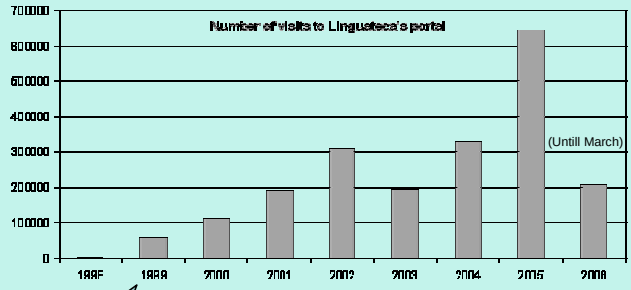
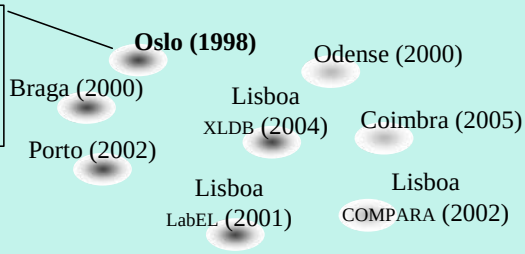
language technology for Portuguese

www.linguateca.pt

SINTEF ICT
Cooperative and Trusted Systems
Diana Santos
Luís Costa
Luís Miguel Cabral





IRE model: Information, Resources and Evaluation

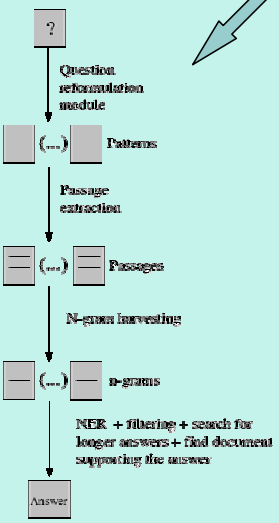
Information

❑We maintain a large web portal on the computational processing of the Portuguese language, with more than 2,000,000 visits so far.

❑ We list resources, tools and services, as well as actors and publications, and we offer a repository in the area.

❑ We also answer questions and help users about any related subject.

❑We make available already existing resources and develop new, as well as their full documentation.



The architecture of Esfinge

Resources

❑ Corpora (large bodies of text):
- ➢ AC/DC: allows one to query syntactically annotated texts (up to 250 million words) online
- ➢ COMPARA: the largest post-edited parallel corpus in the world: Portuguese and English source texts and their translations
- ➢ Floresta sintá(c)tica: treebank
- ➢ CETEMPúblico, CETENFolha

❑ IR collections
- ➢ WPT03: all Portuguese Web
- ➢ CHAVE: newspaper doc.s and topics

❑ Tools
- ➢ Question answering (Esfinge)
- ➢ Named entity recognition (SIEMÊS)
- ➢ Tokenizers, sentence separators
- ➢ Morphological analysers (AnELL)
- ➢ Spellcheckers (Jspell)
- ➢ Word aligners (NATools)

❑ Other resources
- ➢ Corpógrafo (a full-fledged system for terminology and knowledge management)
- ➢GKB (*Geographic Knowledge Base*) and Geo-Net-PT01
- ➢ REPENTINO: a NER gazetteer
- ➢ BACO: database of collocations

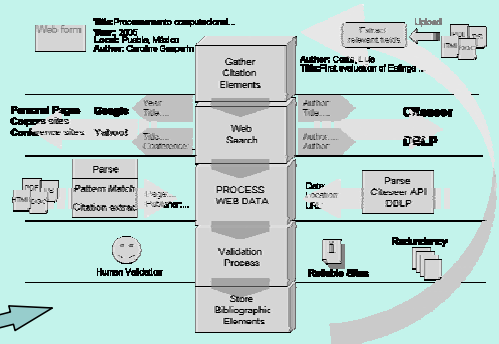
❑ Research tools or resources
- ➢ Example-based machine translation
- ➢ Ontology extraction from text
- ➢ Ontology building from dictionaries
- ➢ SUPERB: Extraction and quality checking of publication citations

Evaluation

❑ Organization of evaluation contests
- ➢ Compare several systems around a shared task
- ➢ Create evaluation resources
- ➢ Create evaluation programs
- ➢ Organize a workshop to discuss the results and the evaluation

❑ Evaluation contests
- ➢ Morfolimpíadas (morphological analysis out of context): 2003
- ➢ CLEF for Portuguese (Cross-language Information Retrieval, QA, geographic IR, WebIR, ImageIR): 2004, 2005, 2006
- ➢ HAREM (Named entity recognition): 2005, 2006

❑ Other evaluation activities
- ➢ MT from English into Portuguese: evaluating the performance of actual Web translation engines
- ➢Unobstrusive user evaluation of Web services
- ➢Component evaluation of Esfinge



The architecture of SUPeRB

SINTEF Diana.Santos@sintef.no Luis.Costa@sintef.no Luis.M.Cabral@sintef.no