Linguateca: seven years working for the computational processing of Portuguese

Diana Santos, Luís Miguel Cabral & Luís Costa
Linguateca
www.linguateca.pt

Linguateca in a nutshell

- a project whose aim is to considerably improve the conditions of the community who deals with the computational processing of the Portuguese language
- ~~processing of Portuguese = NLP specialized to Portuguese~~
- ~~just by financing individual research projects you build a community~~
- you have to build a research infrastructure and actively foster collaboration and joint evaluation

Inspiration for the title: Sonet of Camões

Sete anos de pastor Jacob servia
Labão, pai de Raquel, serrana bela;
mas não servia o pai, servia a ela,
e a ela só por prémio pretendia.

Os dias, na esperança de um só dia,
passava, contentando-se com vê-la;
porém o pai, usando de cautela,
em lugar de Raquel lhe dava Lia.

Vendo o triste pastor que com enganos
lhe fora assi negada a sua pastora,
como se não a tivera merecida,

Começa de servir outros sete anos,
dizendo: Mais servira, se não fora
pera tão longo amor tão curta a vida!

Seven years as shepherd Jakob served...

In the beginning

- There was a little project at SINTEF (1998-1999, 1999-2000), Diana Santos and Signe Oksefjell
- which produced a white paper
  Diana Santos. "Computational processing of Portuguese: working memo". 1999.

  written for a general discussion in Portugal of what to do to considerably improve this area
- and started what later on was called Linguateca
  - creating a portal for CPP
  - starting corpora services on the Web

Nowadays (seven years later)

- 7 senior members: *Diana Santos, José João Almeida, Eckhard Bick, Belinda Maia, Ana Frankenberg Garcia, Mário J. Silva, Paulo Gomes*
- 6 fulltime and 5 partialtime members: *Nuno Cardoso, Rui Vilela, Luís Miguel Cabral, António Silva, Susana Inácio, Ana Sofia Pinto; Luís Costa, Paulo Rocha, Raquel Marchi, Rosário Silva*
- 5 PhD students: *Marcirio Chaves, Alberto Simões, Nuno Seco; Luís Sarmento, Anabela Barreiro, (Cristina Mota, Susana Afonso)*
- 6 past members: *Rachel Aires* (first Phd student), *Renato Haber, Alex Soares, Pedro Moura, Débora Oliveira, Isabel Marcelino*

26 people involved in and with Linguateca

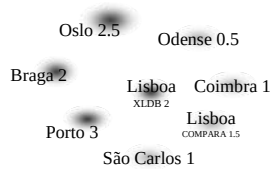
A virtual organization

Linguateca, a project for Portuguese

- A distributed resource center for Portuguese language technology

IRE model
- **I**nformation
- **R**esources
- **E**valuation

www.linguateca.pt



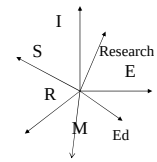
Linguateca highlights, www.linguateca.pt

- > 1000 links More than 2,500,000 visits to the Web site
- AC/DC, CETEMPúblico, COMPARA … Considerable resources for processing the Portuguese language
- *Morfolimpíadas* The first evaluation contest for Portuguese, followed by CLEF and HAREM

- Public resources
- Foster research and collaboration
- Formal measuring and comparison
- One language, many cultures
- Cooperation using the Internet
- Do not adapt applications from English

The IRE model and its evolution

- First: Information, Resources and Evaluation
- But then
  - (resource) Maintenance:
  - Support
  - Research (PhDs)
  - Education



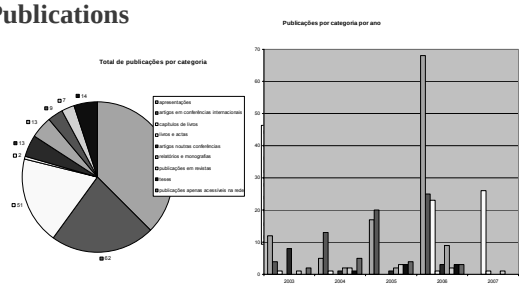
PhD dissertations

- Rachel Aires: Web categorization according to the users' intentions **(August 2005)** [SA & DS]
- Marcirio Chaves: Geographical ontology population and integration [MS & DS]
- Alberto Simões: Example based machine translation [JJ]
- Anabela Barreiro: Bilingual paraphrases for machine translation [BM]
- Nuno Seco: Creation and evaluation of a lexical ontology for Portuguese from published dictionaries [DS & PG]
- Luís Sarmento: Robust semantic analysis, biography construction as an extension of QA [DS & EO]
- Nuno Cardoso: Query reformulation [MS & DS]

Master's dissertations

- Alberto Simões: Word alignment **(September 2004)** [JJ]
- Nuno Cardoso: NER evaluation, statistical validation of HAREM **(November 2006)** [MS & EO]
- Luís Miguel Cabral: SUPeRB [EO] in progress

Publications



ca. 275 publications and presentations since 15 May 2006

Information

- Portal for the computational processing of Portuguese: you find everything that's going on there, there!
- Reasonably visited site, constantly updated
  - 1st November 2006: **2,688,142** hits (excluding robots)
- The usual stuff
  - catalogue
  - forum (news + conference schedule)
  - useful links
  - several dedicated lists
- A repository
- A lot of services on the Web to give direct access to resources

Resources (1)

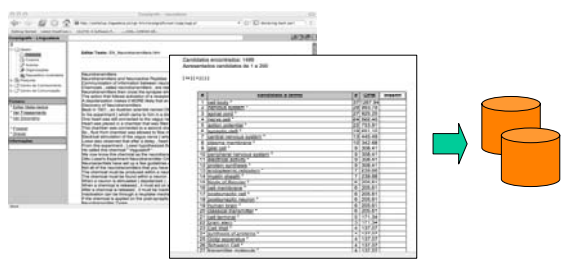
- Give access through the Web
  - Raw material
  - Environments
  - Tools
- Download
  - Corpora
  - Environment
  - Tools
- Note: until 2004 we still distributed corpora on CD ☺

Resources (2)

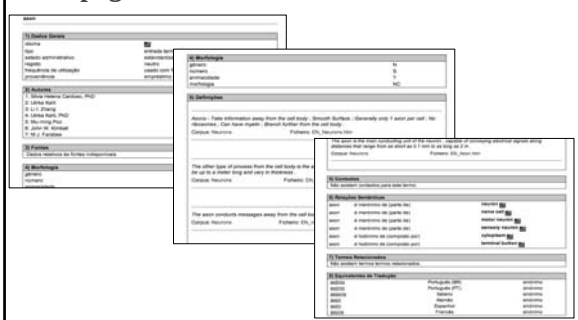
- Corpógrafo: an environment to do terminology in a professional way
- Esfinge: a QA system
- A robust tokenizer and sentence separator for texts in Portuguese
- NATools: word and sentence aligners
- WebJspell: interactive spellchecking, of words, texts and Web pages

- CETEMPúblico, CETENFolha
- Floresta Sintá(c)tica: the first treebank fro Portuguese
- COMPARA: the largest proof-read parallel corpus in the World

Corpógrafo: from text to term database



Corpógrafo: what is stored in the database?



Corpógrafo: Data can be exported in XML

Corpógrafo: success indicators

- More than 1,000 registered users
- Used for translation and terminology teaching in several departments
  - Porto, Aveiro, Faro, Braga, Lisbon, Coimbra, São Paulo, ...
  - Belinda's travel and teaching around Europe: Finland, Germany, UK, Turkey
- Separately installed in Catalunia, one of the European excellence centers for terminology
- Dissertations using Corpógrafo starting to appear everywhere (Bulgaria, Barcelona, ...)

Evaluation contest (*avaliação conjunta*)

- Jointly agree on a task and discuss the details together
- Create an evaluation setup
  - measures
  - resources
  - procedure
- Compare the performance of the several systems and get a state of the art
- Make public both resources, programs and systems' outputs for
  - external validation
  - research on both the task and the evaluation methodology
  - organization of future evaluation contests
  - training of newcomers

Further advantages of an evaluation contest

- Agree on details that generally make individual evaluation measures incommensurable
- Raise awareness about a particular task, its problems and solutions: community building
  - several new systems were born with HAREM
- Produce a wealth of documentation that otherwise would never have been produced
  - cf. HAREM guidelines; cf. the wide discussion of particular morphological problems and solutions; the discussion around QA systems in CLEF
- Can provide baselines and resources (systems, gazetteers) for other work

Evaluation: Morfolimpíadas (2003-2004)

- Compare the output of morphological analysers for Portuguese (for isolated words)
- Golden collection
- Dedicated evaluation architecture and measures
- 7 systems (3 Portugal, 2 from Brazil, 1 France, 1 Denmark)

- 1 book (in print): Diana Santos (ed.). *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*, Lisbon: IST Press, 2007.

CLEF: 2004, 2005, 2006
Main results of Linguateca@CLEF

- CLEF: Cross Language Evaluation Forum (> 100 international participants in 2006)
- The number of systems trying their luck in Portuguese has increased each year (numbers next stuntlunch)
- Creation of the CHAVE collection, freely available.
- In addition to the text (full documents from major Portuguese and Brazilian newspapers, 2004-2005), this collection makes available topics and their relevance judgements, as well as questions and their answers, providing plenty of evaluation material for new systems.

Evaluation: HAREM

- **N**amed **E**ntity **R**ecognition: identify a proper name (NE) and classify it in context according to a set of semantic categories.

Example:

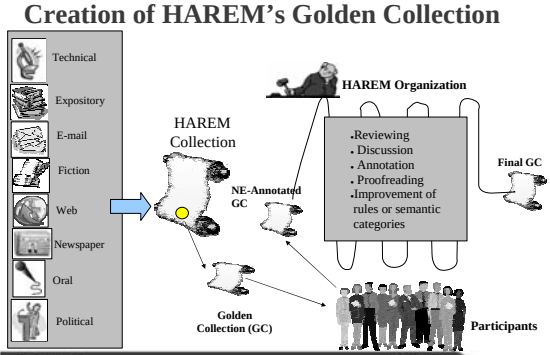
```
Eça de Queirós nasceu na Póvoa de Varzim em 1845, e
faleceu 1900, em Paris. Estudou na Universidade de
Coimbra.
```

```
Eça de Queirós nasceu na Póvoa de Varzim em 1845, e
faleceu em 1900, em Paris. Estudou na Universidade de
Coimbra.
```

Semantic categories:

Place*, *Time*, *Person*, *Organization

Creation of HAREM's Golden Collection



HAREM's success criteria

- Higher participation: 10 systems
  - 1 from Mexico, 1 from Spain
- Statistical validation
- New semantic model, new measures
- Relevant empirical studies on their way
- A new book under development

  Diana Santos & Nuno Cardoso (eds.), *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*. Linguateca: 2006.

SUPeRB (Sistema Uniformizado de Pesquisa de Referências Bibliográficas, ou SUPERBibliotecário)

- Primary motivation:
  - help manage, (currently 1,300 entries)
  - improve (we estimate 10 times more out there)
  - foster participation of users
- of a publication catalogue (o *Catálogo de publicações da Linguateca*)

- Easy to understand the general interest, beyond our practical problem: a bibliographical manager assistant for specific areas or purposes
- three kinds of further users: librarians, teachers and students

SUPeRB: An assistant for a bibliographic manager

- **Finds** relevant references on the Web
- **Parses** those references assigning meanings to their parts
- **Gathers** further bibliographic information
- **Validates** that information
- **Merges** with previous partial or known info
- **Classifies** the references and standardizes them
- **Updates** the catalogue

A modular toolbox

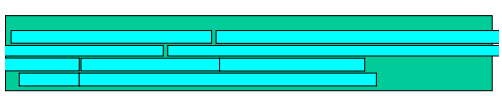
with rich display facilities

The sad life of a publications catalogue manager

- Manually adding references from:
  - Team members and visitors to Linguateca via Web form/mail
  - Unsystematically searching in e-mail lists, conference proceedings on the Web, home pages
- Manually validating references by
  - completing missing elements from references
  - finding the documents or the proceedings to clarify unclear data
- Updating information
  - To be published -> when and what is the last version
  - Change of URL
  - Republished in...

SUPeRB: Some terminology

- Bibliographic references and elements

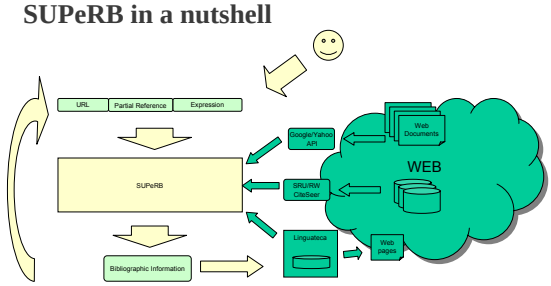


Bibliographic references
Bibliographic elements

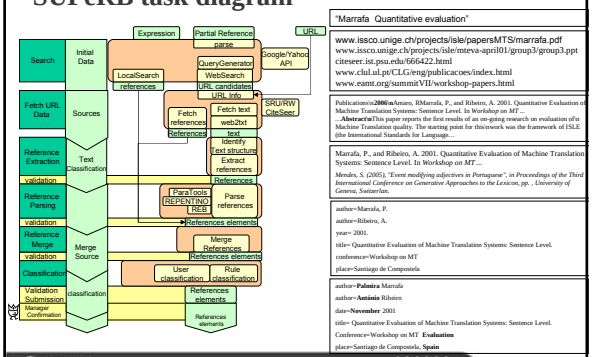
- Bibliographic formats (BibTeX, RIS, EndNote)

```
@inproceedings{marrafa:ribeiro:MIS01,
        author = {Palmira Marrafa and António Ribeiro},
        title = {Quantitative Evaluation of Machine Translation},
      booktitle= {MI Summit VIII: Fourth ISLE Workshop},
      year = 2001,
       page= {39--43}
}
```


SUPeRB in a nutshell



SUPeRB task diagram



SUPeRB internals: or why can this be a Master's Thesis

- Internet search (through Web services)
  - general search engines
  - specialized publication search engines
- Information extraction
  - references from text
  - semi-structured information from the references
- Classification (approaching Web 2.0)
  - tags proposed by the users
  - different (publication) classification schemes ("ontologies")
- Integration (of data and methods)
  - crucial in any software engineering project, generally despised by academia

Concluding remarks – and future

- 2000-2003 We first tried to establish us as important in a Portuguese (language) context
- 2003-2005 Then we became well-known in international circles as the "guardians of Portuguese"
- 2006-2008 Now we are trying to do ground breaking work at an international level, with e.g. HAREM, QolA and Queryonomy

Previous presentations of Linguateca at SINTEF

- Nuno Seco. "Building a Large Scale Lexical Ontology for Port." (16 Aug 2006)
- Diana Santos, Luís Costa & Luís Cabral. "Linguateca, a distributed resource center for language technology for Portuguese". (Hafjell, April 2006).
- Luís Sarmento. "Taming the Web... and other less wild subjects" (November 2006)
- Paulo Rocha & Diana Santos. "Portuguese at CLEF". (27 April 2004)
- Belinda Maia & Luís Sarmento. "Linguateca@Porto". (10 Sept 2003)
- Diana Santos & Luís Costa. "Linguateca activities". (14 May 2003)
- Rachel Aires. "Who am I? What am I doing here?" (May 2002)
- Paulo Rocha. "Creating CETEMPúblico" (?) (2000?)
- Diana Santos. "Internet access to Portuguese corpora" (November 1999)

www.linguateca.pt

- More information on www.linguateca.pt
- Next stuntlunch: next Wednesday, 29 November 2007

On Question Answering in general and Linguateca's role in it in CLEF

Questions? Comments? Feedback?