27e Congrès international de linguistique et de philologie romanes
Nancy, 2013

Universidade do Minho

Procura-PALavras (P-PAL):

Uma aplicação web para uma base de dados lexical do português europeu

Álvaro **Iriarte** S.[1], Ana Paula **Soares**[2], Alberto **Simões**[3], José João de **Almeida**[3], Montserrat **Comesaña**[1] , Ana **Costa**[4], João Filipe **Machado**[4] & Patrícia **França**[2]

[1]Instituto de Letras e Ciências Humanas, [2]Escola de Psicologia, Universidade do Minho, Universidade do Minho, [3]Escola de Engenharia, Universidade do Minho, [4]Centro de Investigação em Psicologia, Universidade do Minho

Projecto PTDC/PSI-PCO/104679/2008 financiado pela Fundação para a Ciência e a Tecnologia (FCT) e co-finaciado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) no âmbito dos programas COMPETE (Programa Operacional Factores de Competitividade) e QREN (Quadro de Referência Estratégico Nacional).

Universidade do Minho

P-PAL procura-palavras

Conteúdos

1. O projeto P-Pal
2. Contextualização
3. Corpus
4. Caraterísticas
5. Interface

Universidade do Minho

projeto Procura-PALvras

aplicação web

métricas lexicais e sublexicais

corpus > 227 milhões de palavras

≈209.000 formas e ≈52.000 lemas

português europeu

Universidade do Minho

projeto Procura-PALvras

N-Watch (Davis, 2005)

BuscaPalabras (Davis & Perea, 2005)

Universidade do Minho

projeto Procura-PALvras

Linguística

Processamento da Linguagem Natural

Psicolinguística

Universidade do Minho

projeto Procura-PALvras

aplicação web, aberta e de acesso livre:

http://p-pal.di.uminho.pt/tools

Contextualização

Universidade do Minho

No PE as bases lexicais existentes são escassas e limitadas:

- Português Fundamental (1984)
  - corpus oral de pequenas dimensões (700,000 palavras), anos 70.

- PORLEX (Gomes & Castro, 2003)
  - I. gráfica, fonológica, fonética, morfo-sintáctica e de vizinhança
  - 29.238 palavras
  - Frequência: ≈5% entradas lexicais

- CORLEX (Bacelar do Nascimento et al, 2000)
  - I. frequência para 26.980 lemas e 140.976 formas, proveniente de um subcorpus do *Corpus de Referência do Português Contemporâneo* (CRPC))
  - Informação morfo-sintáctica

O Corpus

Universidade do Minho

Fontes:

corpora do PE, anotados e disponíveis livremente

LINGUATECA:

Avante	FrasesPP
CETEMPúblico	Museu da Pessoa
DiaCLAV	Natura/Minho
ECI-EE	Vercial

CORLEX

O Corpus

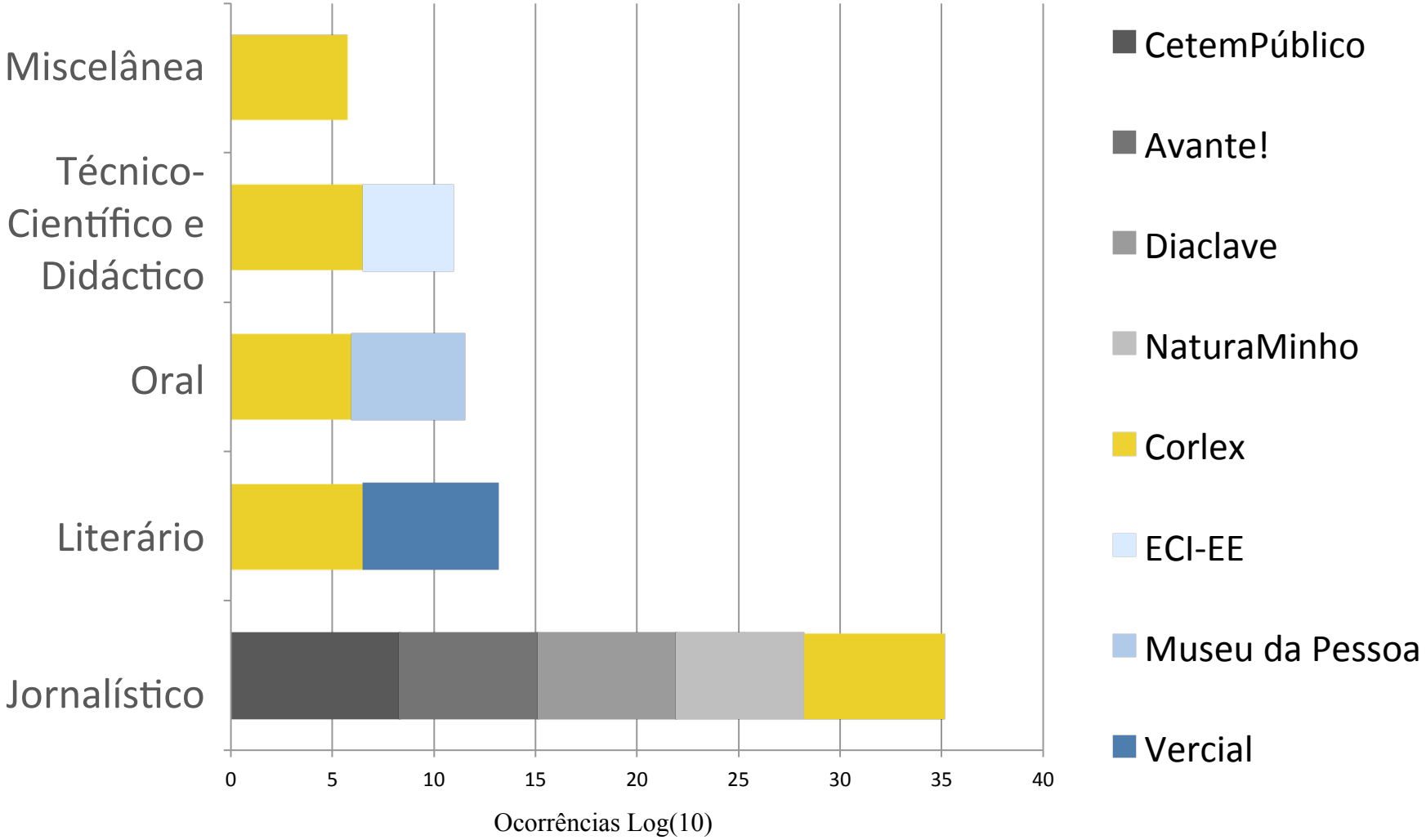


Figura 1: Distribuição dos corpora do P-PAL por género e tipo linguísticos

Tamanho total do corpus – 227.770.752 palavras

Problemas

Universidade do Minho

Diferentes corpora

1. Diferentes sistemas de anotação
2. Diferentes sistemas de lematização

Problemas

Universidade do Minho

Diferentes sistemas de anotação

LINGUATECA			CORLEX		P-PAL		
DET (Artigos Pronomes Adjectivos)	ART	Def.	ART	Def.	DET	Artigo	Def.
		Indef.		Indef.			Indef.
	Relativo					Demonstrativo	
	Interrog.					Possessivo	
						Indef.	
						Relativo	
						Interrogativo	
Pronome Pessoal			PRON	Pessoal	PRON	Pessoal	
Especificador (Pronomes Adjectivos)	Demonst.			Demonst.		Demonstrativo	
	Possessivo			Indefinido		Indefinido	
	Interrogativo			Possessivo		Possessivo	
	Relativo			Interrog.		Interrogativo	
				Relativo		Relativo	

Problemas

Universidade do Minho

Diferentes corpora

1. Diferentes sistemas de anotação
2. Diferentes sistemas de lematização

P-PAL procura-palavras Problemas

Universidade do Minho

Diferentes sistemas de lematização

	Linguateca	Corlex	P-Pal
Nomes	Masc. e fem. singular	Masc. singular	Masc. singular
Adjectivos	Masculino singular (excepto adjectivos com função de nome)	Masc. singular	Masc. singular
Pronomes possessivos, relativos, interrogativos e demonstrativos	Masculino singular (pronomes pessoais têm como lema o pronome pessoal recto: eu é lema de me, ele lema de lhe)	Masculino e feminino singular	Masculino e feminino singular (tal como os determinantes)

Universidade do Minho

Características

(a) obter palavras/lemas que obedeçam a determinados requisitos

(b) analisar palavras/lemas num conjunto requisitos

Universidade do Minho

P-PAL procura-palavras

Características

frequência lexical

informação estrutural

extensão da palavra em letras ou sílabas, divisão silábica, categoria morfo-sintáctica, etc

informação derivada

similaridade ortográfica ou fonológica com outras palavras; bigramas, trigramas, bifones, etc.

informação subjectiva

familiaridade, imaginabilidade, concreteza, ….

P-PAL procura-palavras

Características

frequência lexical

informação estrutural:
morfossintáctica, ortográfica, fonético-fonológica, silábica

informação derivada:
de vizinhança, bigramas, bifones, sílabas

informação subjectiva:
familiaridade, imaginabilidade, concreteza, ….

P-PAL procura-palavras

Características

frequência lexical

informação estrutural:

morfossintáctica, ortográfica, fonético-fonológica, silábica

informação derivada:

de vizinhança, bigramas, bifones, sílabas

informação subjectiva:

familiaridade, imaginabilidade, concreteza, …

Universidade do Minho

Características

frequência lexical

informação estrutural:

morfossintáctica, ortográfica, fonético-fonológica, silábica

informação derivada:

de vizinhança, bigramas, bifones, sílabas

informação subjectiva:

familiaridade, imaginabilidade, concreteza, …

P-PAL procura-palavras

Características

frequência lexical

informação estrutural:

morfossintáctica, ortográfica, fonético-fonológica, silábica

informação derivada:

de vizinhança, bigramas, bifones, sílabas

informação subjectiva:

familiaridade, imaginabilidade, concreteza, …

O Interface


1. Base de dados
P-PAL
SUBTLEX
2. Ferramenta
Obter palavras
Analisar palavras
Pesquisa por:
Lemas
Formas

O Interface

Voltar à aplicação

Expandir filtros Expandir opções

+ Índices de frequência lexical

+ Índices morfo-sintácticos

+ Índices ortográficos

+ Índices fonológicos

+ Índices subjectivos

Submeter

O Interface

☑ Categoria morfo-sintáctica [morf_cat] ⓘ — Filtrar

☐ N ☐ V ☐ ADJ ☐ PREP ☐ CONJ ☐ ADV ☑ PRON ☐ DET ☐ INT ☐ QUANT

☐ Subcategoria morfossintáctica [morf_type] ⓘ — Filtrar

☐ ART_DEF ☐ ART_IND ☐ POSS ☐ SUB ☐ IND ☐ EXIS ☐ NUM_CARD ☐ NUM_MULT ☐ NUM_ORD ☐ NUM_FRAC ☐ COOR ☐ PESS

☑ DEM ☐ INTR ☐ REL ☐ UNI ☐ NONE

☐ Categoria com frequência mais alta [morf_max_cat] ⓘ — Filtrar

☐ Percentagem de ocorrência da categoria mais alta [morf_max_d] ⓘ — Filtrar

☐ Frequência da categoria mais alta [morf_max_freq] ⓘ — Filtrar

☐ Restantes categorias da palavra [morf_others_cat] ⓘ

☐ Percentagem de ocorrência das restantes categorias da palavra [morf_others_d] ⓘ

☐ Frequência das restantes categorias [morf_others_freq] ⓘ

☑ Género [morf_gen] ⓘ — Filtrar

☑ Número [morf_num] ⓘ — Filtrar

☐ Estrangeirismo [morf_est] ⓘ — Filtrar

O Interface

P-PAL procura-palavras

Universidade do Minho

Voltar à aplicação

A mostrar 15 de 25 registos. Resultados **insensíveis** à categoria.

lema ▲	freq_corp_mil	morf_gen	morf_num
a	88046.6292		
aquela	315.9064	Feminino	
aquele	472.7943		
aquilo	172.7038		
as	0.2440	Feminino	Plural
demais	31.8713		
dessas	0.0188	Feminino	Plural
desta	0.0329	Feminino	Singular
destas	0.0422	Feminino	Plural
essa	589.0581		
esse	678.2377		
esta	1890.0555	Feminino	Singular
este	2357.9079		
isso	895.0291		
isto	340.2968		

« Anterior 1 / 2 Seguinte »

Estrutura

- [x] Número de letras [ort_nlet] — Filtrar
  - Média actual: 9.3002 — Mínimo: 4 Máximo: 6
- [ ] Estrutura Consoante/Vogal [ort_cv] — Filtrar
- [ ] Letras repetidas [ort_let_rep] — Filtrar
- [ ] Representação gráfica invertida [ort_inv] — Filtrar

Sílabas

- [x] Número de sílabas ortográficas do estímulo [ort_syl_num] — Filtrar
- [x] Estrutura silábica ortográfica [ort_syl_cv] — Filtrar
- [x] Divisão silábica ortográfica [ort_syl_div] — Filtrar
- [ ] Número de palavras que partilham a estrutura silábica ortográfica do estímulo [ort_syl_cv_tp] — Filtrar
- [ ] Soma da frequência da estrutura silábica ortográfica [ort_syl_cv_tk] — Filtrar
- [ ] Frequência média da estrutura silábica ortográfica [ort_syl_cv_tk_med] — Filtrar
- [ ] Número de palavras que partilham sílabas ortográficas com o estímulo [ort_syl_p_tp] — Filtrar
- [ ] Média das palavras que partilham sílabas ortográficas com o estímulo [ort_syl_p_tp_med] — Filtrar
- [ ] Soma das frequências das sílabas ortográficas [ort_syl_p_tk] — Filtrar
- [x] Média da frequência das sílabas ortográficas [ort_syl_p_tk_med] — Filtrar

O Interface

P-PAL procura-palavras

Universidade do Minho

Voltar à aplicação

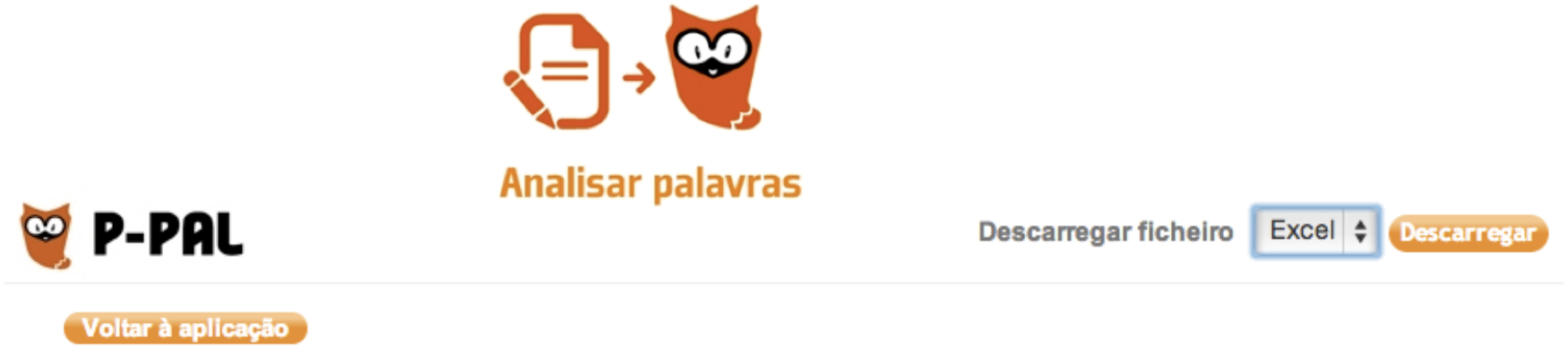
A mostrar 15 de 8543 registos. Resultados **insensíveis** à categoria.

lema ▲	freq_corp_mil	ort_nlet	ort_syl_num	ort_syl_cv	ort_syl_div	ort_syl_p_tk_med
ábaco	0.1502	5	3	V-CV-CV	'á|ba|co	12.7305
abada	0.0375	5	3	V-CV-CV	a|'ba|da	19.0833
abade	3.3462	5	3	V-CV-CV	a|'ba|de	20.0955
abadia	0.9996	6	4	V-CV-CV-V	a|ba|'di|a	5.3473
abafar	7.2791	6	3	V-CV-CVC	a|ba|'far	19.7478
abafo	0.1596	5	3	V-CV-CV	a|'ba|fo	19.9328
abaixo	61.8278	6	3	V-CVV-CV	a|'bai|xo	20.0443
abajur	0.0657	6	3	V-CV-CVC	a|ba|'jur	20.2122
abalar	19.0824	6	3	V-CV-CVC	a|ba|'lar	17.8825
abalo	4.9607	5	3	V-CV-CV	a|'ba|lo	18.7072
abanão	0.9433	6	3	V-CV-CVV	a|ba|'não	20.1245
abanar	6.5564	6	3	V-CV-CVC	a|ba|'nar	17.9997
abano	0.4130	5	3	V-CV-CV	a|'ba|no	19.2481
abanto	0.0047	6	3	V-CVC-CV	a|'ban|to	18.6200
abarca	0.1079	6	3	V-CVC-CV	a|'bar|ca	18.5969

« Anterior 1 / 570 Seguinte »

O Interface


Obter palavras
P-PAL
Voltar à aplicação
Escolha o ficheiro a carregar
Choose File
No file chosen
Expandir opções


Analisar palavras
P-PAL
Descarregar ficheiro
Excel
Descarregar
Voltar à aplicação

Universidade do Minho

projeto Procura-PALvras

http://p-pal.di.uminho.pt/

Universidade do Minho

Obrigado!

Ana Paula Soares
Álvaro Iriarte S.
Alberto Simões
José João de Almeida
Montserrat Comesaña
Ana Costa
João Filipe Machado
Patrícia França

http://p-pal.di.uminho.pt/

Fundação para a Ciência e a Tecnologia
UNIÃO EUROPEIA
FEDER
COMPETE
QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL

Projecto PTDC/PSI-PCO/104679/2008 financiado pela Fundação para a Ciência e a Tecnologia (FCT) e co-financiado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) no âmbito dos programas COMPETE (Programa Operacional Factores de Competitividade) e QREN (Quadro de Referência Estratégico Nacional).