Corpus analysis for indexing: when corpus-based terminology makes a difference

Débora Oliveira
Luís Sarmento
Belinda Maia
Diana Santos
Linguateca

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

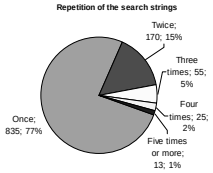
- A search-engine for our site:
  1. Person Search (simple database query)
  2. Publication Search (simple database query)
  3. Simple keyword search (Free-text Search):
     - Processing of rtf, ps and pdf files included
     - Whole system based on CQP: "Site as a corpus"
     - All words are "alike":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

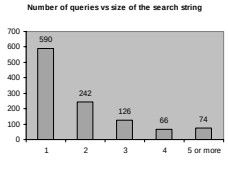
How could we improve Busca?

- 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:





What was being searched in Busca?

search string	#
Variaçoes	10
Adjunto	9
Cabeça	8
Verbos	7
Corpus	5
corpus da folha de são Paulo	5
linguagem natural	5
Peniche	5
registros doque é Conjuções coordenadas	5
Sexo	5
Tesouro	5
Tradução	5
Trail	5
About	4
Adjetivos	4
Admir	4
Árvore	4
Autor	4
Concordância	4
Consultoria	4

search string (2 or more tokens)	#
corpus da folha de são paulo	5
linguagem natural	5
Registros doque é Conjuções coordenadas	5
creme de legumes	4
ele é nada mais nada menos que um idiota	4
há momentos	4
lingua portuguesa 7%AA série	4
o cortiço	4
redação coerência e coesão	4
singno linguistico	4
Vanguardaeuropeia	4
verbos irregulares	3
adjunto adniminal	3
cetem publico um milhao de palavras	3
comparable corpora	3
concordancia verbal	3
dicionário técnico	3
emprego do artigo	3
ensino%2C portugues%2C lingua estrangeira	3
floresta sintactica	3

What was being searched in Google to get to Linguateca's site?

Search string	# queries
linguateca	832
dicionario ingles portugues on line	812
literatura infantil	625
livrarias	602
portugues para estrangeiros	582
priberam	463
compara	457
avalon	451
editoras	431
power translator	431
livrarias portugal	424
dicionario portugues ingles on line	392
dicionario portugues aurelio	391
português para estrangeiros	384
dinalivro	381
dicionario portugues	360
curriculum vitae	349
dicionario portugues ingles	334
dicionario portugues on line	315
Enciclopedias	310

Word in search string	# ocorrences
de	36151
portugues	18102
dicionario	14228
dicionário	11725
ingles	10920
download	8757
português	8419
on	8270
line	7966
para	7941
em	6746
da	5612
inglês	5349
do	5063
e	5054
online	4953
portuguesa	4230
lingua	3350
tradução	3034
Termos	2895

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"

Overview of queries found in logs

- **Specific fields or knowledge domains.**
  - E.g.: "extracção 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)".

Some conclusions

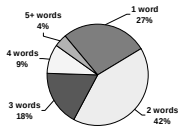
- **All six cases suggest that users have:**
  - **different goals in mind**
  - **different knowledge about the content of the site**
- **Users ARE familiar with terminological units:**
  - **especially noun phrases**
  - **use them in search expressions naturally**
    - **even if the TUs are inappropriate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

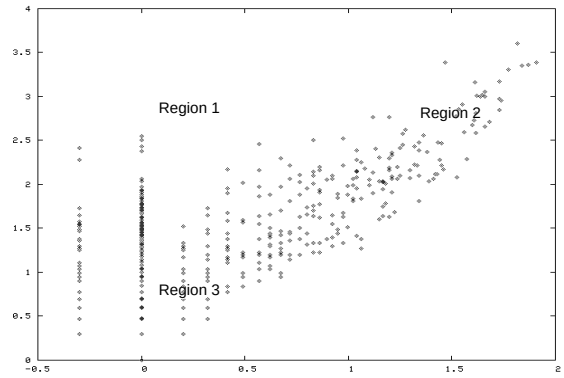
Initial improvements for Busca

- 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





Frequency and Distribution of the 1209 TUs extracted. The axis are set to logarithmic scale.

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”, “linguística”, 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 semática”,”vibração macromecânica”.

Items to help searches

- Synonyms Portuguese (53 pair) - E.g.: “adjectivo: adjetivo”, “bibliografia: documento: publicação”;
- Translation equivalents between Portuguese-English (107 pairs)- E.g.: “dicionário: dictionary”;
- Synonyms English (23 pair)- E.g.: “parsing system: parser”;
- 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)

POS	occur.	%	Examples
CN + ADJ	504	41,6	vagueza grammatical, sumarização automática
CN	226	18,7	dicionário, gramática
CN + PRP + CN	178	14,7	sistema de tradução, sinal de fala
PN	52	4,3	COMPARA, Corpógrafo
CN + PRP + CN + ADJ	37	3,1	reconhecimento de dígitos isolados, resolução da ambigüidade lexical
CN + PN	35	2,9	dicionário Aurélio, sistema Edite
CN + PRP + CN + PRP + CN	28	2,3	arquitectura do sistema de interrogações, processo de aquisição de vocabulário
CN + ADJ + PRP + CN	20	1,7	Legendagem automática de notícias, reconhecimento óptico de caracteres
CN + PRP + PN	19	1,6	modelo de Kanis-Deboer, teorema de Bayes, rede de Elman
Acronym/abbreviation	14	1,2	bd, cce, IA, lil
CN + ADJ + PRP + CN + ADJ	9	0,7	processamento automático da linguagem natural, criação semi-automática de recursos lexicais
CN + ADJ + PRP + PN	3	0,2	modelo auditivo de Seneff, modelo coclear de Goldstein
Other POS structures	84	7	

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”, “extracção 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: dmoliveira@letras.up.pt*
- *Luís Sarmento: las@letras.up.pt*
- *Belinda Maia: bmaia@mail.telepac.pt*
- *Diana Santos: Diana.Santos@sintef.no*