Corpus-based indexing of a specialized Web portal in PT & EN

- Interdisciplinary work
  - Information retrieval
  - Corpus-based terminology
- Corpógrafo
  - Web-based environment for terminology work
- Busca
  - Linguateca’s site search engine

LINGUATECA

- Linguateca is a distributed language resource centre for Portuguese
- Aim: contributing to the quality of NLP resources for Portuguese
- Increasingly large website at http://www.linguateca.pt since mid 1998
  - Several on-line resources (corpora, tools, publications, etc) produced by Linguateca
  - Catalogue of resources produced by other researchers
  - 1300 web documents and 2500 external links

Busca: a simple search engine

- A search-engine for our site:
  1. Person Search (simple database query)
  2. Publication Search (simple database query)
  3. Simple keyword search (Free-text Search):
     - Processing of rtf, ps and pdf files included
     - Whole system based on CQP: “Site as a corpus”
     - All words are “alike”: no TF/IDF, no document clustering, no terminological knowledge
- Search Systems 1 and 2 are OK but not System 3 (too naive! too simple...)

How could we improve Busca?

- Our group has an extensive experience in terminology
- Terminology and IR/search-engines seem a “perfect-match”
  - BUT terminology has not been widely accepted in IR
- Our question: is the knowledge of terminologically relevant units going to help us improve Busca?
  - At indexing stage
  - At query processing stage
  - At result ranking stage
  - ....

Looking at Busca logs

- January 2003 - April 2005
- 1527 “free-text searches” queries:
  - Excluding own searches
  - Very few queries for more than 2 years!!
- Some statistics:

  Number of queries in case of the search string

<table>
<thead>
<tr>
<th>String Size</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Queries</td>
<td>200</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

  Repartition of the search strings

<table>
<thead>
<tr>
<th>Search Strings</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Searches</td>
<td>200</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Overview of queries found in logs

- **Informatics in general**
  - E.g.: “CAD”, “Pascal”, “Java”, “Autocad 2000”

- **Topics concerning Portuguese language (literature, grammar, use)**
  - E.g.: “figuras de estilo”, “verbos”, “Tipos de Sujeito Indeterminado e Oração sem Sujeito”, “verbo inacusativo”, “expressões idiomáticas”.

- **General tools or resources.**
  - E.g.: “corpora”, “dicionário”, “conjugador de verbos”

Some conclusions

- All six cases suggest that users have:
  - different goals in mind
  - different knowledge about the content of the site
- Users ARE familiar with terminological units:
  - especially noun phrases
  - use them in search expressions naturally
  - even if the TUs are inappropriate in respect to the content of our website
- Sometimes users type incomplete, ill-defined or misspelled terminological units.

Specific fields or knowledge domains.
- E.g.: “extração de informação”, “terminologia”, “semântica lexical”, “Portuguese language history”.

Queries about specific tools or resources.
- E.g.: “Cetempúblico”, “Cetenfolha” (two corpora from Linguateca), “COMPARA”, “Corpógrafo”

Queries that seem to be intended for our on-line concordance tools rather than for the search engine.
- E.g.: “sem nada”, “abonad.+”, “ansioso para”, “porém (ocorrências)”.

Overall improvements for Busca

- Each document in the site should be indexed using only the TUs it contains
- Quite easy if complete list of TUs known: the Corpógrafo may help us in this!
- Knowing all possible variants and synonyms of a given TU
- For more problematic search strings (ambiguous, incomplete) > set of TUs suggesting re-formulation to user
Empirical work

- Subcorpus - 178 files in Portuguese
- Total number of tokens approximately 1M.
- Corpógrafo > extracted and manually validated 1209 TUs

Explanation of chart

- Region 1: frequent but not widely distributed TUs. E.g.: “modelo coclear”, “taxa de disparos” - usually compound words.
- Region 2: frequent and widely distributed TUs. E.g.: “análise”, “corpus”, “modelo”, “línguística”, etc. - usually very generic TUs, and /or single words (they nevertheless have multiple possible modifiers).
- Region 3: where less frequent and less distributed TUs may be found. E.g.: “verbo intransitivo”, “relação semântica”, “vibração macromecânica”.

Items to help searches

- Synonyms Portuguese (53 pair) - E.g.: “adjetivo: adjetivo”, “bibliografia: documento: publicação”;
- Translation equivalents between Portuguese-English (107 pairs) - E.g.: “dicionário: dictionary”; “Syntaxis: syntax”;
- Acronyms in Portuguese and English (81) - E.g.: “RI: Recuperação de Informação”.

The distribution of existing POS structures (ADJ – adjective; CN – common name; PN – Proper Name; PRP - Preposition)

<table>
<thead>
<tr>
<th>POS</th>
<th>occur.</th>
<th>%</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN</td>
<td>504</td>
<td>13,4</td>
<td>vagos, graminat. sumarização automática</td>
</tr>
<tr>
<td>CN</td>
<td>226</td>
<td>13,1</td>
<td>diagrama, grama</td>
</tr>
<tr>
<td>CN + PRP + CN</td>
<td>218</td>
<td>12,6</td>
<td>corpora, dicionário</td>
</tr>
<tr>
<td>CN + PRP + CN + ADI</td>
<td>57</td>
<td>2,7</td>
<td>reconhecimento de linguagem natural, resolução da ambigüidade lexical</td>
</tr>
<tr>
<td>CN + PRP + PRP + CN + ADI</td>
<td>18</td>
<td>1,7</td>
<td>reconhecimento de língua natural, processos de aprendizado de vocabulário</td>
</tr>
<tr>
<td>CN + PRP + PN</td>
<td>19</td>
<td>1,4</td>
<td>modelo de forma/dimensão, sistema de busca, rede de Elman</td>
</tr>
<tr>
<td>CN + AD</td>
<td>20</td>
<td>1,2</td>
<td>euror, etc.</td>
</tr>
<tr>
<td>CN + AD</td>
<td>10</td>
<td>0,7</td>
<td>processamento automático de linguagem natural, resolução semi-automática de recursos lexicais</td>
</tr>
<tr>
<td>CN + AD</td>
<td>21</td>
<td>0,6</td>
<td>modelo de busca-em-bloco, modelo de busca-em-bloco</td>
</tr>
<tr>
<td>CN + AD</td>
<td>11</td>
<td>0,3</td>
<td>modelo de busca-em-bloco</td>
</tr>
<tr>
<td>Other POS structures</td>
<td>84</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

Semantic Classification 1

- Language resources. E.g.: “corpora”, “CETEMPÚblico”, “dicionário”, “Wordnet”, “COMPARA” etc.
- Tools and systems. E.g.: “anotador”, “analisador morfológico”, “Corpógrafo”, etc.
- Actions and processes. E.g.: “aquisição de vocabulário”, “extração de terminologia”, “anotação de corpora".
Semantic Classification 2

- Specific theories and models. E.g.: “modelo auditivo de Seneff”, “algoritmo de Earley”, etc.
- Linguistic concepts and phenomena. E.g.: “polissemia”, “ambiguidade lexical”, “verbo incusativo”, “advérbio de tempo”, “adjectivo”, etc.
- Disciplines or knowledge fields. E.g.: “lexicografia”, “engenharia da linguagem”, “inteligência artificial”, “semântica lexical”, etc.

Suggestions

- For:
  - Improvement of Busca’s search capabilities
  - User satisfaction.

Easier searching

- Single words
  - Suggest possible modifiers of word
  - With names of resources > to resource – e.g. COMPARA
- Mechanism to cope with different varieties of spelling in Portuguese
- Lists of synonym lists, acronym lists and translation equivalents
- Clustering of results

More suggestions

- Semantic classification of keywords + pragmatic rules of thumb
- If interested in a particular technology/tool/resource, > systems that apply or implement such a technology or function
  - E.g. - “morphology” > choice
    - “scientific discipline”
    - “applications that deal with morphology” (morphological analysers, stemmers, morphological generators, POS taggers)
    - “specific systems that perform any of these tasks” (Palavroso, PALMORF, etc.)
  - “evaluation”

More suggestions

- Manually select correct semantic classification of each TU (partially done)
- Automatic text categorization system
- Corpógrafo tools for finding semantic relations and building thesaurus/ontologies for helping navigation
- ETC

Conclusions on Interdisciplinary work

- Requires
  - Mutual understanding
  - Tolerance
  - Mental gymnastics
- Exemplified here with
  - Computer science
  - Computational linguistics
  - Terminology
Thank You!

Contact:
- www.linguateca.pt
- www.linguateca.pt/corpografo

Débora Oliveira: dmoliveir@letras.up.pt
Luis Sarmento: las@letras.up.pt
Belinda Maia: bmaia@mail.telepac.pt
Diana Santos: Diana.Santos@sintef.no