Working with Portuguese corpora

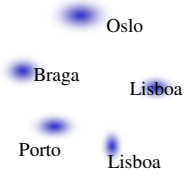
Diana Santos
Linguateca
www.linguateca.pt

Linguateca, a project for Portuguese

- A distributed resource center for Portuguese language technology
- POSI project with FCCN as main contractor (2000-2006)
- First node at SINTEF ICT, Oslo, started in 2000 (work at SINTEF started 1998 as the *Computational Processing of Portuguese* project)

IRE model
- **I**nformation
- **R**esources
- **E**valuation

Contact: Diana.Santos@sintef.no



Linguateca highlights, www.linguateca.pt

- > 1000 links More than 1,100,000 visitors to the Web site
- AC/DC, CETEMPúblico, COMPARA … Considerable resources for processing the Portuguese language
- *Morfolimpíadas* The first evaluation contest for Portuguese

- Public resources
- Foster research and collaboration
- Formal measuring and comparison
- One language, many cultures
- Cooperation using the Internet
- Do not adapt applications from English

Working with language corpora: Advantages and disadvantages

- Allow controlled exposure
- Untiring source for dialog
- Real text as opposed to artificial examples

- A lot of interference from other phenomena
- No clean text
- Too much and yet not enough

Outline of presentation

- Brief overview of corpus projects within Linguateca
  - AC/DC www.linguateca.pt/ACDC/
  - CETEMPúblico www.linguateca.pt/CETEMPublico/
  - CETENFolha www.linguateca.pt/CETENFolha/
  - COMPARA www.linguateca.pt/COMPARA/
  - Floresta Sintá(c)tica www.linguateca.pt/Floresta/
  - Corpógrafo www.linguateca.pt/Corpografo/
  - AnELL www.linguateca.pt/AnELL/
  - CorTA and TrAva www.linguateca.pt/CorTA/
- Three modes: Web access, download, upload

Web access: AC/DC and COMPARA

- Just type the URL and ask questions
- AC/DC: Web access to several different corpora, including CETEMPúblico and CETENFolha
  - newspaper text, fiction, email, non-fiction, ...
  - CETEMPúblico: 200 million words from Público (1991-1998)
  - CETENFolha (included in the NILC/São Carlos corpus): 24 million words from Folha de São Paulo (1994)
  - automatically annotated by PALAVRAS (Eckhard Bick)
- COMPARA: a parallel corpus Portuguese-English, with originals and their translations in both languages (several varieties)
  - 1 million words in each language
  - manually revised alignment

Examples of using AC/DC

- "Difficult" words for students

preterir, *premonição*, *intervindo*

- Meaning subtleties

grade vs. *gradeamento; argumento* vs. *guião*

- Collocations

claro, engraçado

- Comparative constructions

AC/DC (cont.)

- Use of modals in reported speech
- Productive sufixing
- Aspectualizers (*andar a*, *estar a*, *ir –indo*)
- Awareness of the other variant

absolutamente, *imenso*

- English interference

alegadamente, *suposto*

What's the right translation of *skuffelse*?

- skuffelse: decepção, desapontamento, desencantamento, desencanto, desgosto, desilusão, desengano, engano, frustração

Procura:
"decepção|desapontamento|desencantamento|desencanto|desgosto|desilusão|desenganо|engano|frustração".

Pedido: Distribuição das formas
desilusão 1783 frustração 1513 engano 1288 decepção 1017 desencanto 743 desgosto 715 desapontamento 354 desencantamento 30 desengano 14

Advanced AC/DC

- Relationships in Portuguese (*cujo*)
- Use of *seus* vs. *deles*
- Kinds of fights (*renhido*)
- Textual and lexical organization
  - Accident descriptions
  - Concert descriptions
  - Description of people

Contrastive studies (one half)

- Adjectives vs PP (wooden – de madeira)
- *-ing* deverbal nouns (*the moving*, *the establishing*, *the grouping*, *the computing*...)
- *viajado*, *passado* as adjectives
- formal vs. informal (noun vs. personal infinitive)
- Movement or lack of it

AC/DC internals

- The text is tokenized
- The text is sentence separated
- To each token, a set of features is automatically assigned by a parser

```
lema pos morf func deriv
```

- The corpus can thus be queried not only by the word forms but also by the values of these features

AC/DC syntax

- [feature="value"]
- [feature!="value"]
- Sequence of the above, possibly modified by {min, max}, or by * or +, meaning any number including 0 or not

```
[lema="comer"][pos="DET.*"]* [pos="N.*" & func="<ACC"]
```

- The values are described by regular expressions
  - `. any character`
  - `[a-d] any of the characters a, b, c, or d` } characters
  - `[,:?.] any of the characters ,, :, ?, or .` } characters
  - `+ one or more`
  - `* zero or more` } modifiers
  - `{2,7} at least two and at most seven` } modifiers

Published examples of using AC/DC

- Description of the project
  - Diana Santos & Eckhard Bick. 2000. "Providing Internet access to Portuguese corpora: the AC/DC project".
- Examples of using the corpora
  - Diana Santos & Elisabete Ranchhod. 1999. "Ambientes de processamento de corpora em português: Comparação entre dois sistemas".
  - Diana Santos. 2002. "*Med* e *com*: um estudo contrastivo português - norueguês".
  - Diana Santos & Luís Sarmento. 2003. "O projecto AC/DC: acesso a corpora/disponibilização de corpora"
  - Susana Cavadas Afonso. 2003. "*Clara e sucintamente*: um estudo em corpus sobre a coordenação de advérbios em *-mente*".
  - Diana Santos. 2004. "Breves explorações num mar de língua"

COMPARA

- The largest revised parallel corpus in the world
- Collaboration with Ana Frankenberg-Garcia (ISLA)
- The DISPARA system + a set of 52 text pairs
- Allow searches by
  - alignment type
  - translation notes
  - varieties of the languages
  - titles, named entities, etc.
- Fully parallel interface in English and Portuguese
- Currently being annotated

Texts	Source	Transl
Portuguese	31	20
English	18	32

Published examples of using COMPARA

- Introduction to the corpus
  - Ana Frankenberg-Garcia & Diana Santos. 2003. "Introducing COMPARA, the Portuguese-English parallel translation corpus ".
- Introduction to the project
  - Diana Santos. 2002. "DISPARA, a system for distributing parallel corpora on the Web".
- COMPARA's page of publications
  - www.linguateca.pt/COMPARA/COMPARAPublications.html
- and all on-line documentation

Other sources related to this presentation

- acdc.linguateca.pt/acesso/passeios_guiados.html
- www.linguateca.pt/Diana/usos_corpora.html
- acdc.linguateca.pt/acesso/exemplos.html
- acdc.linguateca.pt/acesso/implementacao.html

- All papers cited in
  - www.linguateca.pt/documentos/