II Jornadas Internacionais Per-Fide – Corpora e Terminologia
I 6-7 Junho, 2013 I Braga: Universidade do Minho, Portugal

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

Procura-PALavras (P-PAL):

Da constituição do corpus à base lexical

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

[1]Escola de Psicologia, Universidade do Minho, [2]Instituto de Letras e Ciências Humanas, 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, Escola Psicologia
Grupo de Investigação Psicolinguística

Conteúdos

1. Contextualização
2. Aplicações
3. Caracterização
4. O corpus
5. Problemas
    a) O sistema de anotação
    b) O sistema de lematização
6. O interface

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

Contextualização

- Construção de uma ferramenta rápida e versátil que disponibilize propriedades objectivas (lexicais e/ou sublexicais) e subjectivas das palavras do Português Europeu (PE)
- A investigação depende da análise, controlo e/ou manipulação das propriedades linguísticas em tarefas de desempenho
- Ausência de bases lexicais põe em causa a realização de estudos com falantes do Português Europeu

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

Contextualização

No PE as bases lexicais existentes são escassas e limitadas:

- PORLEX (Gomes & Castro, 2003)
  - Informação lexical de tipo gráfico, fonológico, fonético, morfo-sintáctico e de vizinhança para um total de 29.238 palavras
  - Informação de frequência para ≈5% das entradas lexicais (importadas de um corpus oral de pequenas dimensões - Português Fundamental)

- CORLEX (Bacelar do Nascimento et al, 2000)
  - Informação de frequência absoluta para 26.980 lemas e 140.976 formas (extraídas de um sub-*corpus* do *Corpus* de Referência do Português Contemporâneo (CRPC) do Centro de Linguística da Universidade de Lisboa)
  - Informação morfo-sintáctica

Caracterização

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

- **frequência lexical**: ocorrência das palavras por milhão de palavras, logarítmica (base 10)
- **informação estrutural**: informação de tipo linguístico determinada pela análise da própria palavra (ex. extensão da palavra em letras e fones, classe gramatical)
- **informação derivada**: informação que resulta da análise da relação da palavra com as restantes do léxico tanto a nível lexical (ex. vizinhança), como sublexical (ex. bigramas, bifones, sílabas)
- **informação subjectiva**: informação que reflecte as experiências dos indivíduos com o uso da própria língua (ex. familiaridade, imaginabilidade, concreteza, valência)

Aplicações

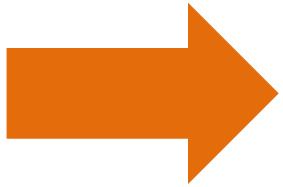
- **Linguística**: estudos sobre a língua (ex., análise empírica das características ortográficas, fonológicas e morfo-sintácticas do PE)
- **PLN**: construção de ferramentas lexicográficas e instrumentos de análise linguística (ex. silabificação)
- **Psicolinguística**: Estudo da arquitectura funcional e do processamento linguístico humano

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

O Corpus

Critérios de selecção:

- Corpora existentes do PE
- Disponíveis livremente
- Anotados



Linguateca

Corlex

O Corpus

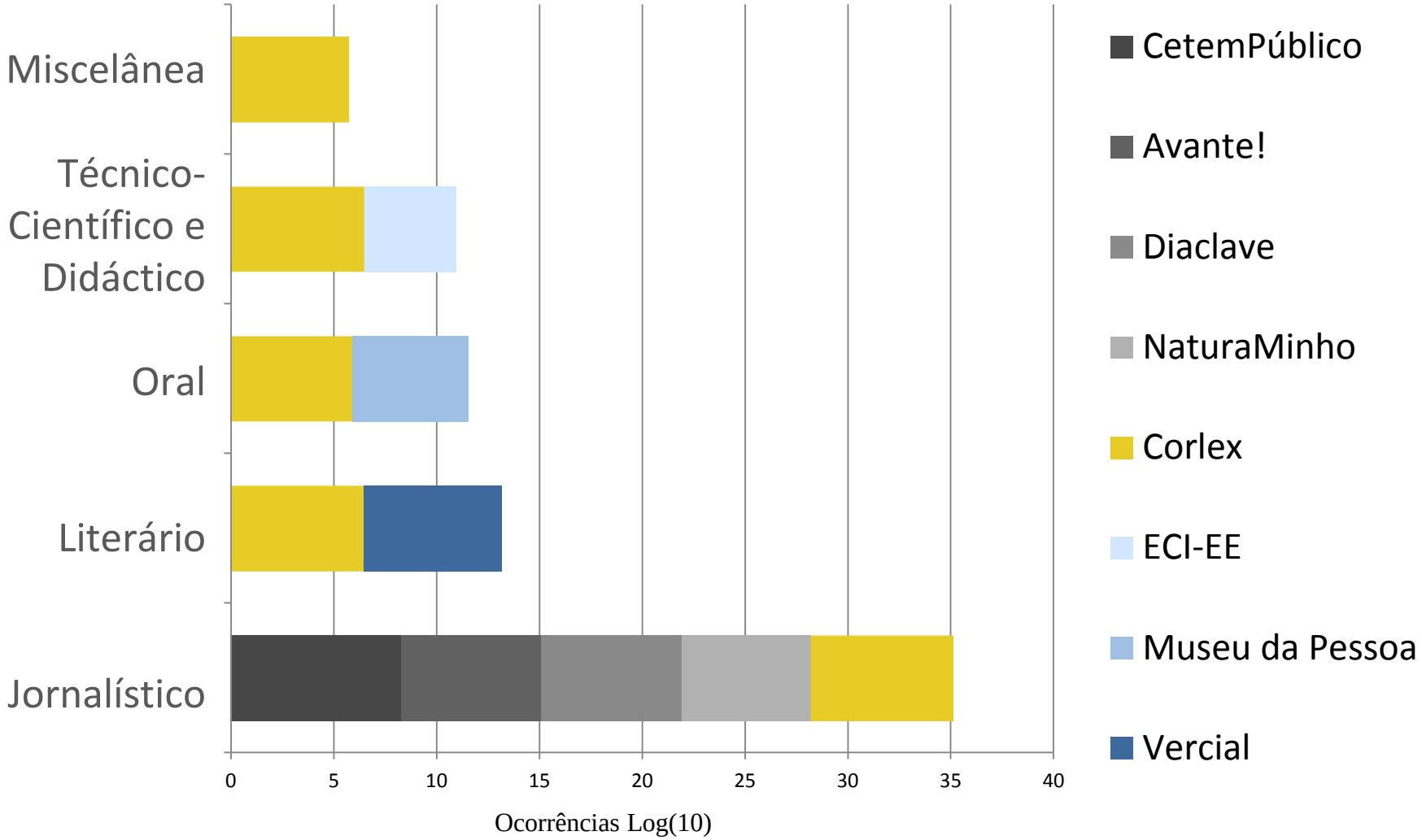


Figura 1: Distribuição dos corpora do P-PAL por género e tipo linguísticos

Tamanho total do corpus – 227.770.752 palavras

Problemas

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

Diferentes corpora

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

<table>
<tr><th colspan="3">Corpora da Linguateca</th><th colspan="2">CORLEX</th><th colspan="3">P-PAL</th></tr>
<tr><td colspan="3">Nomes próprios e comuns</td><td colspan="2">Nome</td><td colspan="3">Nome</td></tr>
<tr><td rowspan="3">DET (Artigos
Pronomes
Adjectivos)</td><td rowspan="2">ART</td><td>Def.</td><td rowspan="3">ART</td><td rowspan="3">Def.
Indef.</td><td rowspan="3">DET</td><td>Artigo</td><td>Def.
Indef.</td></tr>
<tr><td>Indef.</td></tr>
<tr><td colspan="2">Relativo
Interrog.</td><td colspan="2">Demonstrativo
Possessivo
Indef.
Relativo
Interrogativo</td></tr>
<tr><td colspan="3">Pronome Pessoal</td><td rowspan="2">PRON</td><td rowspan="2">Pessoal
Demonst.
Indefinido
Possessivo
Interrog.
Relativo</td><td rowspan="2">PRON</td><td colspan="2" rowspan="2">Pessoal
Demonstrativo
Indefinido
Possessivo
Interrogativo
Relativo</td></tr>
<tr><td>Especificador
(Pronomes
Adjectivos)</td><td colspan="2">Demonst.
Possessivo
Interrogativo
Relativo</td></tr>
</table>

<table>
<tr><th colspan="3">Corpora da Linguateca</th><th>CORLEX</th><th colspan="2">P-PAL</th></tr>
<tr><td colspan="3">Adjectivo</td><td>Adjectivo</td><td colspan="2">Adjectivo</td></tr>
<tr><td rowspan="2">Verbo</td><td>Principal</td><td>Intransit.
Transitivo
Transit. dir.</td><td rowspan="2">Verbo</td><td colspan="2" rowspan="2">Verbo</td></tr>
<tr><td>Copulativo</td><td>vK e vtK</td></tr>
<tr><td colspan="3">Preposição</td><td>Preposição</td><td colspan="2">Preposição</td></tr>
<tr><td colspan="3">Advérbio</td><td>Advérbio</td><td>Advérbio</td><td>Interrog.</td></tr>
<tr><td>Conjunção</td><td colspan="2">Subordinativa
Coordenada</td><td>Conjunção</td><td rowspan="2">Conjunção</td><td>Subordinativa</td></tr>
<tr><td colspan="3">Interjeição</td><td>Interjeição</td><td colspan="2" rowspan="3">Interjeição</td></tr>
<tr><td colspan="3">Contracção</td><td>Contracção</td></tr>
<tr><td colspan="3">Itens multilexicais (não hifenizados)</td><td>Elemento de Locução</td></tr>
</table>

Tabela 1: O sistema morfo-sintáctico do P-Pal

Problemas

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

Diferentes corpora

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

	Linguateca	Corlex	P-Pal
Nomes	Masc. e fem. singular	Masc. singular	Masc. singular
Preposições	Invariável	Invariável	Invariável
Conjunções	Invariável	Invariável	Invariável
Advérbios	Invariável	Invariável	Invariável
Interjeições	Invariável	Invariável	Invariável
Verbos	Infinitivo Pessoal	Infinit. pessoal	Infinit. pessoal
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)

	Linguateca	Corlex	P-Pal
Palavras compostas hifenizadas	Masc. sing. ou invariáveis	Inconsistente: - abelha-mãe pertence ao lema abelha; abelha-mestra tem entrada própria - à-vontade é lema de à e vontade	- Masc. singular ou invariáveis - Palavras formadas por derivação prefixal com alteração de significado, referente ou classe gramatical
Itens multilexicais não hifenizados	Masc. sing. e invariáveis com "="	Não incluídos	Decompostas: frequência somada a cada constituinte
Estrangeirismos	Singular	Singular	Singular

Tabela 2: A integração das bases lexicais de lemas

O Interface

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

O Interface

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

Voltar à aplicação

Expandir filtros Expandir opções

+ Índices de frequência lexical

+ Índices morfo-sintácticos

+ Índices ortográficos

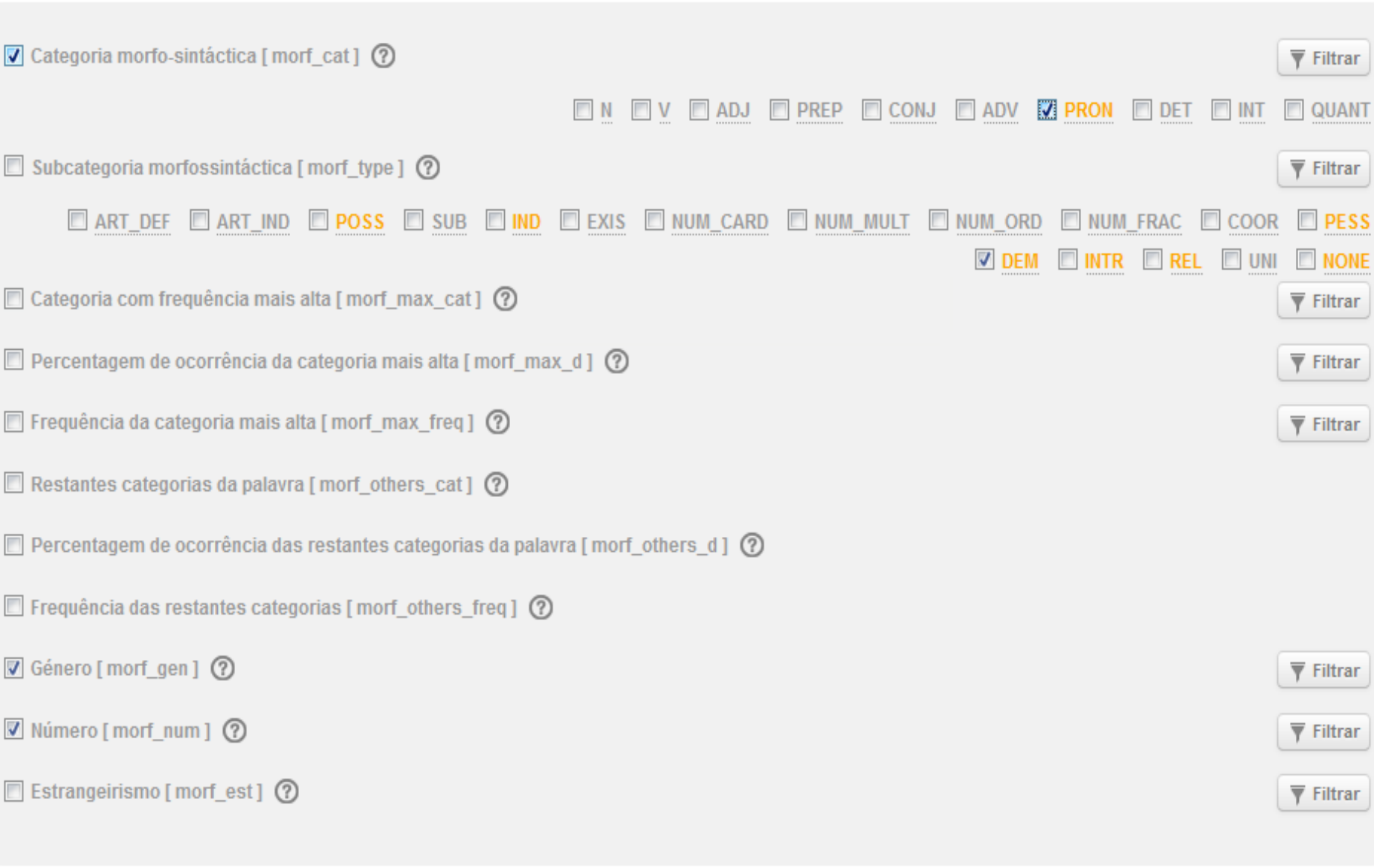
+ Índices fonológicos

+ Índices subjectivos

Submeter

O Interface

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística



P-PAL procura-palavras O Interface

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

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

Universidade do Minho, Escola Psicologia
Grupo de Investigação Psicolinguística

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 »

Obrigada!

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

FCT 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-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).