O projecto Processamento computacional do português

Balanço e perspectivas

Diana Santos

21 de Novembro de 2000

Three main areas

- Catalogue of NLP/CL of Portuguese

  the most visible

- Resource distribution and creation

  the heaviest workload

- Evaluation

  the most difficult

Catalogue

- a large collection of links
- maintained on an almost daily basis
- with a search system associated BUSCA
- several utilities
  - consistency checking
  - new category introduction
  - visit statistics

Resources

- AC/DC project

  access to all available corpora of Portuguese with a common Web interface

  parsed corpora available for interrogation (collaboration with Eckhard Bick)
- CETEMPúblico
- COMPARA/DISPARA project

  English-Portuguese-English parallel corpus (collaboration with Ana Frankenberg-Garcia)

More resources

- Sizeable English-Portuguese translation lexicon for MT (Logos corporation)
- A treebank for Portuguese
- A set of corpus processing tools
  - sentence separation, tokenization, etc.
  - CGI for Web serving of monolingual or paralell corpora
- Documentation about corpora and their use

Evaluation

- Narrow: English-Portuguese alignment; comparing two corpus processing systems
- Broad: Verb conjugators for Portuguese

- Usability of COMPARA
- Parsing schemes for Portuguese

New projects

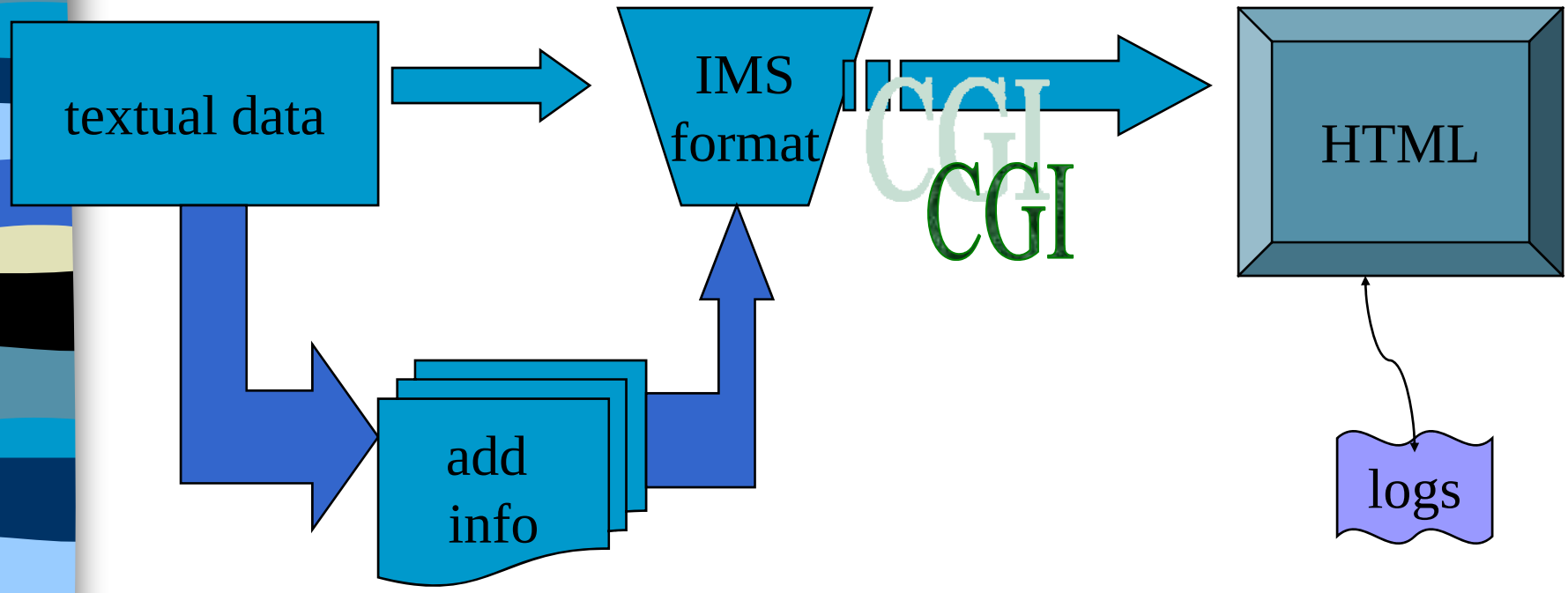
- Floresta sintáctica: A treebank for Portuguese
- Book + site on the use of Portuguese corpora
- Evaluation contests

We need to involve **all** the community working in the processing of Portuguese

Book & site on using corpora

- emphasis on methodology
- teaching how to use corpora
- eclectic in choice of corpora and encoding
- not constrained in terms of authors / subject
- inspiration: Aston/Burnard 96, Manning/Schütze 99
- international publisher (two parallel editions?)
- call and grant for chapter writing
- suggested exercises

AC/DC in one picture



Textual data ranges from simple text to SGML

Information added: from paragraph marking to parsing the text

Original features of DISPARA

- Inspection of translator's notes (N.T.)
- Search by alignment type (1:2, 1:1/2, etc.)
- Quantitative display of parallel distribution
- Search in the two directions
- Arbitrarily complex query string in both sides

COMPARA features that differ from ENPC

Different sizes and parts of the texts

Alignment unit always one source sentence

Not TEI-compliant. <title>, <foreign>, <emph>, <note> and <add> tags