Carla Simões,

t-carlas@microsoft.com

Microsoft | Development Center
Portugal

Survey on European and Brazilian Portuguese Speech and Text Corpora

Portuguese Corpora

- **Where can we find it?**
  - **World Wide Web**
    - www.elra.info , European Language Resources Association
    - www.elda.org, Evaluations and Language resources Distribution Agency
    - www.ldc.upenn.edu, Linguistic Data Consortium
    - www.iltec.pt, Instituto de Linguística Teórica e Computacional
    - www.clul.ul.pt, Centro de Linguística da Universidade de Lisboa
    - www.l2f.inesc-id.pt, Laboratório de Sistemas de Língua Falada, INESC
    - www.linguateca.pt, Language resource center for Portuguese
    - devoted.to/corpora, Bookmarks for Corpus-based Linguists
    - www.appen.com.au, Appen Speech Technologies

ELRA

www.elra.info

- Spoken corpus
  - **Desktop/microphone**
    - C-ORAL-ROM - Integrated reference corpora for spoken romance languages
    - FASiL Portuguese “fasil-pt” corpus
    - Portuguese Speecon database
    - GlobalPhone Portuguese (Brazilian)
  - **Telephony**
    - Portuguese SpeechDat(M) database
    - Portuguese SpeechDat(II) FDB-4000

- Written corpus
  - **Monolingual Lexicon**
    - LusoLEX European Portuguese Lexicon
    - BrasiLEX Brazilian Portuguese lexicon
    - PAROLE Portuguese Lexicon
    - LABEL-LEX (MW)
    - LABEL-LEX (SW)
  - **Written corpora**
    - PAROLE Portuguese Corpus
    - ECI/MCI (European Corpus Initiative/Multilingual Corpus I)
    - MLCC - Multilingual and Parallel Corpora

Spoken corpus – Desktop/microphone

- **C-ORAL-ROM - Integrated reference corpora for spoken romance languages**
  - The corpus consists of four comparable recording collections of Italian, French, Portuguese and Spanish **spontaneous speech sessions** (around **300,000 words for each Language**)
  - It provides the acoustic source of each session together with the following main annotations:
    - The **orthographic transcription**, in CHAT format, enriched with the tagging of terminal and non terminal prosodic breaks
    - Session metadata
    - The text to speech synchronization, in WIN PITCH CORPUS format, based on the alignment of each transcribed utterance
  - Package:
    - uncompressed .WAV files (Win PCM: **22,050 hz; 16 bit**)
    - Transcription files in .TXT and .XML format
    - transcription files with PoS tagging in .TXT files
    - The frequency list of lemmas for each language collection in TXT files
    - Measurements of spoken language variability in EXCEL files

- **Catalog Reference :** S0172
- **Source Channel :** Microphone, Radio, Telephone, Television
- **Members Prices**
  - Academic - Commercial 10000.00 EUR
  - Academic - Research 1500.00 EUR
  - Commercial - Commercial 10000.00 EUR
  - Commercial - Research 10000.00 EUR
- **Non Member Prices** :
  - Academic - Commercial 20000.00 EUR
  - Academic - Research 3000.00 EUR
  - Commercial - Commercial 20000.00 EUR
  - Commercial - Research 20000.00 EUR

Spoken corpus – Desktop/microphone

- **FASiL Portuguese unimodal "fasil-pt" corpus**
  - The corpus was collected in the context of the FASiL project, EU FP5 IST-2001-38685 (http://www.fasil.co.uk), as a wizard-of-oz experiment
  - There are sound recordings of subject and wizard. A total of **70 subjects were recorded**
  - The woz experiment is about the voice interaction with a Virtual Personal Assistant (VPA) for an email, calendar and contacts task
  - **.wav** files (u-law) **for audio**, plain **ASCII text (.txt) for transcriptions**
    - **Catalog Reference :** S0174-02
    - **Distribution medium :** CD-ROM, DVD
- Also available the **FASiL combined unimodal "fasil-all" corpus** and the **FASiL multimodal "fasil-mm" corpus,** where the subjects were recorded in three project languages: Swedish, Portuguese and English
- Demo

- **Members Prices**
  - Academic - Commercial 8000.00 Academic - Research 4000.00 EUR
  - Commercial - Commercial 8000.00 EUR
  - Commercial - Research 8000.00 EUR
- **Non Member Prices**
  - Academic - Commercial 10000.00 EUR
  - Academic - Research 8000.00 EUR
  - Commercial - Commercial 10000.00 EUR
  - Commercial - Research 10000.00 EUR

Spoken corpus – Desktop/microphone

- **Portuguese Speecon database**
  - It's a Portuguese speech corpus recorded in Portugal
  - Recorded at **16 KHz, 16 bit – linear**
  - **87 hours**
  - Microphone: CloseTalk/ FarTalk
  - Divided in two sets:
    - The first set comprises the recordings of 553 adult Portuguese speakers (266 males, 287 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place)
    - The second set comprises the recordings of 52 child Portuguese speakers (19 boys, 33 girls), recorded over 4 microphone channels in 1 recording environment (children room)
    - This database is partitioned into 29 DVDs (first set) and 4 DVDs (second set)

- **Catalog Reference :** S0180
- **Source Channel :** Microphone

- **Members Prices**
  - Academic - Commercial 67000.00
  - Academic - Research 50000.00 EUR
  - Commercial - Commercial 67000.00
  - Commercial - Research 67000.00

- **Non Member Prices**
  - Academic - Commercial 75000.00 EUR
  - Academic - Research 60000.00 EUR
  - Commercial - Commercial 75000.00 EUR
  - Commercial - Research 75000.00 EUR

Menu

Spoken corpus – Desktop/microphone

- **GlobalPhone Portuguese (Brazilian)**
  - provides **transcribed speech** data for the development and evaluation of large vocabulary continuous speech recognition systems in the most widespread languages of the world
  - The Portuguese (Brazilian) corpus was produced using the Folha de São Paulo newspaper
  - The entire GlobalPhone corpus contains over **300 hours of speech spoken** by more than **1500 native adult speakers**
  - It contains recordings of 102 speakers (54 males, 48 females with different age distribution) recorded in Porto Velho and Sao Paulo, Brazil
  - In each language about 100 adult native speakers were asked to read 100 sentences
  - About **2 Gb** for each language
  - **close-speaking microphone, PCM encoding, mono quality, 16-bit quantization, and 16 kHz sampling rate**

- **Catalog Reference :** S0201
- **Info**
  - http://www.cs.cmu.edu/~tanja/GlobalPhone
- **Distribution medium :** DVD

- **Members Prices**
  - Academic - Commercial 3000.00
  - Academic - Research 600.00
  - Commercial - Commercial 3000.00 EUR
  - Commercial - Research 3000.00
- **Non Member Prices**
  - Academic - Commercial 3600.00 Academic - Research 700.00
  - Commercial - Commercial 3600.00 EUR
  - Commercial - Research 3600.00

Spoken corpus – Telephony

- **Portuguese SpeechDat(M) database**
- An INESC project under a subcontract with Portugal Telecom – first phase
  - Contains the recordings of **1,001 speakers** (453 males, 548 females). This speech database was collected by Portugal Telecom within the European SpeechDat project
  - It has a good representation of many regional accents, and age distribution
  - It is also included a pronunciation lexicon with a phonemic transcription in SAMPA
  - **8 kHz, 8-bit A-law**
  - Each speaker uttered the following items:
    - natural numbers
    - digits
    - money amounts
    - dates
    - time phrase
    - application words
    - spelled-out words
    - word spotting phrases
    - sentences
    - yes/no questions
    - spontaneous date
    - spontaneous time
    - region name

- **Catalog Reference :** S0068
- **Source Channel :** Telephone
- **Distribution :** CD-ROM
- **Members prices :**
  - Academic - Commercial 14000.00
  - EURAcademic - Research 11000.00
  - EURCommercial - Commercial 14000.00
  - EURCommercial - Research 14000.00 EUR

- **Non member prices :**
  - Academic - Commercial 20000.00
  - EURAcademic - Research 14000.00
  - EURCommercial - Commercial 20000.00
  - EURCommercial - Research 20000.00 EUR

Menu

Spoken corpus – Telephony

- **Portuguese SpeechDat(II) FDB-4000**
- An INESC project under a subcontract with Portugal Telecom – second phase
  - Comprises **4027 Portuguese speakers** (1861 males, 2166 females) recorded over the Portuguese fixed telephone network. It has a good representation of many regional accents, and age distribution
  - **8-bit 8 kHz A-law**
  - A **pronunciation lexicon** with a phonemic transcription in **SAMPA**
  - Each speaker uttered different items:

    - digits, numbers
    - currency money amount
    - dates
    - time phrases :
    - spelled words :
    - directory assistance utterances
    - yes/no questions
    - application words
    - phonetically rich words
    - phonetically rich sentences

- **Samples -** http://speechdat.phonetik.uni-muenchen.de/speechdt/speechDB/FIXED1PT/HTML/index.html

- **Catalog Reference :** S0092
- **Source Channel :** Telephone
- **Distribution :** CD-ROM
- **Members prices** :
  - Academic - Commercial 40000.00 EUR
  - Academic - Research 28000.00 EUR
  - Commercial - Commercial 40000.00
  - Commercial - Research 40000.00 EUR
- **Non member prices :**
  - Academic - Commercial 56000.00 EUR
  - Academic - Research 48000.00 EUR
  - Commercial - Commercial 56000.00
  - Commercial - Research 56000.00 EUR

Menu

Written corpus - Monolingual Lexicon

- **LusoLEX European Portuguese Lexicon** and **BrasiLEX Brazilian Portuguese lexicon**
  - Available at Microsoft Language Resources

- **PAROLE Portuguese Lexicon**
  - It's constituted by 20 000 entries morpho-syntactically and syntactically encoded
  - **Distribution medium :** CD-ROM
    - **Members Prices**
      - Academic - Commercial 10500.00 EUR
      - Academic - Research 1400.00 EUR
      - Commercial - Commercial 10500.00 EUR
      - Commercial - Research 3500.00 EUR
    - **Non Member Prices**
      - Academic - Commercial 15000.00 EUR
      - Academic - Research 2000.00 EUR
      - Commercial - Commercial 15000.00 EUR
      - Commercial - Research 5000.00 EUR

Written corpus – Monolingual Lexicon

- **LABEL-LEX (MW)**
  - It is a Portuguese formalized lexicon, containing **88 619 multiword lexical units** (formally, sequences of simple words)
  - Often, are used to express ideas and concepts
  - It is important the treatment of multiword lexical units, to improve the automatic text analysis
    - **Catalog Reference :** L0054

- **LABEL-LEX (SW)**
  - It's a Portuguese formalized lexicon, containing 1.545.156 simple words
  - Each entry is associated to syntactic and semantic information
    - **Catalog Reference** : L0055

- **Members Prices**
  - Academic - Commercial 10000.00 EUR
  - Academic - Research 3000.00 EUR
  - Commercial - Commercial 10000.00 EUR
  - Commercial - Research 10000.00 EUR
- **Non Member Prices**
  - Academic - Commercial 15000.00 EUR
  - Academic - Research 5000.00 EUR
  - Commercial - Commercial 15000.00 EUR
  - Commercial - Research 15000.00 EUR

- **Members Prices**
  - Academic - Commercial 10000.00 EUR
  - Academic - Research 2500.00 EUR
  - Commercial - Commercial 10000.00 EUR
  - Commercial - Research 10000.00 EUR
- **Non Member Prices**
  - Academic - Commercial 15000.00 EUR
  - Academic - Research 5000.00 EUR
  - Commercial - Commercial 15000.00 EUR
  - Commercial - Research 15000.00 EUR

Label – Lex project, developed by Laboratório de Engenharia da Linguagem of Faculdade de Letras da Universidade de Lisboa. They are used in the AnELL (Anotador Electrónico LabEL-Linguateca), http://www.linguateca.pt/AnELL/

Menu

Written corpus – Written corpora

- **PAROLE Portuguese Corpus**
  - contains approximately 3 million running words of European Portuguese
    - Newspaper: about 65%, covering the period 1996-1997 of 3 titles
    - Book: about 20%, concerning 12 titles from 3 editing houses
    - Periodical: about 5%, concerning 7 weekly issues of 1 title, 1996
    - Miscellaneous: about 10%, concerning several files distributed by 8 titles
  - **Catalog Reference** : W0024-01

- **Members Prices**
  - Academic - Commercial 2450.00
  - Academic - Research 875.00 EUR
  - Commercial - Commercial 2450.00
  - Commercial - Research 1575.00
- **Non Member Prices**
  - Academic - Commercial 3500.00
  - Academic - Research 1250.00
  - Commercial - Commercial 3500.00
  - Commercial - Research 2250.00

- **PAROLE Portuguese Corpus - tagged subset**
  - A subcorpus of the PAROLE Portuguese Corpus, which reproduces approximately the whole Corpus distribution (Newspaper: about 65%, Book: ab. 20%, Periodical: ab. 5%, Miscellaneous: ab. 10%)
  - It has about 250,000 words morpho-syntactically tagged
  - **Catalog Reference** : W0024-02

- **Members Prices**
  - Academic - Commercial 1750.00
  - Academic - Research 525.00 EUR
  - Commercial - Commercial 1750.00
  - Commercial - Research 875.00
- **Non Member Prices**
  - Academic - Commercial 2500.00
  - Academic - Research 750.00 EUR
  - Commercial - Commercial 2500.00
  - Commercial - Research 1250.00

Menu

Written corpus – Written corpora

- **ECI/MCI (European Corpus Initiative/Multilingual Corpus I)**
  - ECI has produced the Multilingual Corpus I (ECI/MCI) covering most of the major European languages
  - The primary focus in this effort is on textual **material of all kinds**, including **transcriptions of spoken material**
  - The ECI/MCI is available from ELSNET (http://www.elsnet.org/eci.html)
  - **98 million words**
    - **Catalog Reference :** W0004
    - **Distribution medium :** CD-ROM
    - **Platform :** PC, Unix, Macintosh

- **Members Prices**
  - Academic - Research 50.00 EUR
  - Commercial - Research 50.00 EUR

- **Non Member Prices**
  - Academic - Research 50.00 EUR
  - Commercial - Research 50.00 EUR

Written corpus – Written corpora

- **MLCC - Multilingual and Parallel Corpora**
  - 2 sets:
    - The first set contains articles from 6 European newspapers (93.38 million words total)
    - the second set consists of a parallel corpus of translated data in the nine European official languages (1992-1994) divided into 2 sub-corpora:
      - written questions (10.2 million words)
      - parliamentary debates (5 to 8 million words per language in addition to the specific amounts indicated for each language).
    - For European Portuguese **1100000 words**
      - Catalog Reference : ELRA-W0023
      - Distribution medium : CD-ROM

- **Members Prices**
  - Academic - Research 450.00 EUR
  - Commercial - Research 1600.00 EUR

- **Non Member Prices**
  - Academic - Research 1200.00 EUR
  - Commercia- Research 3600.00 EUR

Menu

Linguistic Data Consortium

www.ldc.upenn.edu

Spoken Corpus – Wideband Speech

- LDC2006S16 **CSLU: Spoltech Brazilian Portuguese Version 1.0**
  - Contains **microphone speech** from a variety of regions in Brazil with **phonetic and orthographic transcriptions**
  - The utterances consist on read speech and responses to questions
  - The corpus contains **477 speakers and 8080 separate utterances**
  - **2540 utterances** have been transcribed at the word level (without)
  - **5479 utterances** have been transcribed at the phoneme level (with time alignments)
  - The data have been recorded at **44.1 kHz (mono, 16 bit)** and stored in RIFF format, **1 hour**
    - **Distribution** 1CD
    - **Non-member Fee:** US$150.00
    - **Extra-Copy Fee:** US$150.00

Spoken Corpus - Telephony

- LDC2005S26 **CSLU: 22 Languages Corpus**
  - Consists on telephone speech from 21 languages
  - It has a total of **50191 utterances**
  - Contains fixed vocabulary utterances (e.g. days of the week), and fluent continuous speech. The utterances are verified by native speakers
  - For this release, approximately **19758 utterances** have corresponding **orthographic transcriptions**
  - **8-bit ulaw at 8 KHz, 1 hour each language**
    - **Catalog Number** LDC2005S26
    - **Microphone Type** Telephony
    - **Language Dialect** Brazilian
    - **Distribution** 2DVDs
    - **Non-member Fee**: US$150.00
    - **Reduced-License Fee:** US$150.00
    - **Extra-Copy Fee:** US$150.00

Written corpus - Text

- LDC2001T62 **CETEMPúblico Version 1.7 [CETEMPúblico]**
  - Described in the Linguateca section

- LDC94T5 **ECI/MCI (European Corpus Initiative/Multilingual Corpus I)**
  - ECI has produced the Multilingual Corpus I (ECI/MCI) covering most of the major European languages
  - The primary focus in this effort is on textual **material of all kinds**, including **transcriptions of spoken material**
  - **98 million words**
    - **Non-member Fee:** US$35.00
    - **Reduced-License Fee:** US$35.00
    - **Extra-Copy Fee**: US$35.00
    - **Distribution**: 1 CD

Written corpus - Text

- **LDC95T11 European Language Newspaper Text**
  - This corpus includes roughly 100 million words of French, 90 million words of German and **15 million words of Portuguese** and has been marked using SGML
    - **Data Type:** text
    - **Data Source(s):** newswire
    - **Distribution:** 1 CD
    - **Application(s):** information retrieval, language modeling

- **LDC99T40 Portuguese Newswire Text**
  - This corpus is built on the Portuguese data published previously in the European Language Newswire Text Corpus and contains the previously published material, as well as more recent material
    - **Data Type**: text
    - **Data Source(s):** newswire
    - **Application(s):** information retrieval, language modeling
    - **Distribution:** 1 CD
    - **Non-member Fee**: US$400.00
    - **Reduced-License Fee**: US$200.00
    - **Extra-Copy Fee**: US$150.00

Instituto de Linguística Teórica e Computacional

www.iltec.pt

The REDIP Corpus

- **Speech and written corpus**
- http://www.iltec.pt/?action=concord

- REDIP (Rede de Difusão Internacional do Português: rádio, televisão e imprensa), it's a **speech and written corpus** containing data from:
  - radio, television and press

- Project developed by Instituto de Linguística Teórica e Computacional (ILTEC), with the cooperation of the Centro de Linguística da Universidade de Lisboa (CLUL) and Universidade Aberta,

- Contains **330.000 words**, distributed by six themes:
  - News, science, culture, economy, sports and opinion

- It's possible to access to the **Redip corpus** through the extractor **SimpleConcord;**
  - this allows the search by words and parts of words, from the available list of texts
    - http://www.iltec.pt/handler.php?action=concord&act=list

MorDebe

- **Monolingual lexicon**
- http://www.iltec.pt/mordebe/

- The **MorDebe,** is a **lexicon database** that gives information about the orthography and inflection of most Portuguese words

- The majority of words belongs to the **European Portuguese variant**

- Those that appear on the text sources, but don’t belong to the European Portuguese variant, are also included and marked as Portuguese from Brazil, Mozambique, Angola or Cabo Verde

- The data was generated and is frequently updated from **Portuguese newspapers**, as **Público** and **Diário de Notícias:**
  - **Number of lemmas: 125994**
  - **Number of wordforms: 1265661**

- It’s a free distribution database without commercial purposes

Centro de Linguística da Universidade de Lisboa

www.clul.ul.pt

Corpora

- **Corpus de Referência do Português Contemporâneo - CRPC**

  http://www.clul.ul.pt/sectores/projecto_crpc.html
- The CRPC is an electronically based linguistic corpus containing at the present

  92 000000 words taken by sampling from several types of **written speech** (literary, newspaper, technical, scientific, didactic, economics, decisions of the supreme court of justice, parliament) and **oral speech** (formal and informal)
  - These samplings refer to national and regional varieties of Portuguese
    - European, Brazilian, African, Macau, and East-Timor
    - contains texts from the second half of the XIX century up until now, mostly after 1970
  - The CRPC **main goal** is the **continous construction of a balanced corpus** and its availability and dissemination to make this resource easily and accessible
- **Resumed table with the available corpora in CLUL**

CRPC

Applications and Projects that use the CRPC Corpus

- **Concluded Projects:**
  - Recursos Linguísticos para o Português: um *corpus* e instrumentos para a sua consulta e análise
  - REDIP - Rede de Difusão Internacional do Português: rádio, televisão e imprensa (partnership)
  - Léxico Multifuncional Computorizado do Português Contemporâneo (partnership)
  - Português Falado - Variedades Geográficas e Sociais
  - SIMPLE – Semantic Information for Multifunctional Plurilingual Lexicon (partnership)
  - LE-PAROLE (partnership)
  - ELAN – European Language Activity Network (partnership)
  - Propriedades Sintácticas e Semânticas de Predicados Verbais Polissémicos: o caso dos verbos psicológicos
  - DCP – Dicionário de Combinatórias do Português
  - AUDIOLING-LP Língua Portuguesa: som e pronúncia
  - Estudo do uso e do Significado da Preposição DE em Contextos Nominais SN DE SN
  - Língua Portuguesa: Ensino Assistido por Computador (partnership)

- **On going Projects:**
  - Gramática do Português
  - C-ORAL-ROM – Integrated Reference *Corpora* for Spoken Romance Languages (partnership)
  - ENABLER - European National Activities for Basic Language Resources (partnership)
  - Aspectos da Sintaxe do Sintagma Nominal em Português numa perspectiva comparada
  - VARPORT - Análise Contrastiva de Variedades do Português (partnership)

Corpora

- **Dialectal oral corpus of the Atlas Linguístico-Etnográfico de Portugal e da Galiza (ALEPG) -** www.clul.ul.pt/english/sectores/projecto_alepg.html
  - The CLUL Dialectal Studies Research Group has collected approximately **3500 hours** of **recorded speech in 176 inquiries carried out in the Portuguese continental territory**
    - This corpus contains mainly **directed speech**: answers to a linguistic questionnaire specially built for a national Atlas.
    - there is also an important part of **spontaneous speech**
  - Applications:
    - ALEAç: Linguistic and Ethnographic Atlas of Azores
    - ALLP: Linguistic Atlas of the Portuguese Coast
    - ALE: Atlas of the European Languages
    - ALiR: Linguistic Atlas of the Romance Domain
    - Border Languages: Mirandese
    - Border Languages: Barranquenho

ALEPG

Applications and Projects that use the ALEPG Corpus

- Other applications
  - **VarV: Inflectional Variants of the Verb, in Spoken Continental Portuguese** www.clul.ul.pt/english/sectores/projecto_estudo_variantes.html
    - This project intends to make the inventory of the variants of verbal inflexion observed in the continent in order to establish the variant patterns analyzing and characterizing them mostly in terms of morphological and phonological features and to define the main dialectal/geographical areas of each inflexion pattern and the inflexion between areas
  - **Syntax-oriented Corpus of Portuguese Dialects (CORDIAL-SIN)** www.clul.ul.pt/english/sectores/cordialsin/projecto_cordialsin.html
    - This project is aimed at developing and enhancing research activity on syntactic dialect variation. The project outcome will be an electronic medium-size database – of about 30 000 sentences – integrating a detailed syntactic annotation

Laboratório de Sistemas de Língua Falada, INESC

www.l2f.inesc-id.pt

Current Corpora

- LECTRA - Classroom lectures
- IPSOM - Aligned spoken books
- ALERT - Broadcast news
- CORAL - Spoken dialogues (map task)
- BD-PÚBLICO - Large vocabulary, speaker-independent, continuous speech
- SPEECHDAT - Multi-purpose telephone speech database
  - Available in the Elra repository
- BDFALA - Speech analysis / synthesis
- EUROM.1 - Multi-Lingual speech corpus for phonetic comparison

LECTRA - Classroom lectures

- Project in progress
- Two different courses have been selected:
  - "Economic Theory I“ - ETI
    - 17 classes
  - "Production of Multimedia Contents“ - PMC
    - 20 classes

- Two different positions for microphones
  - lapel microphone
    - ETI and 6 classes of PMC
  - head-mounted microphone
    - 14 classes of PMC
- The classes had variable duration, ranging from 40 to 90 minutes
- Only the professor’s dialogs were transcribed

IPSOM - Aligned spoken books

Microsoft | Development Center
Portugal

- Pilot corpus consisting of different types of aligned books, to test the software developed in IPSOM
  - **Fiction:**
    - O Senhor Ventura, by Miguel Torga, read by Isabel Bahia
  - **Poetry:**
    - Um Momento de Palavras, David Mourão Ferreira, read by himself
    - Manifesto Anti-Dantas, Almada Negreiros, read by Mário Viegas
    - Fado Falado, Aníbal Nazaré, read by João Villaret
  - **Children's stories:**
    - O Monge Desastrado, included in Histórias de Belém, Ana Maria Magalhães e Isabel Alçada, read by Jorge Moreira and Juva Batella (European and Brazilian Portuguese, respectively)
  - **School books:**
    - Chapter of a History book (2nd cycle), read by Jorge Moreira and Ana Bernardino

Menu

ALERT - Broadcast news

- Corpus for training and evaluating several components of the ALERT media watch system for European Portuguese
- RTP, as the Portuguese data provider in this project, was responsible for collecting the data at their premises
- 3 main parts:
  - **Speeh Recognition Corpus** (SRC) The main goal of this corpus was the training of the acoustic models and the adaptation of the language models used in the large vocabulary speech recognition component of our system
    - include **122 programs of different types** and schedules
    - 76h of audio data, with transcriptions
  - **Topic Detection Corpus** (TDC) The main goal of this corpus was to have a broader coverage of topics and associated topic classification for training our topic indexation module
    - data related to 133 TV broadcast of the 8 o'clock evening news program.
    - 300 hours of recordings, automatic orthographic transcriptions
  - **Textual Corpus** (TC) daily extraction of newspaper texts, L2F is keeping up with this daily collection activity, now reaching close to 450 million words

CORAL - Spoken dialogues

Microsoft | Development Center Portugal

- **Labelled spoken dialogue corpus -** Transcript speech corpora
  - The purpose of this project is the collection of a **spoken dialogue corpus**, with **several levels of labelling**: orthographic, phonetic, prosodic, syntactic and semantic
  - Number of speakers: **32**, grouped into 8 quartets, amounting to 64 dialogues
  - It is available in **5 CDs**, amounting to **1.6 Gb.**
  - The frequency is **16kHz**. Its availability in ***wav* format is also possible**.
  - All dialogues have been **annotated orthographically**
    - Example: **orthographic labelling**

Menu

BDFala – EUROM.1

- **BDFala** - Speech analysis/ synthesis
  - It was jointly developed by INESC and CLUL.
  - Enlargement of the EUROM.1 corpus and improvement of speech synthesis systems
  - The purpose was collecting a database of spoken European Portuguese. It was structured in 5 tasks:
    - Normalization, statistic research of the European Portuguese, specification of the database, collect the database and recording it
  - It was included, **5000 speakers from the telephony network** that were recorded in the SpeechData project
  - **It has spontaneous speech**, from television resources
  - The data was stored in **4CDs**, amounting to **2 Gb**
  - The sampling frequency was **16 kHz**
    - 4600 isolated words
    - 350 sentences for prosodic studies
    - 18 phonetically-complete paragraphs
    - 60 read paragraphs extracted from television debates
    - 600 phonetically rich sentences

- **EUROM.1 -** Multi-Lingual speech corpus for phonetic comparison
  - It was jointly developed by INESC and CLUL
  - 11 languages, 4 types of corpus material were collected:
    - CVC material (totalling 121 different logatomes) in isolation and in context (5 carrier phrases)
    - 100 selected numbers from 0-9999
    - 40 short passages each containing 5 thematically connected sentences; 50 filler sentences to compensate for the phoneme-frequency imbalance in the passages
  - **5 CDs** and totals **2.6 Gb**
  - The sampling frequency was 20 kHz

Linguateca

www.linguateca.pt

Linguateca

- Linguateca is a distributed language resource center for **Portuguese Language**
  - Provides:
    - Written text corpora
    - Transcript speech corpora
    - Speech corpora
    - Parallel/ Comparable corpora
    - Projects for corpora availability

Written corpus

- **CETEMPúblico: a large corpus of Portuguese newspaper language**
  - Corpus in European Portuguese from the Portuguese daily newspaper PÚBLICO;
  - The corpus includes the text of around 2,600 editions of PÚBLICO, written (stored) between 1991 and 1998, amounting to approximately **180 million words**
  - CETEMPúblico 1.7 contains **1,504,258 extracts** (CETEMPúblico 1.0 had 1,567,625), bearing the information about section of origin and semester. Each extract is divided in paragraphs and sentences, and titles and authors are marked as such. See some examples of extracts
  - The last version is **CETEMPúblico 1.7,** in **annotated** or **not annotated** form
  - The corpus is available by 4 different ways:
    - **AC/DC project**, granting Web access to corpora
    - Through ftp or http download
    - in CD-ROM
    - Through the Linguistic Data Consortium (LDC), CETEMPúblico Version 1.7
- For more information about **CETEMPúblico** and how to get it:
  - http://acdc.linguateca.pt/cetempublico/whatisCETEMP.html

Written corpus

- CETENFolha
  - It’s a corpus in Brazilian Portuguese from the Brazilian newspaper Folha de S. Paulo
  - The corpus includes approximately **24 million words**
  - CETENFolha 1.0 contains **340.947 extracts**, from 365 editions of 1994 some exemples
  - The corpus is available by 2 different ways:
    - Through ftp or http download
    - **AC/DC project**, as a part of the NILC/São Carlos corpus
- For more information about **CETEMFolha** and how to get it:
  - http://acdc.linguateca.pt/cetenfolha/#rede

Written corpus

- Newspaper corpus **Natura-PUBLICO** http://natura.di.uminho.pt/~jj/pln/pln.html
- This corpus contains extracted sentences from PÚBLICO newspaper
- The paragraphs were extracted from the 1991 to 1994 editions
- It is also available wordlists, for orthographic correction and grammatical classification
- These resources are available for free download in the website of **Projecto Natura,** or for access in http://acdc.linguateca.pt/acesso/

Written corpus

- **Corpus NILC/São Carlos** - http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm
  - The main goal was building corpora for supporting NLP researches, especially on Brazilian Portuguese
  - a **40 million word corpus** consisting of prose texts in Brazilian Portuguese
  - it is available for access in http://acdc.linguateca.pt/acesso/
  - Here is the list of words and their frequencies collected from the corrected texts

Written corpus

- **Lácio-Ref -** http://www.nilc.icmc.usp.br/lacioweb/corpora.htm
  - Open reference corpus, containing **texts in Brazilian Portuguese** within the boundaries of standard orthography and accepted grammar
  - no PoS, syntactic or semantic tags, but there is an indication as to the existence of graphic elements in the original texts
  - The majority of the texts are **available in their integral forms**
  - The texts go through three phases before they are made available in the corpus:
    - **Compilation-Formatting:** consists on selecting the texts, storing them in electronic media in a .txt format
    - **Naming:** Text files are named taking into consideration the text's genre, media and other features that allow for the automatic construction of subcorpora
    - **Cataloguing:** a header is inserted in the file with ordinary bibliographic information and descriptive information
  - **Access to the corpora and tools** is granted **after the registration** form is completed

Menu

Written corpus

- **Mac-Morpho -** http://www.nilc.icmc.usp.br/lacioweb/corpora.htm
  - Closed (complete) annotated corpus
  - It is composed of newspaper articles published by *Folha de São Paulo* in 1994, in the following sections:
    - sports, money, science, agriculture, information technology, culture and arts, weekend supplement of culture and arts, international, brazil and local news
  - It consists of **1.174.206 words**, which have been automatically annotated with Eckhard Bick's parser "Palavras“
  - **Access to the corpora and tools** is granted **after the registration** form is completed

Written corpus

- **Corpus TeMário (acronym for ‘TExtos com suMÁRIOs’)** http://www.linguateca.pt/Repositorio/TeMario/
  - It was created for the project EXPLOSA of NILC
  - It consists on **100 newspaper articles** (**61.412 words),** along with both their manual summaries and ideal extracts (these have been automatically generated)
  - The texts are written **in Brazilian Portuguese** to support evaluation of Automatic Summarization tasks
  - A technical report describing TeMário is available in pdf format and should be read for clarification on its organization
  - This corpus can be **downloaded** in the **linguateca website** or in the **Explosa project website**

Written corpus

- **Corpus Informatizado de Português Medieval -** http://cipm.fcsh.unl.pt/
  - this corpus includes **Latin-romances texts** from the IX century to the XII century, and **Portuguese texts** from the XII century to the XVI century
  - The texts were annotated using the **CIPM methodology**, CIPM annotation
  - it is available for access through the CIPM website

Written corpus

- **The Tycho Brahe Parsed Corpus of Historical Portuguese** www.ime.usp.br/~tycho/corpus/files/index.html
  - The *Tycho Brahe Parsed Corpus of Historical Portuguese* is an electronic **annotated corpus**, formed by texts written between 1500-1900, freely available for researchers
- **Wordtheque - multilingual library** http://www.logos.it/literature/literature.html
  - Interface with a massive database (currently **707.737.941 words**) containing multilingual novels, technical literature and translated texts

- **O Corpus do Portugês -** www.corposdoportugues.org
  - It's a website where we can search more than **45 million words** in more than 50,000 Portuguese texts from the 1300s to the 1900s (Brazilian and European)
  - The search can be made using **exact words or phrases, lemmas, part of speech, or any combinations of these**. We can also search for **surrounding words**
  - It's easy to compare the frequency and distribution of words, phrases, and grammatical constructions across texts
  - There are 3 ways to do it::
    - **By register:** comparisons between spoken, fiction, newspaper, and academic
    - **By dialect**: Portugal compared with Brazil
    - **By historical period**: compare different centuries from the 1300s to the 1900s

Written corpus

- **XLDB Node of Linguateca - Faculty of Sciences - University of Lisbon** - WPT 03 (Portuguese Web collection of 2003)
  - It's the most complete portuguese web collection at the moment
  - Near 12 Gb
  - It was collected by the tumba! search engine crawlers.
  - Contains all files of type TXT, HTML, RTF, PDF, PS, DOC, XLS, PPT and SWF, that are hosted:
    - under the .pt domain
    - under the .com, .org, .net or .tv domain, given that they are written in portuguese, and have at least one inlink from a file hosted under the .pt domain
  - the documents are in **plain text**, without HTML tags. For binary documents such as PDF, PS, DOC, XLS, PPT or SWF, the text was extracted by proper converters
  - two alternative WPT 03 formats:DVD, Sql database
    - available over request into XLDB Node of Linguateca
  - Demo

Transcript speech corpora

- **Nexing Corpus -** http://www.di.fc.ul.pt/~ahb/nexingcorpus.htm
  - Written transcriptions:
    - Verbal data (30 hours) elicited during an experiment on syllogistic reasoning (each of 27 participants x the 64 syllogistic problems)
  - 27 files (ISO-8859-1) with around 15000 tokens each
  - it is available for visualize and download in the Nexing project website
  - A detailed description of the Nexing Corpus:
    - Branco, António, José Leitão, João Silva and Luís Gomes, 2002, "Nexing Corpus: a corpus of verbal protocols on syllogistic reasoning". In Proceedings of LREC2002-Third International Conference on Language Resources and Evaluation, Las Palmas, May 29-31

Speech corpora

- **European Portuguese DataBase for Speech Synthesis FEUP/IPB-BD**
- an example track - http://www.linguateca.pt/Repositorio/EuroSpeechIPB/
  - This work was produced, in part under the PhD dissertation of João Paulo Teixeira and in part under the ANTIGONA Project (Program IC-PME)
  - Every track has been carefully examined and segmentation marks placed using the Speech Filing System (SFS) software tool
  - For each track phrase, word and phoneme labels were attached
  - The phonetic level labels are based in the SAMPA
  - **44.1 KHz, 16 bits, mono**
  - It is available for download in the above website

Speech corpora

- **BDFALA (Base de Dados Falada para o Português Europeu)** http://www.speech.inesc.pt/projects/bdfala/bdfala_pt.html
  - The purpose of this project was to collect a database of spoken European Portuguese. It was structured in 5 tasks:
    - Normalization, statistic research of the European Portuguese, specification of the database, collect the database and recording it
    - 5000 speakers, were included, from the telephony network that were recorded in the SpeechData project
  - It has spontaneous speech, from television resources
  - The data was stored in 4CDs, amounting to 2 Gb
  - For further information about the project, contact Isabel Trancoso - INESC (Instituto de Engenharia de Sistemas e Computadores), Lisboa

Parallel/ Comparable corpora

- COMPARA (Portuguese-English Parallel Translation Corpus) http://www.linguateca.pt/COMPARA/
- Comp-C http://www.nilc.icmc.usp.br/lacioweb/corpora.htm
- Corpus Técnico-Científico (CorTec) http://www.fflch.usp.br/dlm/comet/projeto.html#cortec
- Europarl: European Parliament Proceedings Parallel Corpus 1996-2003 http://people.csail.mit.edu/koehn/publications/europarl/
  - Available in the Microsoft Language Resources
- Par-C http://www.nilc.icmc.usp.br/lacioweb/corpora.htm
- The JRC-Acquis Multilingual Parallel Corpus http://langtech.jrc.it/JRC-Acquis.html
- Oslo Multilingual Corpus http://www.hf.uio.no/german/sprik/english/corpus.shtml
- The English-Norwegian Parallel Corpus http://www.hf.uio.no/iba/prosjekt/languages.html

AC/DC Project

Microsoft | Development Center Portugal

- **Main goals**
  - provide one place where access to all corpora is given
  - further improve the information associated with these corpora
  - develop a good user interface
- AC/DC URL
  - http://acdc.linguateca.pt/acesso/
- There are variants for the CETEMPúblico (versão 1.7) and for CETEMPúblico anotado (versão 1.7 anotado 2.0)
- It makes available all the information about the given corpora (annotated/ not annotated)
  - For more information about the process involved in this project
    - http://acdc.linguateca.pt/acesso/info_acesso_English.html#proc

- **Possible Corpora:**
  - Primeiro milhão do CETEMPúblico - anotado
  - Corpus Natura/Público – anotado
  - Parte portuguesa pública do ENPC - anotado
  - Corpus Natura/Diário do Minho – anotado
  - Corpus ECI-EBR – anotado
  - Corpus ECI-EE- anotado
  - Corpus NILC/São Carlos - anotado
  - Corpus FrasesPP - anotado segunda vez
  - Corpus FrasesPB - anotado
  - Corpus ANCIB - anotado
  - Corpus DiaCLAV - anotado
  - Corpus Avante! - Anotado
  - AmostRA-NILC - anotado segunda vez
  - Corpus CoNE - anotado
  - Corpus Clássicos da Literatura Portuguesa / Porto Editora - anotado
  - Corpus Museu da Pessoa - anotado
  - Corpus CONDIVport – anotado

Bookmarks for Corpus-based Linguists

http://devoted.to/corpora

Reuters Corpora

- **Written corpora**
- In 2000, Reuters Ltd made available a large collection of Reuters News stories for use in research and **development of natural language processing**, **information retrieval,** and **machine learning systems**

- **RCV2 -** Reuters Corpus, Volume 2, **Multilingual Corpus**, 1996-08-20 to 1997-08-19 (Release date 2005-05-31, Format version 1, correction level 0)

- This is distributed on **1 CD** and contains over **487,000 Reuters News stories** (1996-08-20 to 1997-08-19) in thirteen languages (Dutch, French, German, Chinese, Japanese, Russian, **Portuguese**, Spanish, Latin American Spanish, Italian, Danish, Norwegian, and Swedish)

- To request a copy of the Reuters Corpus:
  - Email: reuters-request@nist.gov
  - To get more information about it: http://trec.nist.gov/data/reuters/reuters.html

Appen Speech Technologies

http://www.appen.com.au

Written Corpora

- Only brazilian and european Portuguese lexicons are available

- Portuguese (EU)
  - 15,000 Common Words (approximately 600 Proper Nouns - more can be added on request)
  - Available
  - **Catalog number** PTP_LEX001

- Portuguese (EU)
  - 9,700 Mixture of Given Names and Family Names. Includes 1,795 Foreign Given Names
  - Available
  - **Catalog number** PTP_LEX002

- Portuguese (EU)
  - 2,000 Place Names
  - Available
  - **Catalog number** PTP_LEX003

- Portuguese (Brazil)
  - 5,000 Given Names
  - Available
  - **Catalog number** PTB_LEX001

- Portuguese (Brazil)
  - 10,000 Family Names
  - Available
  - **Catalog number** PTB_LEX002

- Portuguese (Brazil)
  - 26,000 Common words
  - Available
  - **Catalog number** PTB_LEX003

- Portuguese (Brazil)
  - 14,000 Common Words
  - In Development
  - **Catalog number** PTB_LEX004