

SUPeRB: using an automated publication helper in 9012 at SINTEF ICT

Luís Miguel Cabral
Diana Santos & Luís Costa

Goals of this presentation

- Present SUPeRB as a useful tool for the department
 - For everyone
 - For project and group leaders specially
 - For Stine and/or publication managers
 - For Bjørn and other forskningsjefer
- Request feedback from the audience

In a nutshell

- SUPeRB is (among other things)
 - a publication manager
 - a set of tools for publication-related tasks
 - a publication repository
 - a publication-related environment
 - publication-related ontologies
 - a Web service that provides publication related info

History of SUPeRB

- Born from the need to maintain a catalogue of publications devoted to the computational processing of Portuguese
- Development started in 2000 by Paulo Rocha, who basically created a home-made database and HTML interface pages with a set of predefined obligatory attributes
- Luís Miguel took over since he joined Linguateca's node at SINTEF in 2005
 - wrote his MSc dissertation on it, 2007
 - continued development since both for the Natural Language Technologies Group at SINTEF and for Linguateca

Publications or presentations on SUPeRB

- Luís Miguel Cabral. SUPeRB - **S**istema **U**niformizado de **P**esquisa de **R**epreferências **B**ibliográficas. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto. Março 2007.
- Luís Miguel Cabral, Diana Santos & Luís Fernando Costa. "SUPeRB: Building bibliographic resources on the computational processing of Portuguese". In *Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies* (Curia, Portugal, 10 September, 2008).
- Luís Miguel Cabral, Diana Santos & Luís Fernando Costa. "SUPeRB - Gerindo referências de autores de língua portuguesa". In *VI Workshop Information and Human Language Technology (TIL'08)* (Vila Velha, ES, Brazil, 28-29 October 2008).

Relevant design options

- Each task can be user-supervised
- Multilingual capabilities (PT, EN, NO) to present references
 - places, dates
 - conventions
 - sorting
- Scheduling for publication-related tasks
 - checking missing fields
 - making available electronic versions

SUPeRB features (1)

1. Browse a complex publication database
 - Multiple views
 - Graphical aggregation
 - Faceted navigation
2. Ease of import/export
 1. Create a SUPeRB database with a minimum work
 2. Export it to many formats
 3. Invoke in on the fly as a Web service

SUPeRB features (2)

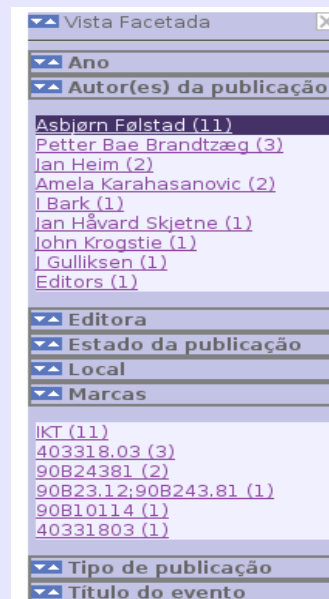
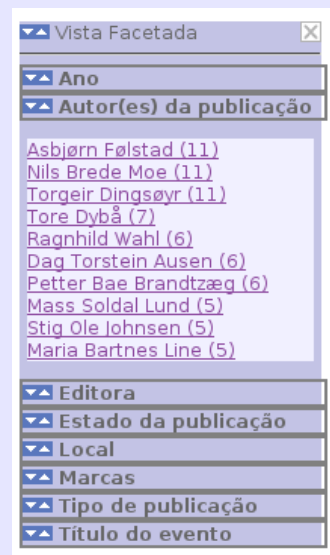
1. Add user-generated content
 - Tags such as in del.icio.us
 - Ontology translation (user-tampered-with categories)
 - On-the-fly categories
2. Allows easy reporting
 - abstract of complex questions
3. Helps maintenance
 - URLs, temporal periods where a paper cannot be published on the Web, name change

Demo / presentation

- Disclaimer: 😊
 - Some of the features may still include bugs
 - Some things may have usability problems
 - Translation into other languages has not been properly reviewed
- After all, our aim is that you think it is worthwhile to develop (=put money on) SUPeRB further on
- For this demo we gathered Ca. 240 Publications from IKT and created a new catalogue

Faceted navigation

- Allows users to refine a search based on faceted classification oriented by authors, publication category, places, date, etc.
- Allows viewing one refinement at a time instead of all search fields at once



Publications lists

- Are generate both by administrator and users as a complement of the search interface allowing to personalise search restrictions as well as the configuration for design of page

Superb can generate publications lists in different shapes and sizes

The screenshot shows the COMPARA website interface. On the left is a navigation menu with links for Home, Search, Help, Texts in COMPARA, General information, Specific documentation, Building COMPARA, The DISPARA system, Publications, Grammar annotation, Linguateca, News, and Contact us. The main content area is titled 'Publications' and includes a 'Print' button. Below this, there are sections for 'Forthcoming' and 'Regularly updated' with entries for various papers and reports.

Written documentation

[Linguateca](#)

Here you can find papers, reports and presentations created under the scope of *Linguateca* (or the former *Computational Processing of Portuguese project*). Also available are a list of [presentations](#).

1998

1 [Oksefjell & Santos 1998] Signe Oksefjell & Diana Santos. "Breve panorâmica dos recursos de português mencionados na Web". In Vera Lúcia Strube de Lima (ed.), *III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98)* (Porto Alegre, RS, 3-4. November 1998), pp. 38-47. [pdf Resumo](#)

1999

- 2 [Santos 1999] Diana Santos. "Porquê processamento computacional do português e não processamento de linguagem natural?". 24. March 1999. [html](#)
- 3 [Santos 1999] Diana Santos. "Disponibilização de corpora de texto através da WWW". In Palmira Marrafa & Maria Antónia Mota (eds.), *Linguística Computacional: Investigação Fundamental e Aplicações. Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27. May 1999). Lisboa: Colibri, pp. 323-335. [pdf Resumo](#)
- 4 [Santos 1999] Diana Santos. "Towards language-specific applications". *Machine Translation* **14.2** (1999), pp. 83-112. Dordrecht: Kluwer Academic Publishers. ISSN: 0922-6667.
- 5 [Santos 1999] Diana Santos. "Comparação de corpora em português: algumas experiências". 17. September 1999. [pdf Resumo](#)
- 6 [Santos & Ranchhod 1999] Diana Santos & Elisabete Ranchhod. "Ambientes de processamento de corpora em português: Comparação entre dois sistemas". In Irene Rodrigues & Paulo Quaresma (eds.), *Actas do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99)* (Évora, 20-21. September 1999), pp. 257-268. [pdf Resumo](#)
- 7 [Santos 1999] Diana Santos. "O computador e a tradução". 22-24. November 1999. [pdf Resumo](#)
- 8 [Santos 1999] Diana Santos. "Processamento computacional da língua portuguesa: Documento de trabalho". 1999. Versão base de 9 de Fevereiro de 1999; revista a 13 de Abril de 1999 [html](#)
- 9 [Santos 1999] Diana Santos. "Computational processing of Portuguese: working memo". 1999. [html](#)
- 10 [Santos & Oksefjell 1999] Diana Santos & Signe Oksefjell. "Using a parallel corpus to validate independent claims". *Languages in Contrast* **2.1** (1999), pp. 117-120. Amsterdam/Dordrecht: John Benjamins Publishing. ISSN: 1387-6700.

Journals

[Linguateca](#), [Publications on the computational processing of portuguese](#)

113 publications ordered alphabetically by author

```
@article{2006:vrel,
  year= {2006},
  journal= {(V)eredas : (R)evista de (E)studos (L)inguísticos},
  publisher= {Universidade Federal de Juiz de Fora},
  volume= {10}
}

@article{::sigirf,
  journal= {(SIGIR) (F)orum},
  number= {2}
}

@article{abdulkader2004:carvinglanguage:croprilina,
  author= {Inácio Abdulkader},
  title= {{C}arving (L)anguage at its (J)oints: {O}s {C}orpora como {F}erramentas de {(D)es}{T}rinchar a (L)íngua},
  year= {2004},
  a E. O. Tagnin},
  ROP} - {(R)evista da Área de (L)íngua e (L)iteratura (I)nglesa e (N)orte-(A)mericana},
  o Paulo, Brasil},
  aculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo},
  },
  2},
  www.linguateca.pt/documentos/crop10Abdu.pdf
}

@O2:restrincoesgradientes-cel,
mona Cavalcante Albano},
rições gradientes sobre relações entre vogais no léxico português},
}

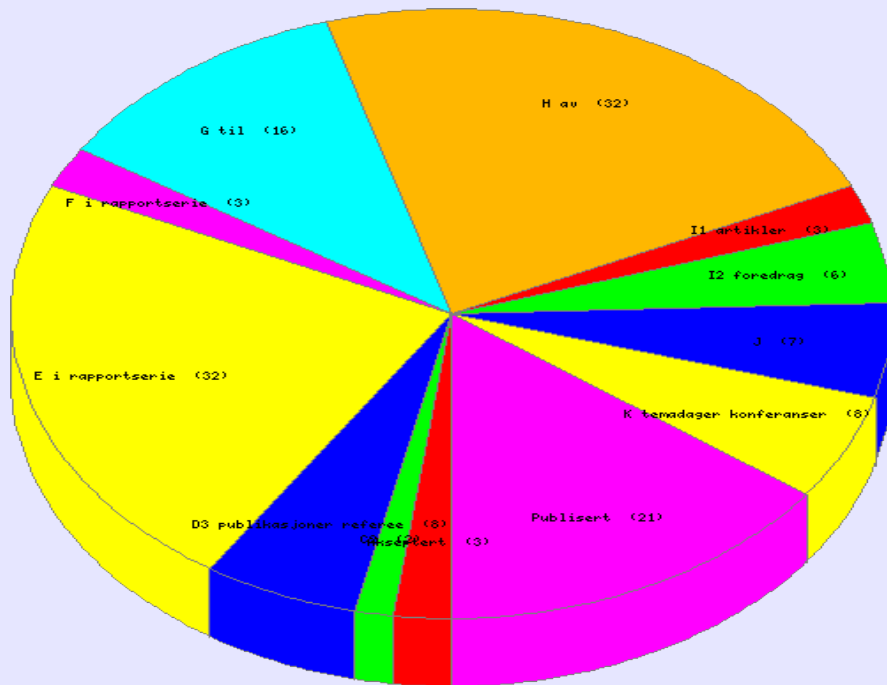
}aderno de (E)studos (L)inguísticos},
EL/Unicamp},
},
www.lafapes.iel.unicamp.br/Publicações/VV\_lex.pdf
}

@96:logicprogramming:jar,
Júlio Alferes and Carlos Damásio and Luís Moniz Pereira},
.}logic {P}rogramming {S}ystem for {N}on-monotonic {R}easoning},
}ournal of {A}utomated {R}easoning},
```

[Esta página em português](#)

Viewing results in graphics

- Publications by year (co-author, category)



Import / export facilities so far

- Supports several formats (BibTex, RIS, EndNote, Text)
- Extract references from Web documents & pages

```
@misc{cabral2006:siemesbreve:uem,  
  author={Luís M. Cabral},  
  title={{SIEMES} e uma breve introdução ao {REPENTINO}},  
  year={2006},  
  booktitle={{E}ncontro do {HAREM}},  
  location={Porto, Portugal},  
  month={15 de Julho},  
  url="http://www.linguateca.pt/documentos/SIEMES\_HAREM.pdf"  
}
```

Linguatoteca's (SUPeRB-based) present catalogue

- References included
 - Ca. 1,600 references
 - Ca. 1200 links for the electronic version(s) of the document (paper or presentation)
- Publication-related data
 - Ca 2,263 authors
 - Ca. 400 conferences
 - Ca. 250 publishers

SUPeRB

Sistema Uniformizado de Pesquisa de Referências Bibliográficas

Is an semi-automatic system which helps in the search and the managing of bibliographic information

- Extraction of new bibliographic information
- Management of the bibliographic information within a catalogue
 - Add and edit
 - Tag
 - Validate
 - Generate summary pages

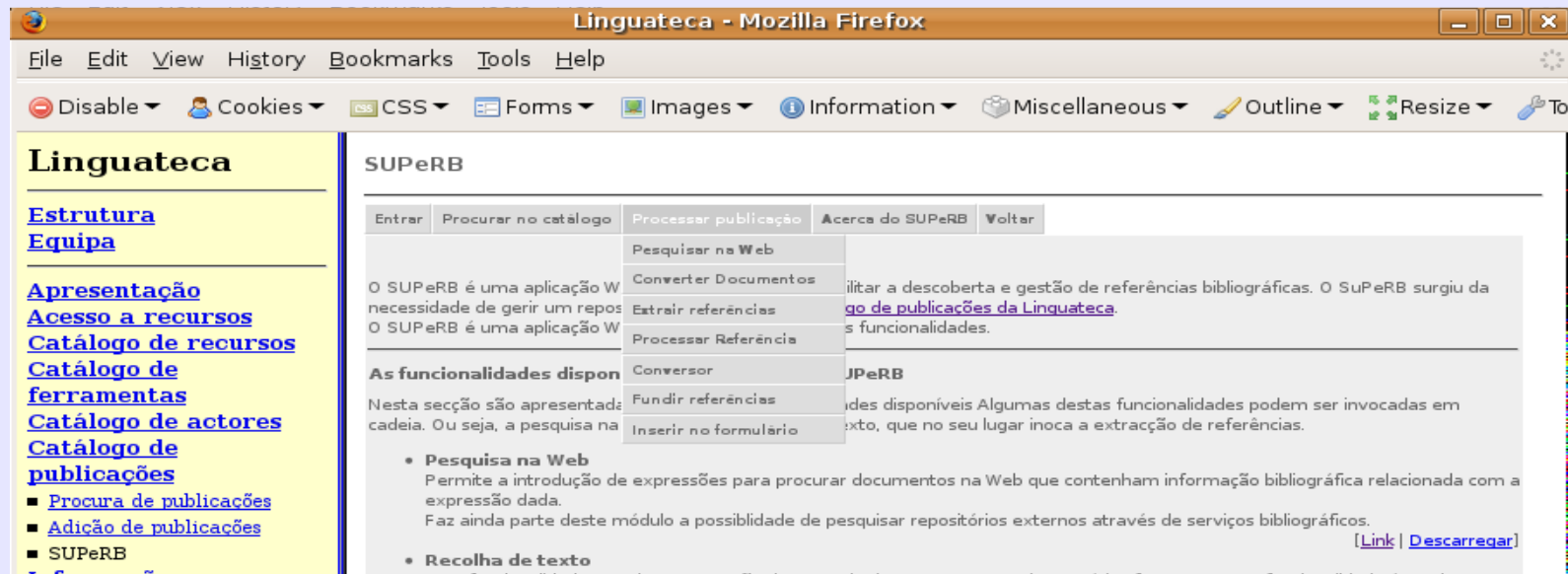
Using SUPeRB

- Two kinds of users
 - Repository user, allowed to search, suggest and tag bibliographic references
 - Repository manager, responsible for validating users actions and generate final data
- Input
 - Expression composed by a set of keywords
 - Partial reference
 - Link to a web document
- Output
 - Relevant bibliographic content, properly formatted for use or storage

Availability

<http://www.linguateca.pt/SUPeRB/>

- Manages Linguateca's catalogue
- Source code available separately
- Full documentation so far only in Portuguese



Example 1: Report to Portuguese FR

- Usually, it would be enough to select all publications marked with the tag *Linguateca*
- However, bureaucrats have often new requirements
 - Two new tags: SEMACK, AMBITO
 - Further info: when was a “to appear” publication submitted, and when was expected to appear
- Plus, reports on consecutive years should not repeat publications in the same category
 - Except for the “regularly updated” category of documentation

Editing books with many authors

- Creating a tag for each new book
- (Re)using the publications already in SUPeRB
- Merging (and validating) from several chapters
- Self-reference
 - This volume

Send group production to Stine

(to be improved)

- Find publications per kvartal (3 months)
- Use a tag for “sent preliminary version to Stine”
 - To tell her it is an update
 - To watch out for updates
- Use SINTEF categories instead of native ones

Publication production in the department

- A tag for each group (and project)
- Graphical distribution of production in the department
- Average number of publications per person, per publication channel, per project, ...
- Preferred publication arenas
- Temporal evolution of publication practice
- (Future) Search level (0,1,2) through access to the corresponding Webservice
- Export in the format required by Doris/SINTEF

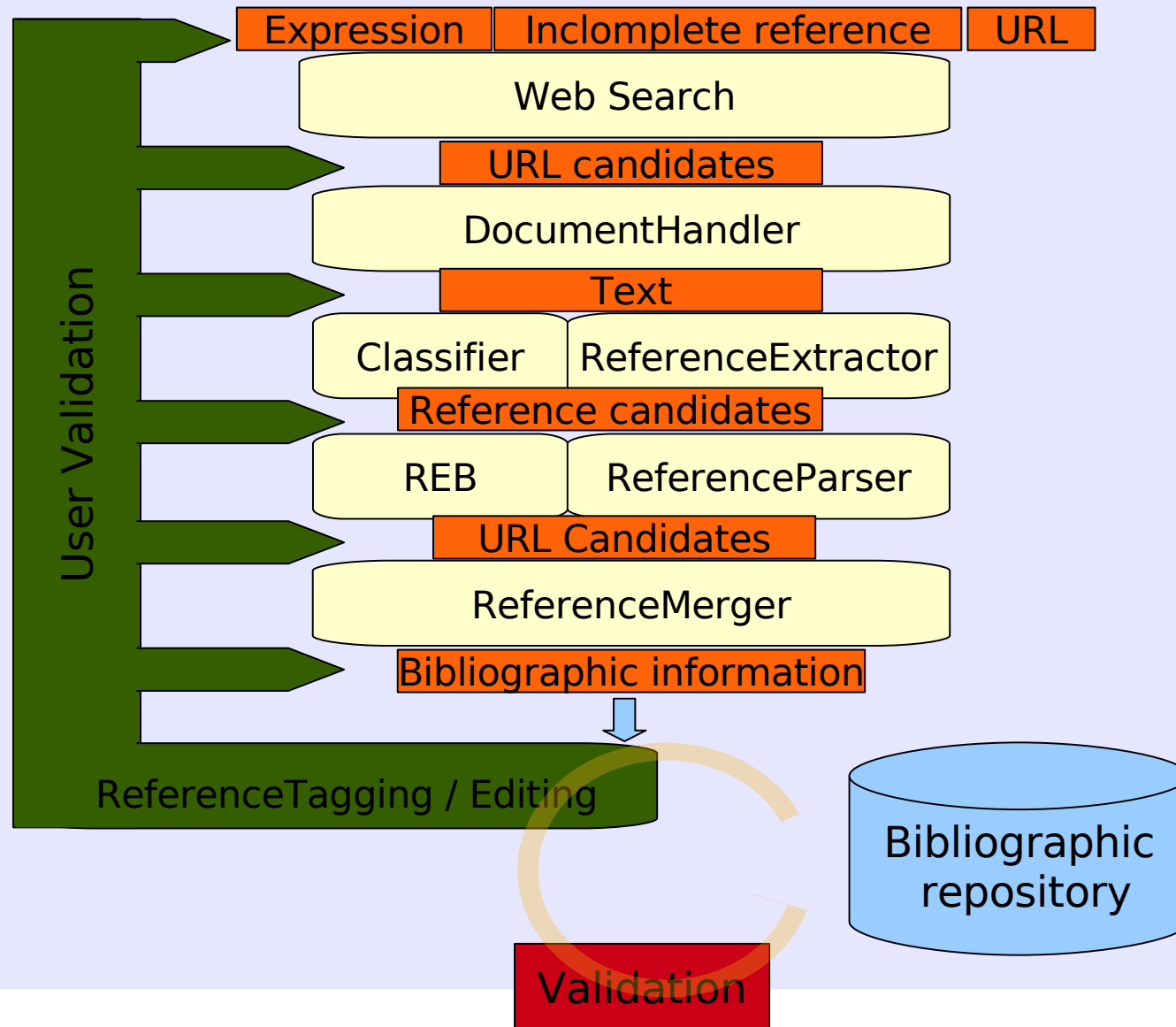
Acknowledgements

- This work was done in the scope of Linguateca, contract no. 339/1.3/C/NAC, jointly funded by the Portuguese government and the European Union,
- and improved with SINTEF internal funding



END!

Architecture



WebSearch

Searching the Web

- Module responsible for processing expressions given by users, rewriting them and generating multiple expressions
 - joins specific domain words into the expressions,
 - queries search APIs (Google API, Yahoo API)
 - Returns URLs to relevant digital documents on the Web

DocumentHandler

Convert documents into text

- Obtain the content from different file formats
 - Gets remote documents if necessary
 - Converts documents formats into text
 - Using publicly available programs
 - Can be easily configured to use other programs

ReferenceExtractor

Finding references from unstructured text

- Capable of identifying and delimiting bibliographic references
 - Relies first on classification methods to match the document structure to expected genres (article, dissertation, homepage, ...)
 - Handles each structure in a different way
 - Article's header
 - Article's reference section
 - List of references

ReferenceParser

Parsing bibliographic references

- Takes one bibliographic reference and obtains its bibliographic elements (author, date, etc.), using a combination of methods
 - Paratools (Jewell)
 - Heuristics
 - Ontology based validation (*REB*) before user validation
- *REB* (Repositório de Elementos Bibliográficos) is both:
 - An ontology of authors, conferences and places allowing relations between elements (dealing for instance with abbreviations)
 - A set of similarity detection methods (both used for updating the ontology and validating the reference)

ReferenceMerger

Avoiding duplicates

- Identifies duplicates or partial references (different parts of the same reference)
- Merges a set of partial references, combining the distinct elements in each
- Checks whether the reference is already in the catalogue and if it needs updating

Other Modules

- *ReferenceTagger*
 - Allows users to provide additional information by assigning tags
- *ReferenceConverter*
 - Provides conversion between different bibliographic formats (BibTeX, EndNote, etc.)

Concluding remarks

- Motivated by a practical problem
- System used in practice
- An example of information extraction of a specific kind of information (references)
- Potential usefulness in virtually any scientific area with Portuguese speaking authors
- Publicly available