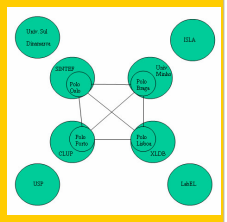
CHAVE: topics and questions on the Portuguese participation in CLEF

Diana Santos & Paulo Rocha
www.linguateca.pt

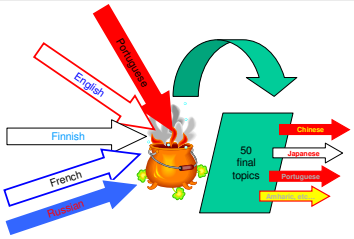


Linguateca, a distributed resource center for the processing of the Portuguese language

Promote NLP of Portuguese through the IRE model

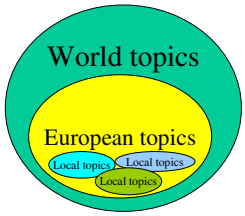
I Information http://www.linguateca.pt
R Resources (CETEMPúblico, COMPARA, AC/DC, Floresta treebank, Corpógrafo, etc.)
E Evaluation (AvalON, Morfolimpíadas, CLEF for Portuguese, etc.)

Topics for CLIR

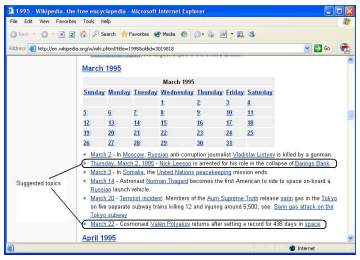


Kinds of topics

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



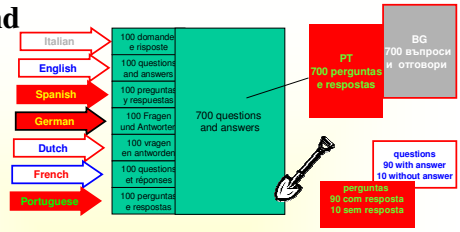
Actual procedure



Questions: 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?
How to assess the impact of one culture in another?

How to formulate a topic?
Maximize variation of formulation
Try to render it in a meaningful way for Portuguese speakers

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

558 F OTHER Qual a nacionalidade do tenista Sergi Bruguera?
1 SEARCH[espanhol]
582 F LOCATION De que país é a escritora Taslima Nasreen?
1 SEARCH[Bangladesh]

480 F OBJECT Que produz a MCC?
1 SEARCH[o automóvel Micro Compact Car]
2 SEARCH[o "carro urbano do futuro" de dois lugares]
3 SEARCH[o carro compacto Smart]
4 SEARCH[veículos] automóveis de dois lugares
5 SEARCH[Swatchmobile]
6 SEARCH[carro urbano]

180 D ORGANIZATION O que é a maçonaria?
1 SEARCH[uma sociedade secreta]
2 SEARCH[movimento]
3 SEARCH[sociedade iniciática]

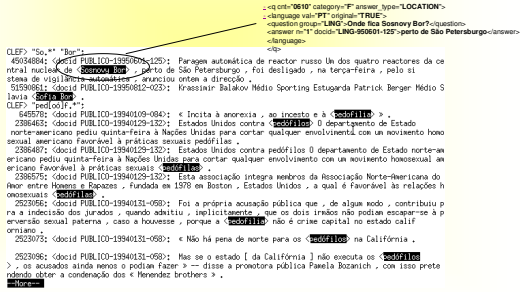
CHAVE, the Portuguese collection

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



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