```

Quantas palavras existem?

Quantas palavras diferentes existem?

PALAVRAS

Na frase:

```
Muito longo para os que lamentam,
muito curto para os que festejam
```

Quantas palavras existem? **12**

Quantas palavras diferentes existem? **8**

PALAVRAS

Na frase:

```
Muito longo para os que lamentam,
muito curto para os que festejam
```

Quantas palavras existem? **12 (tokens)**

Quantas palavras diferentes existem? **8 (tipo/vocabulário)**

PALAVRAS

Na frase:

`Muito longo para os que lamentam,`

`muito curto para os que festejam`

Quantas palavras existem? **12 (tokens)**

Quantas palavras diferentes existem? **8 (tipo/vocabulário)**

Coletânea sobre um determinado assunto.
Plural: corpora

Corpus	Tokens = N	Types = $\|V\|$
Shakespeare	884 thousand	31 thousand
Brown corpus	1 million	38 thousand
Switchboard telephone conversations	2.4 million	20 thousand
COCA	440 million	2 million
Google N-grams	1 trillion	13 million

PALAVRAS

No contexto de PLN, um **corpus** é um conjunto de documentos (ou de frases) geralmente anotados e utilizados para:

- Aprendizado (análise)
- Validação (verificação)

PALAVRAS

o corpus do português

Corpora
Corpus size
Compare to other corpora
Related resources
Problems
Contact us

English Português

Created by Mark Davies, BYU. Funded by the US National Endowment for the Humanities (2004, 2015).

	Corpus	Size	Created	More info
1	**Genre / Historical**	45 million words	2006	Info
2	**Web / Dialects**	1 *billion* words	2016	Info
3	**NOW (2012 - 2019)**	1.1 *billion* words	2018	Info
4	**WordAndPhrase**	Top 40,000 words	2017	Info

The Corpus do Português now has two different parts:

- the (original, smaller) corpus that allows you to look at historical changes and genre-based variation
- the (new, much larger) corpus that you can use to look at dialectal variation (and have 50x as much data for Modern Portuguese).

Click on an [Info] link above for more details.

PALAVRAS

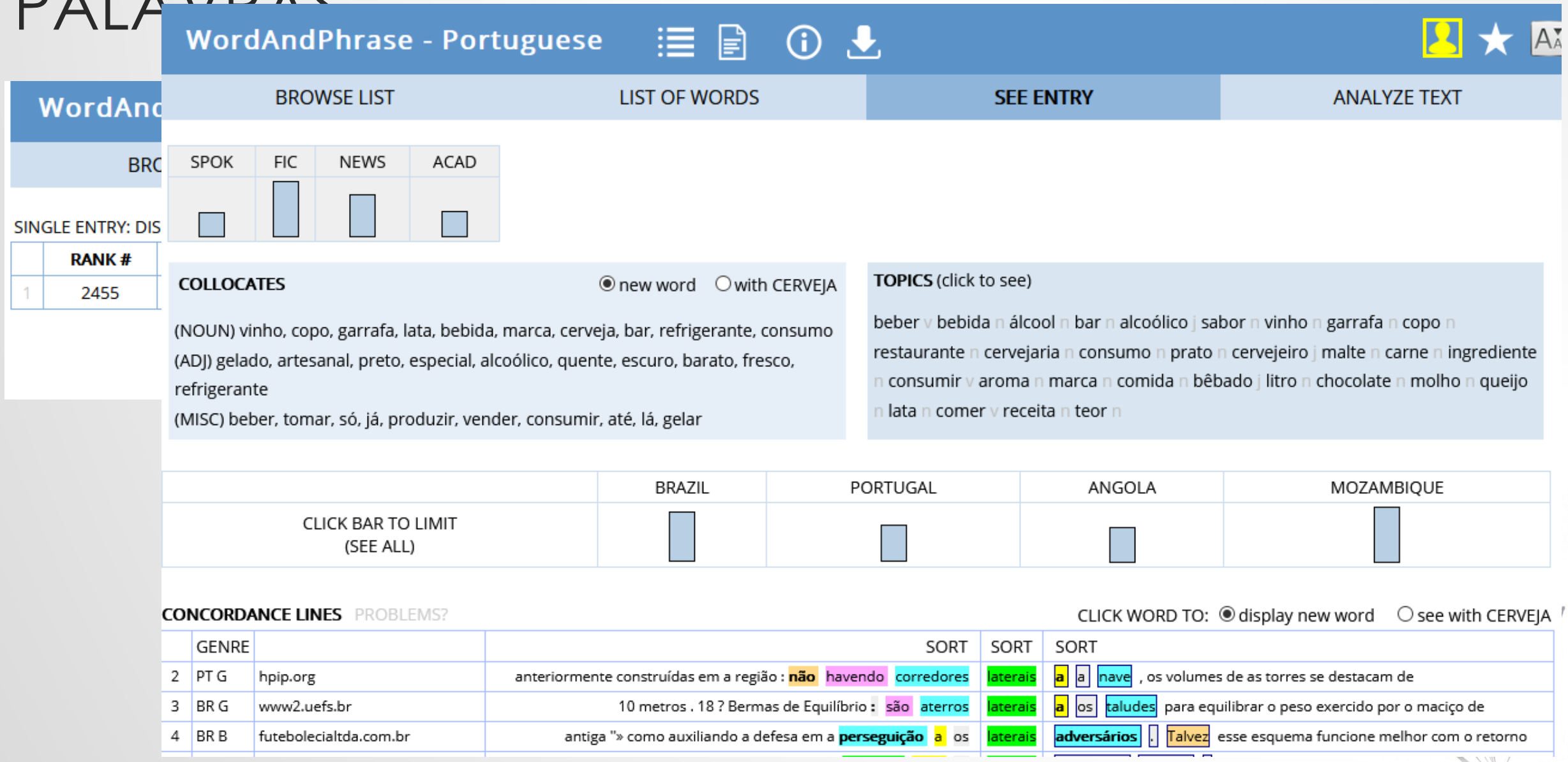
WordAndPhrase - Portuguese

BROWSE LIST | **LIST OF WORDS** | SEE ENTRY | ANALYZE TEXT

SINGLE ENTRY: DISPLAYED IN TAB TO THE RIGHT

	RANK #	PoS	WORD	DEFINITION	MORE DEFIN	IMAGE	TOTAL		ACAD	NEWS	FIC	SPOK
1	2455	N	CERVEJA	beer, (tap)	Ling WR Rev Coll Oxf		30295		13	71	75	8

PALAVRAS



WordAnd

BRC

SINGLE ENTRY: DIS

	RANK #
1	2455

WordAndPhrase - Portuguese

BROWSE LIST | LIST OF WORDS | SEE ENTRY | ANALYZE TEXT

SPOK	FIC	NEWS	ACAD

COLLOCATES ◉ new word ○ with CERVEJA

(NOUN) vinho, copo, garrafa, lata, bebida, marca, cerveja, bar, refrigerante, consumo

(ADJ) gelado, artesanal, preto, especial, alcoólico, quente, escuro, barato, fresco, refrigerante

(MISC) beber, tomar, só, já, produzir, vender, consumir, até, lá, gelar

TOPICS (click to see)

beber v bebida n álcool n bar n alcoólico j sabor n vinho n garrafa n copo n restaurante n cervejaria n consumo n prato n cervejeiro j malte n carne n ingrediente n consumir v aroma n marca n comida n bêbado j litro n chocolate n molho n queijo n lata n comer v receita n teor n

	BRAZIL	PORTUGAL	ANGOLA	MOZAMBIQUE
CLICK BAR TO LIMIT (SEE ALL)				

CONCORDANCE LINES PROBLEMS?

CLICK WORD TO: ◉ display new word ○ see with CERVEJA

	GENRE		SORT	SORT	SORT
2	PT G	hpip.org	anteriormente construídas em a região : **não** havendo corredores	laterais	**a** a nave , os volumes de as torres se destacam de
3	BR G	www2.uefs.br	10 metros . 18 ? Bermas de Equilíbrio : são aterros	laterais	**a** os taludes para equilibrar o peso exercido por o maciço de
4	BR B	futebolecialtda.com.br	antiga "» como auxiliando a defesa em a **perseguição** a os	laterais	**adversários** . Talvez esse esquema funcione melhor com o retorno

PALAVRAS

- Prática 1

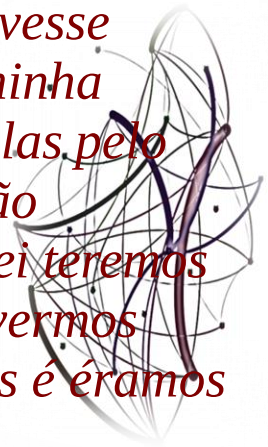
- Prática 2

PALAVRAS

Em **muitos** contextos:

Uma *stopword* pode ser considerada uma **palavra irrelevante** para análise (artigos, preposições).

a ao aos aquela aquelas aquele aqueles aqui aquilo as até aí com como da das de dela delas dele deles depois do dos dá e ela elas ele eles em entre era eram essa essas esse esses esta estamos estas estava estavam este esteja estejam estejamos estes esteve estive estivemos estiver estivera estiveram estiverem estivermos estivesse estivessem estivéramos estivéssemos estou está estávamos estão eu foi fomos for fora foram forem formos fosse fossem fui fôramos fôssemos haja hajam hajamos havemos hei houve houvemos houver houvera houveram houverei houverem houveremos houveria houveriam houvermos houverá houverão houveríamos houvesse houvessem houvéramos houvéssemos há hão isso isto já lhe lhes lá mais mas me mesmo meu meus minha minhas muito na nas nem no nos nossa nossas nosso nossos num numa não nós o os ou para pela pelas pelo pelos por pra qual quando que quem se seja sejam sejamos sem serei seremos seria seriam será serão seríamos seu seus somos sou sua suas são só também te tem temos tenha tenham tenhamos tenho terei teremos teria teriam terá terão teríamos teu teus teve tinha tinham tive tivemos tiver tivera tiveram tiverem tivermos tivesse tivessem tivéramos tivéssemos tu tua tuas tá tém tínhamos um uma vai você vocês vos vou à às é éramos



PALAVRAS

- Prática 3

PALAVRAS

As **stopwords/stoplists** podem ser eliminadas **antes** ou **depois** do processamento/leitura do texto.

Devemos avaliar seu uso para cada aplicação.

Quais das seguintes frases seriam mais informativas?

Rio de Janeiro	Rio Janeiro
Internet das coisas	Internet coisas
Viagem para casa	Viagem casa
Vitamina a	Vitamina

PALAVRAS

```
…
 475 outro
 437 outra
 420 outros
 258 outras
…
 471 homem
 310 homens
…
 145 próprio
 128 própria
  65 próprios
  41 próprias
…
  43 alegria
  25 alegres
  24 alegre
```

Thiago M. R. Dias

thiagomagela@cefetmg.br

CEFET-MG