### Linguateca, a distributed resource center for the processing of the Portuguese language

Promote NLP of Portuguese through the IRE model

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

### Topics for CLIR

#### Kinds of topics

- Cyclic events
- Once-only events
- States of broader events
- Impact measures
- Atemporal events

#### World topics

- European topics

#### Actual procedure

![Diagram of CLIR process]

### Questions and answers for Q&A

#### Comments on question selection

- What is a natural question?
- How to deal with presuppositions?
- What does question classification mean? (ex. 558 582)
- Definition questions are ill-defined (ex. 180)
- Answer strings: articles, gender and redundancy
- Why not confirmation questions (yes/no)?
- Irrelevant questions

#### Comments on the translation task

- Cross-cultural difficulties
- Fragments without context
- Translations without purpose (ex. 480)

#### General comments

- No awareness about questioning a multilingual collection
- No integration with topic preparation
- No grounding on real users’ questions

#### CHAVE, the Portuguese collection

**Texts from daily PÚBLICO, 1994, 1995**

106,821 documents, 348Mb

ID Format: PUBLICO-19950717-032

Used the IMS Corpus workbench to prepare and check both topics and questions

<table>
<thead>
<tr>
<th>Portuguese</th>
<th>English</th>
<th>Finnish</th>
<th>French</th>
<th>Russian</th>
<th>Portuguese</th>
</tr>
</thead>
<tbody>
<tr>
<td>700 questions and answers</td>
<td>100 domande e risposte</td>
<td>100 questions and answers</td>
<td>100 preguntas y respuestas</td>
<td>100 Fragen und Antworten</td>
<td>100 vragen en antwoorden</td>
</tr>
<tr>
<td>90 with answer</td>
<td>100 domande e risposte</td>
<td>100 questions and answers</td>
<td>100 preguntas y respuestas</td>
<td>100 Fragen und Antworten</td>
<td>100 vragen en antwoorden</td>
</tr>
<tr>
<td>10 without answer</td>
<td>100 domande e risposte</td>
<td>100 questions and answers</td>
<td>100 preguntas y respuestas</td>
<td>100 Fragen und Antworten</td>
<td>100 vragen en antwoorden</td>
</tr>
</tbody>
</table>

### Suggestions

- Stronger connection between I&R and Q&A
- Study the weight of local topics in the collections as well as study different kinds of topics
- Create a multilingual corpus to ease the verification of topic references simultaneously
- Review question categorization and choice
- Review the form of answer strings
- Redefine definition track
- Accept confirmation (yes/no) questions
- Do user studies with CLIR topics and real questions asked