

Internet access to Portuguese corpora

Diana Santos
**Computational processing of
Portuguese**
<http://cgi.portugues.mct.pt/acesso/>

09.11.99

Main goals

- **provide one place where access to all corpora is given**
- **further improve the information associated with these corpora**
- **develop a good user interface**

Rationale within the CPP project

- **make the available resources more available**
- **foster development and public availability of others**
- **provide programs to get corpora on-the-fly on the Internet**
- **create sufficiently big corpora that can be used as a reference**

Presentation structure

- **some concepts**
- **the process involved**
- **the result**
 - from the user's perspective
 - from our perspective
- **examples of envisioned use**

FOR MORE INFORMATION...

See <http://cgi.portugues.mct.pt/acesso/exemplos.html>

09.11.99

Internet access to corpora

- **Corpora: definition**
 - collections of texts (oral, spoken)
 - encoded in a format easily searchable
 - for purposes of linguistic investigation and NLP
- **Internet access**
 - Web interface to a corpus workbench

Corpus creation and dissemination

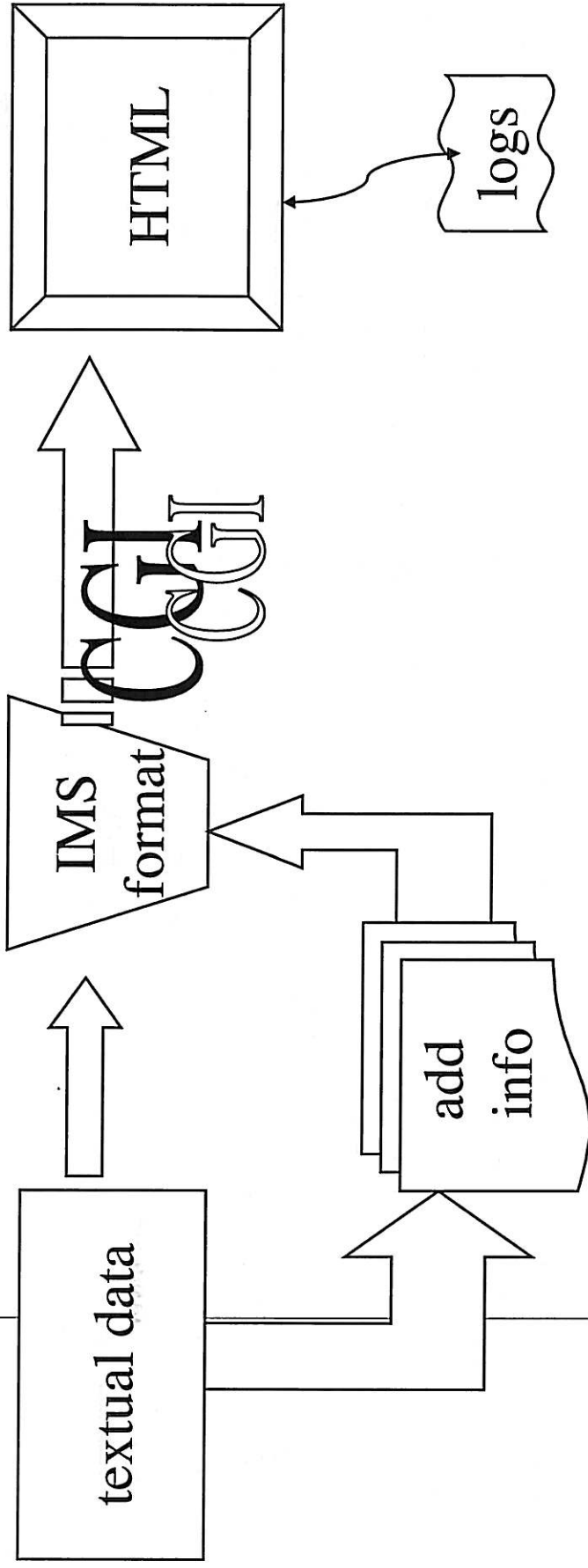
• **Creation**

- legal aspects
- technical aspects
- huge "slave work"

• **Dissemination**

- legal aspects
- technical aspects
- huge "slave work"

The process

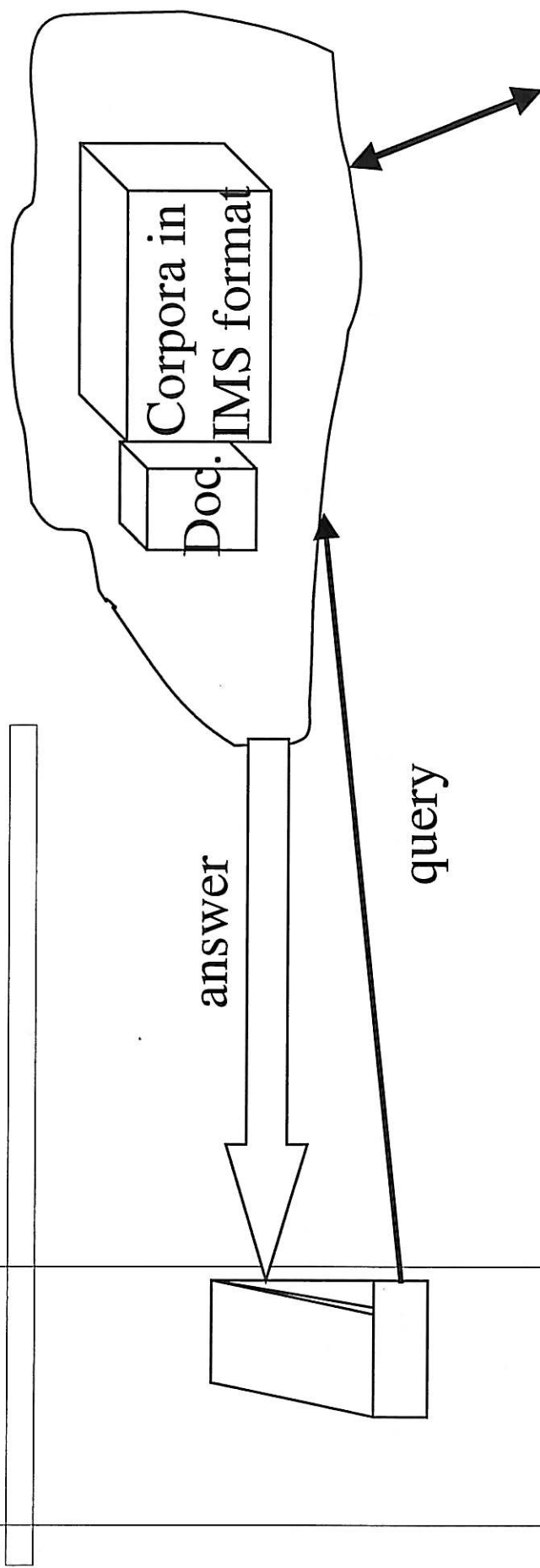


Textual data ranges from simple text to SGML

Information added: from paragraph marking to parsing the text

09.11.99

The result



Try [http://cgi.portugues.mct.pt/acesso ...](http://cgi.portugues.mct.pt/acesso...)

Documentation on the corpora; the encoding; the kinds of queries allowed; the underlying workbench; corpus uses...

09.11.99

Kinds of queries

- **based on forms**
 - **preposition use**
 - **word endings in context**
- **based on morphosyntactic information**
 - **uses of lemma *ser* + adjective**
 - **subjunctive forms not in *que*-clauses**

Kinds of queries (cont.)

- **based on functional information**
 - which kinds are typical objects of *matar*
- **based on prosodic information**
 - what words are preceded by pauses
 - which parts of sentences are overlapped
- **based on discourse information**
 - irony; narrative beginnings

Kinds of queries (cont.)

- **distribution**
 - of form
 - of genre
- **collocational analysis**
 - which verbs collocate with *casa*
- **development help**
 - which words are differently tagged
 - tagger disagreement

