

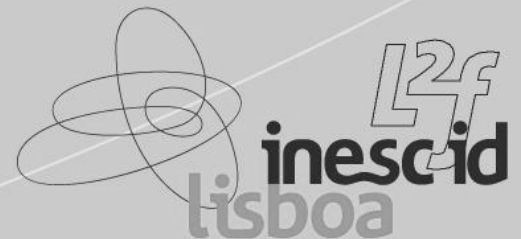
# **Overview of Speech Corpora for European Portuguese**

**Astrid Hagen, Isabel Trancoso**  
**Spoken Language Systems Laboratory**



INSTITUTO  
SUPERIOR  
TÉCNICO

INESC-ID Lisboa, Institute for Systems and Computer Engineering: Research and Development  
L<sup>2</sup>F, Spoken Language Systems Laboratory



# Overview of Speech Corpora for European Portuguese

- Identification of existing speech resources (BDFALA, BD-PUBLICO, SPEECHDAT, ALERT, ...)
- Presentation of the major speech corpora regarding:
  - Partners
  - (Eagles) Typology: Linguistic contents, speaker characteristics, data collection, annotation
  - Research efforts
- Dissemination of Databases
- Recent speech recognition results on Portuguese SpeechDat
  - Using the Reference Recognizer (RefRec)
  - Using our HMM/MLP Hybrid system

# EUROM/SAM


Ontem à noite abri a porta de casa para deixar o gato sair. Estava uma noite tão boa que resolvi dar uma voltinha pelo jardim e apanhar ar. Mal saí, ouvi o ruído da porta a fechar-se atrás de mim. Lembrei-me então que não tinha as chaves comigo. O pior foi que me apanharam a arrombar a porta e fui parar à cade



- Partners: collected w/in SAM\_A European project (INESC / CLUL)
- Typology
  - Linguistic contents: 4 types of material: CVC (121 diff. logatomes), selected numbers, 40 short passages, 50 filler sentences.
  - Number and type of speakers: 3 target corpora: "Many talkers", "few talkers" and "very few talkers" corpus; wide range of age groups; mainly Lisbon accent.
  - Data collection: anechoic room; well monitored; 2.6 Gb.
  - Annotation: orthographic transcription, broad phonetic transcription.
- Research efforts: speech recognition and synthesis, phonetic coding.

# BDFALA

Estás com fome? Vou arranjar-te  
qualquer coisa para comer.

- Partners: INESC and CLUL (Centro de Linguística da Universidade de Lisboa); national project. 
- Typology
  - Linguistic contents: 6 types: logatomes; isolated words, sentences for prosodic studies; phonetically rich sentences; phonetically-complete paragraphs; read paragraphs (from TV debates).
  - Number and type of speakers: 4 male, 4 female speakers; betw. 20 and 50 years old
  - Data collection: sound-proof room; (self-)monitoring; 2.4 Gb.
  - Annotation: orthographic transcription; pronunciation lexicon.
- Research efforts: main goal: enlarging EUROM corpus.

# BD-PUBLICO

Mas o presidente da Câmara de Representantes, o papa da revolução conservadora republicana, Newt Gingrich deu claramente a entender que não faria o mesmo.

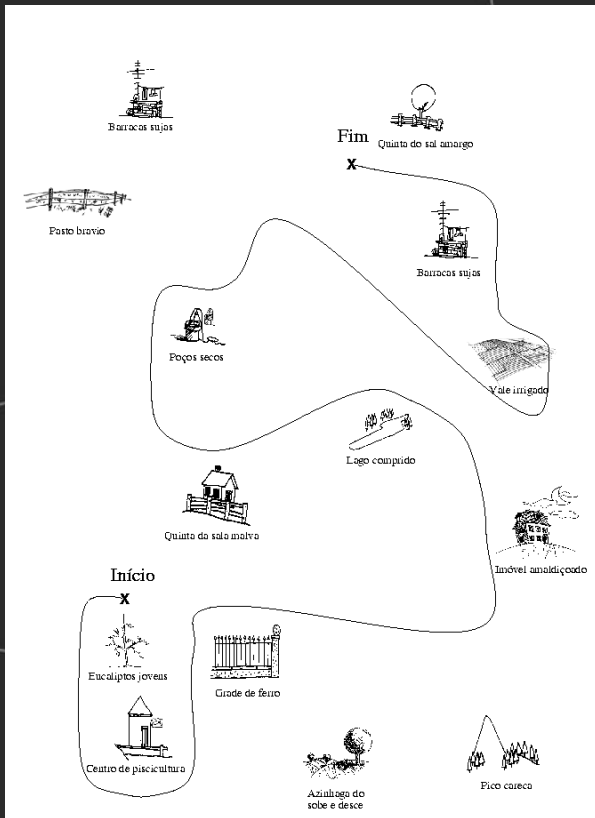
- Partners: INESC, IST (Instituto Superior Técnico), PUBLICO national newspaper; national project (PRAXIS XXI), internat. project: SPRACH.
- Typology
  - Linguistic contents: newspaper text (6 months; 10M words; 156k different words); 15 speaker-adaptation and 3 calibration sentences (phonetically rich).
  - Number and type of speakers: (under)graduate students from IST, betw. 19 and 28 years old; broad coverage of accents. 120 speakers.
  - Data collection: sound-proof room; 2 Gb.
  - Annotation: pronunciation lexicon w/ phonemic transcription for each word (hand-corrected automatic pronunciation).
- Research efforts: develop domain independent AMs, pronunciation dictionaries, and LMs, i.e. domain independent ASR



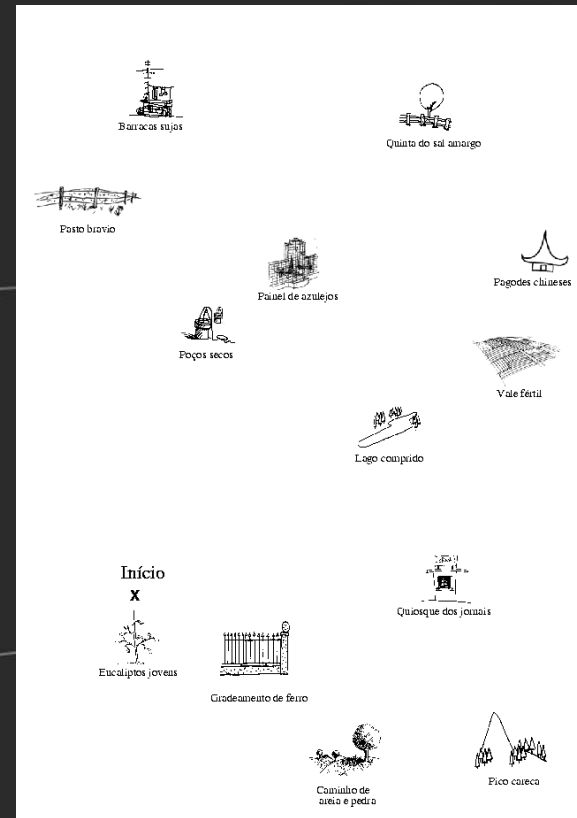
# CORAL

- Partners: INESC, CLUL, FLUL, FCSH-UNL; national project (PRAXIS XXI).
- Typology
  - Linguistic contents: spoken dialogues (giving map directions)
  - Number and type of speakers: 32 speakers; (under)graduate students from Lisbon area.
  - Data collection: sound-proof room, close-talking microphones; no visual contact betw. speakers, no monitoring; 64 unrestricted, spontaneous dialogues. Dual-channel: one for each speaker.
  - Annotation: orthographic, phonetic, phonological, syntactic and semantic labelling of part of corpus.
- Research efforts: study of connected speech phenomena (e.g. sequences of plosives across word boundaries)

# CORAL



Sim. O pasto bravio fica-te à tua esquerda, {p}



{ph/6t"6~w~=Então} {ip|p"Er6=espera}, vou fazer  
vou fazer {pp} um caminho da quinta {ct|p"O=pa

# ALERT



- Partners: INESC-ID, RTP (data provider & collector)
- Typology
  - Linguistic contents: national and regional news programs
  - Number and type of speakers: > 1376 speakers (professional news speakers, reporters and amateurs)
  - Data collection:
    - Speech Recognition Corpus: 122 news programs (76h audio data)
    - Topic Detection Corpus: 133 evening news programs (300h audio data)
  - Annotation: following LDC Hub4 (Broadcast speech) transcription conventions
- Research efforts: BN speech recognizer, system for topic segmentation and indexing.



# Audioling (Sound and Pronunciation)

- Partners:

- CLUL – Centro de Linguística da Universidade de Lisboa (PT)
- Universiteit Gent, Romaanse Taalkunde (BE)
- Université Charles-de-Gaulle, LILLE III (FR)
- INESC – Instituto de Engenharia de Sistemas e Computadores (PT)
- LIDEL - Edições Técnicas (PT)

- Typology

- Linguistic contents: multimedia didactic material to teach Portuguese to intermediate and advanced learners; words, sentences, texts for oral understanding, phonetic curves (F0).
- Number and type of speakers
- Data collection
- Annotation: phone-level automatic annotation?? Etc.??

# Audioling (Sound and Pronunciation)

- Research efforts: automatic F0 detection
- Goals: self teaching of Portuguese.

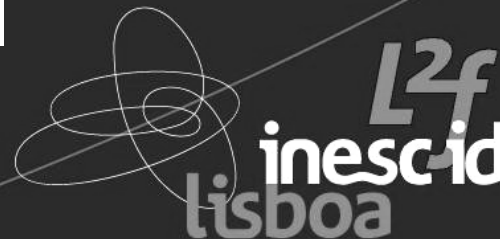


Os próprios estrangeiros que nos visitam também sentem que Portugal é muito mais do que ter uma boa comida e algum sol no Algarve, é sobretudo uma forma de estar e de comunicar.



INSTITUTO  
SUPERIOR  
TÉCNICO

L<sup>2</sup> F — Spoken Language Systems Laboratory



# IPSOM: Indexing, Integration and Sound Retrieval in Multimedia Documents



- Partners: national project IPSOM
  - INESC-ID
  - “Large Scale Informatics Laboratory” of the University of Lisbon (Faculty of Science, Informatics Department)
  - National Library
- Typology
  - Linguistic contents: spoken books (e.g. For the visual impaired) such as “O Senhor Ventura” by Miguel Torga
  - Number and type of speakers: several professional speakers
  - Data collection: high-quality DAT recordings, 2h 15m.
  - Annotation: phonetic transcription (grapheme-to-phone conversion of the lexicon, then hand-correction, followed by automatic alignment).
- Research efforts: provide improved access to spoken information; spoken interface; detection and indexing of units in spoken books.

# SPEECHDAT

- Partners: European SpeechDat project: INESC under subcontract w/ Portugal Telecom
- Typology
  - Linguistic contents: isolated and continuous speech; spontaneous answers to 7 questions; prompt sheet of 33 items: read speech (e.g. application words, connected digits, dates, currency amounts, yes/no questions, time phrases, phonetically rich words and sentences).
  - Number and type of speakers: 2 phases (1000 and 4000 speakers); employees of Portugal Telecom: large demographic coverage; 16 to 60 years old; 47% male, 53% female speakers.
  - Data collection: responsibility of INESCTEL; via telephone network (on location), spoken prompts.
  - Annotation: ascii file containing information on calling session, recording conditions, date and time, speaker info, prompt, orthographic transcription etc.; pronunciation lexica (automatically created and hand-corrected).

Pedir-lhe-emos agora que leia a coluna da direita da seguinte lista:

Pedir-lhe-emos agora que leia a coluna da direita da seguinte lista:

1. Leia o número algarismo a algarismo	3 6 4 8 2
2. Leia a frase	A derrota veio num golo que teve um remate muito bonito.
3. Leia o nome da cidade ou vila	Edimburgo
4. Soletre a palavra (letra a letra)	E, D, I, M, B, U, R, G, O
5. Leia a frase	Pincele tudo com uma gema de ovo misturada com uma colher de sopa de água.
6. Leia as horas	onze horas e cinco minutos
7. Leia a palavra	operador
8. Leia a quantia em dinheiro	18.362\$00

11. Leia a frase	Eu queria telefonar
12. Leia o número por extenso	19.395
13. Leia a frase	O deputado participou, em sessenta e um, na Operação Dulcinea, conduzida por Henrique Galvão.

27. Leia o número	zero
28. Leia a palavra	conferência
29. Leia a frase	O estado apostou sem risco e embolsou mais de dez milhões de contos
30. Leia o código pessoal	1 4 1 4 2 0
31. Leia a palavra	sopro
32. Leia o número algarismo a algarismo	9 0 5 2 7 3 1 8 4 6

25. Leia a frase	O avião que sai às três faz escala em Copenhague.
26. Leia a palavra	metralhadora
27. Leia o número	zero
28. Leia a palavra	conferência
29. Leia a frase	O Estado apostou sem risco e embolsou mais de dez milhões de contos.
30. Leia o código pessoal	1 4 1 4 2 0
31. Leia a palavra	sopro
32. Leia o número algarismo a algarismo	9 0 5 2 7 3 1 8 4 6
33. Leia a frase	O comandante militar respondia desta forma tanto ao bispo de Tinnor, como ao governador Mário Carascalão.

M,O,N,T,A,L,E,G,R,E



**L<sup>2</sup>f**  
**inescid**  
**lisboa**



INSTITUTO  
SUPERIOR  
TÉCNICO

L<sup>2</sup>F — Sp

# SPEECHDAT

- Research efforts:
  - Robust speech recognizers

# Dissemination of Databases

- Databases available via
  - ELRA
  - Exchange for other corpora

# Recent speech recognition results on Portuguese SpeechDat

- Using the SpeechDat reference recognizer
- Using an HMM/MLP hybrid recognizer



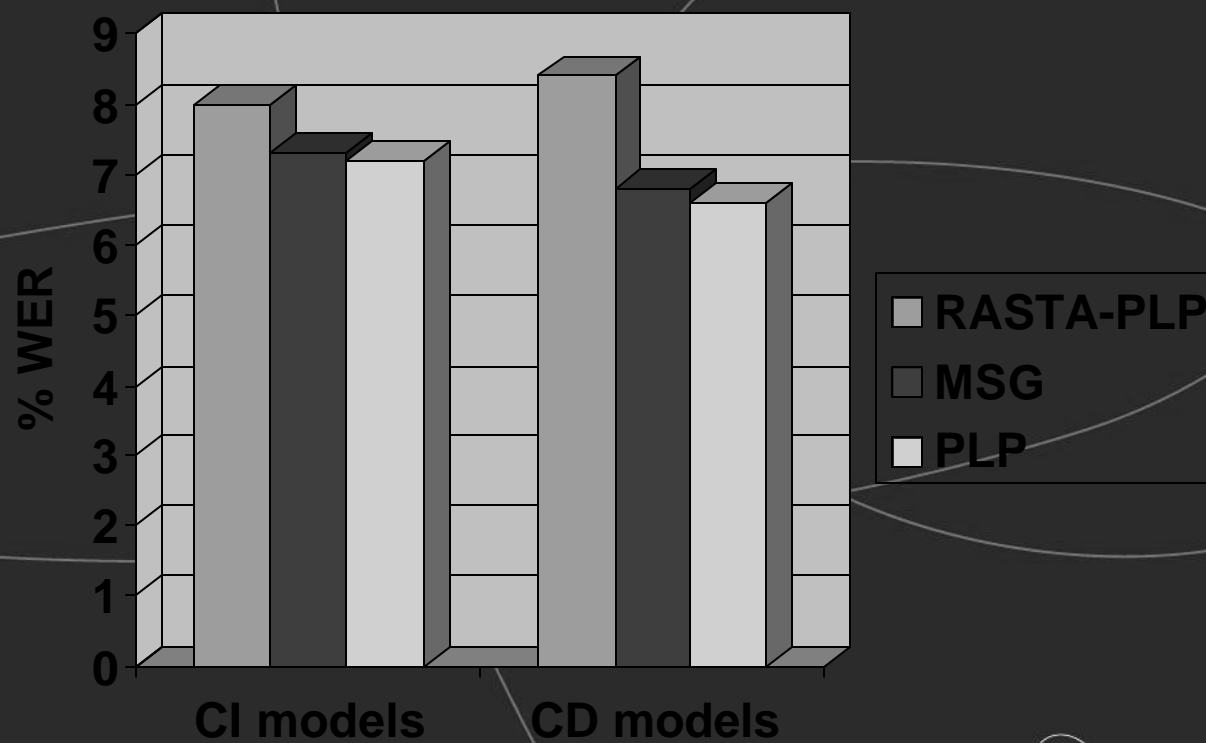
# ASR System (contd.)

- Acoustic Modelling
  - Speech features used: 12 Perceptual Linear Prediction (PLP) cepstra, 12 RASTA-PLP cepstra, and 28 Modulation Spectrogram (MSG) features.
  - Alignment created with Gaussian models, using flat start, then re-aligned with Multi-Layer Perceptron (MLP) several times
  - MLP characteristics:
    - MLP: 7 frames contextual information, 2000 hidden nodes, # output nodes = # speech units
    - Different speech units used:
      - 31 context-independent monophone models ('s' 'af' 'y' 'ch')
      - 151 context-dependent triphone models ('?-s-af' 's-af-y' ...)

# Automatic Speech Recognition System

- Vocabulary and Language Modelling
  - Vocabulary consists of 51 words  
(comprising digits, natural numbers and the 2 females forms: 'uma', 'duas')
  - Language Model was set up on training utterances using the CMU-Cambridge Language Modelling Toolkit (V2.05):  
2601 bigrams

# Experiments and Results



## Comparison to other Speech Corpus

	RASTA-PLP	MSG	PLP
	CD (CI)	CD (CI)	CD (CI)
Portuguese SpeechDat	8.4 (8.0)	6.8 (7.3)	6.6 (7.2)
OGI Numbers Corpus	6.8*	9.8*	7.1°

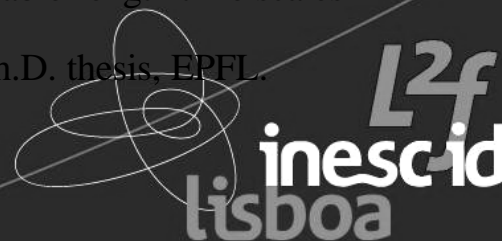
\* Wu, Kingsbury, Morgan, Greenberg “Incorporating information from syllable-length time scales into ASR”, ICSLP, 1:721-724, 1998.

° Hagen, “Robust speech recognition based on multi-stream processing”, Ph.D. thesis, EPFL, Switzerland, 2001.



INSTITUTO  
SUPERIOR  
TÉCNICO

L<sup>2</sup> F— Spoken Language Systems Laboratory



# Comparison to other SpeechDat Corpora

Portuguese B1, C1-C4, I1	6.6
Norwegian B1,C1,C4	5.9*
Slovenian B1,C1	6.1*
English B1,C1,C4	4.3*

\* Lindberg, Johansen, Warakagoda, Lehtinen, Kacic, Žgank, Elenius, Salvi, “A Noise Robust Multilingual Reference Recognizer based on Speechdat (II)”, ICSLP, 3:370-373, 2000.

# HMM Reference Recognizer vs. HMM/MLP Hybrid

	Q1	I1	C1	O3	A1-3
<b>REFREC</b> <b>5000--0.95</b>	<b>0.2</b>	<b>0.5</b>	<b>3.4</b>	<b>2.8</b>	<b>3.0</b>
<b>HMM/MLP</b> <b>Hybrid</b>	<b>0.8</b>	<b>1.8</b>	<b>3.8</b>	<b>5.6</b>	<b>3.4</b>