

# Portuguese at CLEF: Reflections and Challenges

Diana Santos & Nuno Cardoso

[www.linguateca.pt](http://www.linguateca.pt)



**Linguateca** is a distributed resource center for the processing of the Portuguese language, following the **IRE** model

- I** Information <http://www.linguateca.pt>
- R** Resources (CETEMPúblico, COMPARA, AC/DC, Floresta treebank, Corpógrafo, etc.)
- E** Evaluation (AvalON, Morfolimpiadas, **CLEF for Portuguese**, HAREM, etc.)

**The quality of an evaluation campaign depends crucially on the evaluation setup (resources and method). To prepare evaluation resources for Portuguese NLP, one must know well the challenges involved:**

## ImageCLEF

### The most culture- and language-dependent task

Please give me labels able to recover pictures like the following:



Collections: captions are essential to give unity.  
Captions (small sized pieces of text) are extremely difficult to translate.

## GeoCLEF

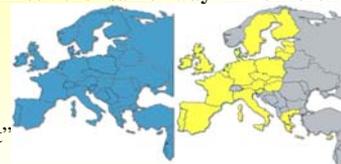
### Conception of the space encoded in language

#### Lack of a formal definition

What are geographical relations?  
What are meaningful geographic questions?

#### Is geographic knowledge conceptualized the same way in different languages?

What is Europe?  
What is Latin America?  
What is North America?  
“near Canada” vs. “at the Canadian coast”



#### Vagueness present in topic formulation

reports on human rights in Latin America  
environmental concerns in and around the Scottish Trossachs

## WebCLEF



Are topics representative of any users' needs?  
Is the subcollection representative of the “government” Portuguese Web?  
Is this subset a good choice in an European context?

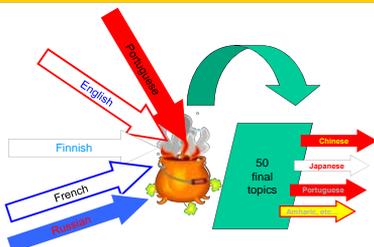
## The organizer's paradox

If one is only doing translation/ adapting queries and answers to queries to a multilingual context, and not also adding a new language collection to the evaluation setup, literal translation (as opposed to idiomatic translation) radically improves system's performance.

So, **the more natural we render a translation, the better for the human user, the worse for a cross-lingual automatic system.**

**Still, many challenges are involved. We mention improvements to the more mature tracks:**

## adhoc CLIR

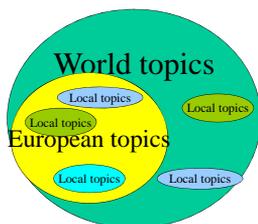


### Kinds of topics

- Cyclic events
- States of broader events
- Impact measures (including same individual)

### Geographical/cultural scope

Same language, different culture  
Balance in Brazilian/Portuguese news collections achieved



### Still not answered:

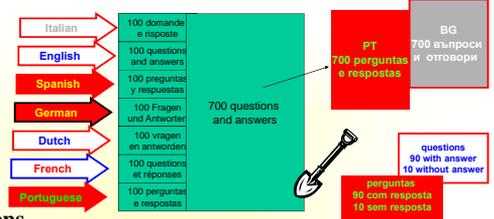
#### What is a CLIR topic?

How to choose topics from a CLIR perspective?

What is a typical user query that is better satisfied by a multilingual collection?

Do kinds of topics vary with language and/or culture?

## QA@CLEF



### Selection of questions

What is a natural question?  
What is a naturally cross-lingual question?  
Definition questions ill-defined.  
Temporally-restricted questions not well argued for.

### Evaluation of questions

Assess question difficulty given the collection.  
Assess question danger.  
Require more specific and natural justification of answers  
Attempt to define precise evaluation criteria for definition questions.



## New CHAVE

Brazilian (*Folha de São Paulo*) collection added  
>200,000 documents  
>90,000,000 words

Drawings created by Paulo Rocha for poster at CLEF'2004