

## HAREM

### The first evaluation contest for Named Entity Recognition in Portuguese

Diana Santos  
Linguatca  
[www.linguatca.pt](http://www.linguatca.pt)

## Structure of the talk

- A light introduction to NL(P)
- Very brief presentation of Linguatca
- Evaluation contests
- Named entity recognition
- HAREM

## What is natural language (processing)?

- Natural language is the oldest and most successful knowledge representation language
- Used for communication, negotiation, and reason (->logic)
- Main features:
  - vagueness
  - context-dependent
  - implicit knowledge
  - evolves/dynamic/creative
- Different natural languages
  - different world view
  - different glue/implicit

## What is NL processing?

- Using computers to do things with natural language
- to be useful for humans
- Most intelligent human tasks involve language
  - as center (communicating, teaching, converting)
  - as periphery (mathematics papers, medical diagnosis)
- Daily tasks
  - writing (and creating or conveying information or affection)
  - reading (and finding information)
  - translating (and mediating)
  - teaching and learning and documenting
- Enormous political impact

## Linguatca, a project for Portuguese

- A distributed resource center for Portuguese language technology
- POSI project with FCCN as main contractor (2000-2006)
- First node at SINTEF ICT, Oslo, started in 2000 (work at SINTEF started 1998 as the *Computational Processing of Portuguese* project)

### IRE model

- Information
  - Resources
  - Evaluation
- [www.linguatca.pt](http://www.linguatca.pt)



## Linguatca highlights, [www.linguatca.pt](http://www.linguatca.pt)

- > 1000 links More than 1,500,000 visits to the Web site
- [AC/DC](#), [CETEMPúblico](#), [COMPARA](#) ... Considerable resources for processing the Portuguese language
- [Morfolimpiadas](#) The first evaluation contest for Portuguese, followed by CLEF and HAREM
- [Public resources](#)
- [One language, many cultures](#)
- Foster research and [collaboration](#)
- Cooperation using the [Internet](#)
- Formal [measuring](#) and comparison
- Do not adapt applications from English

## Linguateca news

- Organizing a summer school about the computational processing of Portuguese: July 10-14th 2006 in Porto
- Organizing CLEF 2006 for Portuguese
- Organizing mini-HAREM at this very moment

## 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

## The task, the problem

- NER = Robust identification and classification of proper nouns in running text -- in Portuguese
- Applications:
  - IR: indexing and retrieving
  - MT: translating properly
  - Text understanding, and building resources from text
  - etc.
- History: well known task from MUC (Message Understanding Conference), used in CoNLL, re-formulated in ACE, TERN etc.
- Our translation/appropriation: REM, *reconhecimento de entidades mencionadas*

## Is it the same task? Just Portuguese

- Is different language relevant?
- Just change of modules (tokenization, spelling) and resources (gazetteers)? Minor adaptations...
- Or a different language has different challenges? Different things people talk about, different typographical conventions, different conceptualization of the world...
- This is basically an empirical question...

## The same task? Methodological questions

- What are the set of classifications we are interested in?
- How do we agree on their interpretation?
- Is extension to other text genres relevant?
- Is the NE concept (*entidade mencionada*) even delimited the same way? the operational criteria are the same?...
  - partial identification
  - ontological nearness
  - spelling errors, different varieties
- Is extension to other sorts of classification relevant?
- How do we handle indeterminacy, and disagreement? (ceiling effects)

## For NLP-ignorants, what's the problem? Flagging proper names in text?

- Well, the same proper name in different contexts...  
O Brasil venceu a Copa (PESSOA<sub>GRUPO</sub>). O Brasil assinou o tratado (ORGANIZACAO<sub>ADMINISTRACAO</sub>). O Brasil tem muitos rios (LOCAL<sub>ADMINISTRATIVO</sub>). Por amor ao Brasil (ABSTRACCAO<sub>IDEIA</sub>). ...
- Or a different one which happens to be equal... *Camilo Castelo Branco*
- Not all occurrences are equally obvious to classify
  - Guimarães tinha muito poder junto do governo naquele tempo
  - Caros amigos dos **Bombeiros**
  - disse ontem em entrevista à revista **Playboy**
  - o certificado **ISO-9001** atestou seu nível de qualidade internacional
  - o **Brasil** da metade do século XIX não diferia muito da...
  - as três repúblicas que surgiram da divisão da **Bósnia**
  - Hoje a **Sé** está completamente diferente por dentro

## What's the problem? (contd.)

- Not all occurrences are equally obvious to identify
  - licenciada pelo Ministério da Indústria do Governo cessante
  - doação de terras a senhores da nobreza, concretamente com as Honras de Cardoso, de Cantim, de Fonseca ...
  - tirada dos Jardins deste Palácio, que era Episcopal, depois passou para Biblioteca Pública e depois para a Universidade do Minho
  - Eu não posso deixar de louvar a atitude de V.Exa., prestando assim esses informes à Casa,
  - de acordo com as Convenções das Nações Unidas
  - para a realização de uma História da Imprensa em Macau
  - não herdei a vontade de ser Monárquico
  - lutou contra a Ditadura de João Franco
  - pegar avião na ponte Rio-São Paulo

## Delimitation criteria

- The abstract goal: extract every thing which has a name, and assign it the correct classification in context
- First problem: most names are part of longer strings
  - constante de Planck
  - ministro da Defesa
  - pasta dos Negócios Estrangeiros
  - dona da barraca das faturas da Feira Popular
- Second problem: names can be compositional and therefore refer to different things simultaneously
  - Centro de Lógica e Computação do Departamento de Matemática do Instituto Superior Técnico

## Delimitation criteria (contd.)

- Third problem: names do not always appear complete
  - a Revolução de 30 e a de 33
  - o ministro da Educação e a da Ciência
  - a Santa Casa
- Fourth problem: capitalization is almost random!
  - que assolam a freguesia de Ferreiró -- um bastião **Socialista** --
  - o Pinto Machado que quis fundar a **faculdade de Medicina** e que agora está à frente.
  - diz ela. (Do artigo **Fonte da juventude**, publicado em *Veja*, 25 de julho de 1990)
- Fifth problem: errors occur...
  - cuja verba ronda os **150 ecudos por metro quadrado**
  - Quantos anos esteve em **Biblau** ?

## HAREM: the first evaluation contest in named entity recognition in Portuguese

- Process
  - Agreement on the categories and subtypes employed, as well as on the tasks
  - Common compilation of a golden resource (manually annotated with NEs)
  - Deploying an evaluation setup architecture, for automatic comparison of system outputs over a large text collection
  - Producing results according to several criteria
- Event
  - Three tasks: identification, morphological and semantic classification
  - Contest run 14-16th February 2005: 10 participants (5 countries), 18 runs
  - Different winners in different measures
  - HAREM workshop scheduled for May 2006
  - repetition of HAREM (mini-HAREM) in April 2006 for studying statistical reliability and systems' progress

## Three main axes

- Compiling the golden collection: what is right, how to express it
- Developing the evaluation environment (a set of general modules with several options in order to try out several ways of ranking systems and dealing with this kind of problem, etc.)
- Making sense of the results
- The three things are obviously connected



## The future of HAREM, February 2006

- We are still organizing the final workshop, after a rerun for statistical testing (mini-HAREM)
- We expect to add further challenges to further editions
- We expect more and more participants also with different research aims: GIR, ontology learning, semantic interpretation, ...
- We hope for more mathematically oriented research round this kind of events, after enough data has been gathered