

10 Anos de Linguateca (PROPOR 2008)

Aveiro, 11 de Setembro de 2008

# CONVERSOR DE GRAFEMAS PARA FONES BASEADO EM REGRAS PARA PORTUGUÊS

Sara Candeias  
Fernando Perdigão

INSTITUIÇÕES ASSOCIADAS:



INSTITUTO  
SUPERIOR  
TÉCNICO



Faculdade de Ciências  
e Tecnologia da  
Universidade de Coimbra



universidade  
de aveiro



Inovação



instituto de  
telecomunicações

*creating and sharing knowledge for telecommunications*

© 2005, IT - instituto de telecomunicações. Todos os direitos reservados.

## Sumário

### → Sistema de conversão Gr2Ph

· Desenvolvimento

· Teste

· Avaliação

### → Sugestões / Desafios

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Objectivo Final

## → Sistema de conversão Gr2Ph

- : conversão de unidades acentuais em fones de forma a definir a sequência de modelos acústicos para um sistema de reconhecimento automático de fala

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



instituto de  
telecomunicações

## Sumário

# → Sistema de Conversão Gr2Ph

- Desenvolvimento

- Teste

- Avaliação

INSTITUIÇÕES ASSOCIADAS:



PROPOR 2008



# Sistema de Conversão Gr2Ph

## ■ Desenvolvimento

### Sistemas Intermédios

- Segmentação silábica
- Marcação de sílaba tónica

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Sistema de Conversão Gr2Ph

- Desenvolvimento

→ **Recurso: Linguateca / UMinho** (Projecto NATURA)

- Lista de 680 000 unidades acentuais (*spelling*)
  - papel fundamental no desenvolvimento dos algoritmos de processamento de linguagem natural
- Análise e verificação de regras
  - Sequência VC's
  - Divisão silábica
  - Marcação de sílaba tónica

INSTITUIÇÕES ASSOCIADAS:



PROPOR 2008



instituto de  
telecomunicações

# Sistema de Conversão Gr2Ph

· **Desenvolvimento**

· **Recurso (Linguateca / UMinho)**

**Projecto NATURA**

**Sistemas Intermédios**

- Segmentação silábica

18 padrões de sequências de grafemas a formar sílaba em português

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Sistema de Conversão Gr2Ph

· **Desenvolvimento**

· **Recurso (Linguateca / UMinho)**

**Projecto NATURA**

**Sistemas Intermédios**

- Segmentação silábica

1 sequência de 1 segmento  
3 sequência de 2 segmentos  
5 sequência de 3 segmentos  
6 sequência de 4 segmentos  
3 sequência de 5 segmentos

18 padrões de sequências de grafemas a formar sílaba em português

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Sistema de Conversão Gr2Ph

· **Desenvolvimento**

· **Recurso (Linguateca / UMinho)**

**Projecto NATURA**

**Sistemas Intermédios**

• Segmentação silábica

1 sequência de 1 segmento  
3 sequência de 2 segmentos  
5 sequência de 3 segmentos  
6 sequência de 4 segmentos  
3 sequência de 5 segmentos

V  
VV, CV, VC  
CVV, VCC, CVC, CCV, VVC  
CVVC, CVCC, VCVC, CCVV, CCVC, CCCV  
CCVVC, CCVCC, CVCCC

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Sistema de Conversão Gr2Ph

· Desenvolvimento

· Recurso (**Linguateca / UMinho**)

Projecto NATURA

**Sistemas Intermédios**

• Segmentação silábica

*á.gua*

*eu, ca.sa,...*

*pai, abs.trair,...*

*mães, subs.crever,...*

*grãos, trans.crever,...*

V

VV, CV, VC

CVV, VCC, CVC, CCV, VVC

CVVC, CVCC, VCVC, CCVV, CCVC, CCCV

CCVVC, CCVCC, CVCCC

INSTITUIÇÕES ASSOCIADAS:



PROPOR 2008



# Sistema de Conversão Gr2Ph

■ **Desenvolvimento**

Análise e verificação  
de regras

**DIFICULDADES  
ENCONTRADAS**

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Sistema de Conversão Gr2Ph

## : Desenvolvimento

### anotação fonética

- SAMPA ?
- recurso a extensões ?
  - [anEI] ou [anEI~] ?
  - [dadu] ou [daDu] ?
- anotação dos ditongos crescentes?
  - [suav@] ou [swav@] ?
  - [awrius] ou [awrijus] ?

**DIFICULDADES  
ENCONTRADAS**

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



## Sumário

# → Sistema de Conversão Gr2Ph

- Desenvolvimento

- **Teste**

- **Avaliação**

VALIDAÇÃO DO SISTEMA

INSTITUIÇÕES ASSOCIADAS:



PROPOR 2008



# Sistema de Conversão Gr2Ph

· Teste

· Avaliação

**PROBLEMAS  
ENCONTRADOS**

VALIDAÇÃO DO SISTEMA

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



# Sistema de Conversão Gr2Ph

## → Teste e Avaliação

- Corpus do SpeechDat (**15 209 unidades acentuais**)
- Problemas:
  - “e.”+ «palatal» [L] | [J] → [6] - [e]
    - [ @Sp6Lu] – [ @SpeLu] ; [f@R6Ju] - [f@ReJu]
  - “ex+ V” → [i] – [e]
    - [ize~plu] - [eze~plu]; [izib@] – [ezib@]
  - [E] – [e] em sílaba tónica
    - + [r]: [muLEr], [n6Sser]
  - [O] – [o] em sílaba tónica
    - + [j]: [bOjn6], [k6lojru]
  - Siglas
  - Prefixos e sufixos (múltipla acentuação)
- Cerca de 66% de concordância

INSTITUIÇÕES ASSOCIADAS:



PROPOR 2008



## Sumário

→ Sistema de Conversão Gr2Ph

→ **Sugestões / Desafios**

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



## Sugestões / Desafios

**CRIAÇÃO DE RECURSOS E DE FERRAMENTAS  
LIGADOS À FONÉTICA DO PORTUGUÊS  
AINDA NÃO DISPONÍVEIS**

→ **Dicionário fonético de domínio público**

→ **Aplicação Gr2Ph *on-line***

⋮ Criação de directivas para a anotação fonética e avaliação

INSTITUIÇÕES ASSOCIADAS:



**PROPOR 2008**



Obrigada

Fim

INSTITUIÇÕES ASSOCIADAS:



INSTITUTO  
SUPERIOR  
TÉCNICO



Faculdade de Ciências  
e Tecnologia da  
Universidade de Coimbra



universidade  
de aveiro



Inovação



SIEMENS  
Communications



instituto de  
telecomunicações

*creating and sharing knowledge for telecommunications*

Ciência.Inovação  
2010

Programa Operacional Ciência e Inovação 2010

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR



© 2010, IT - Instituto de Telecomunicações. Todos os direitos reservados.