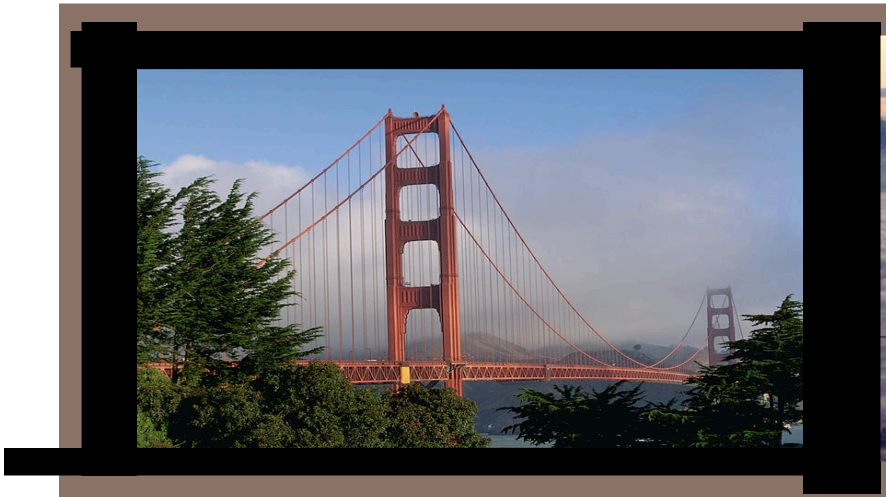
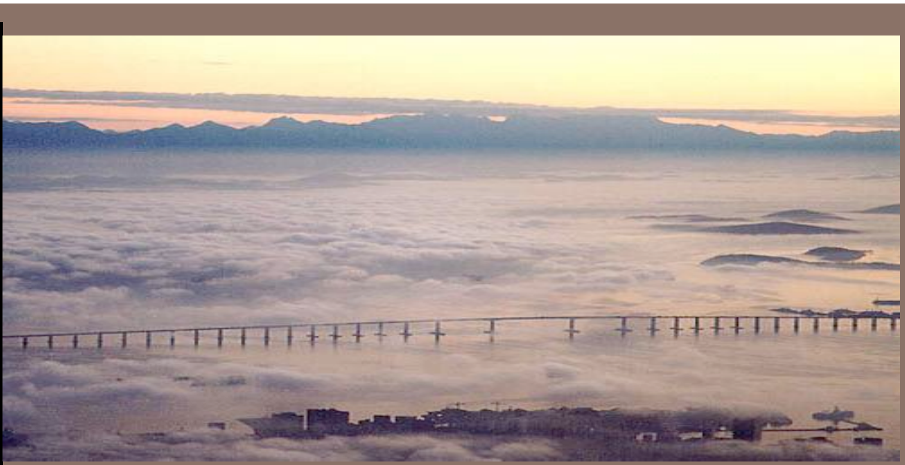




PORTUGUESE NLP FOR CPDOC/FGV?

PARGRAM/PARSEM PROJECT ANNOUNCEMENT

Valeria de Paiva

October 2010

Fundacao Getulio Vargas (FGV)

"**Fundação Getulio Vargas** (*FGV*) is a Brazilian higher education institution founded in December 20, 1944. It offers regular courses of Economics, Business Administration, Law, Social Sciences and Information technology management. Its original goal was to train people for the country's public- and private-sector management.[…] It is considered by Foreign Policy magazine to be a top-5 "policymaker think-tank" worldwide."

CPDOC

Centro de Pesquisa e Documentação de História Contemporânea do Brasil (*Contemporary Brazilian History Research and Documentation Center*) is a private higher education institution founded in 1973 linked to the FGV.

CPDOC

Originally:

To house personal archives of public figures

Develop historic research using privileged archive

Documentation & Research

CPDOC

Now:

Graduate program on History, Political Science and Cultural Artifacts (since 1974)

School of Social Sciences and History of the FGV

(since 2005)

CPDOC

- **two centers Rio/SP**
- **Research:**
- Núcleo de Pesquisa Social Aplicada
- Centro de Relações Internacionais
- **Archives:**
  - Programa de Arquivos Pessoais (PAP) program of personal archives
  - Programa de História Oral (PHO) program of oral history

FGV School of Applied Math

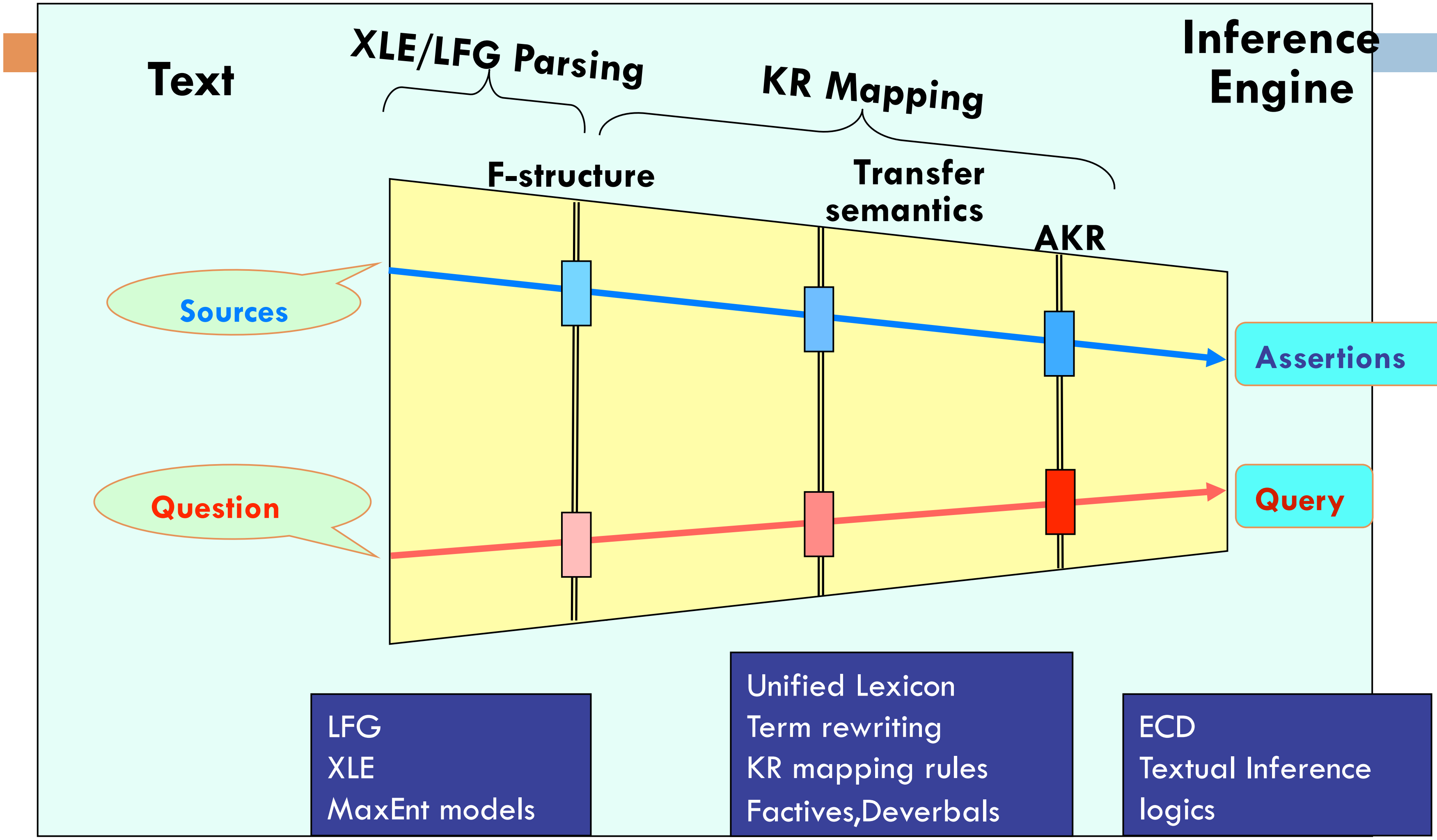
- In the process of being created…
- Experts on image processing, signal/sound processing
- Not much on textual processing
- Possibility of impact…

impa

Portuguese in PARGRAM/PARSEM?

- Alexandre Rademaker, Ph D, PUC-Rio, Mar2010
- Work on Automating Description Logics
- A Bridge-like system for Portuguese?

Bridge Layered Architecture



LFG
XLE
MaxEnt models

Unified Lexicon
Term rewriting
KR mapping rules
Factives,Deverbals

ECD
Textual Inference logics

Basic idea: canonicalization of meanings

Portuguese NLP?

- Lots of Homework to do…
- **STIL** (Simpósio de Tecnologia da Informação e Linguagem Humana) 8th ed
- **PROPOR (bienally, 9th edition)**
- **Linguistica de corpus (7a edition)**
- Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade Maria das Graças V. Nunes, Sandra M. Aluisio, Thiago A. S. Pardo NILC – ICMC – Universidade de São Paulo São Carlos – SP, Brasil Junho de 2010
- Pardo, T.A.S.; Gasperin, C.V.; Caseli, H.M.; Nunes. M.G.V. (2010). Computational Linguistics in Brazil: An Overview. In the *Proceedings of the NAACL-HLT Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pp. 1-7. June 1-6, Los Angeles, CA/USA. Pdf
- **SBC special interest group in NLP, 2007**

Challenges from STIL09 survey

- Lack of large and robust language resources
- Lack of formal models for linguistic description and analysis of Portuguese
- Difficulty in attracting students and researchers
- Lack of multidisciplinary collaboration
- CL/NLP marginalization in both Computer Science and Linguistics.
- Poor interaction between universities and industry
- Insufficient funding.

NILC's view 1

- Recently more robust semantic tools like TeP 2.016, Wordnet.PT17, and MWN.PT18, as well as named entities recognizers, e.g., REMBRANDT19.
- Topics: text summarization, machine translation, text simplification, automatic discourse analysis, coreference and anaphora resolution, information retrieval, text mining, terminology/lexicon research, ontologies and semantic tagging, and corpus linguistics.
- The largest CL/NLP research group in Brazil is NILC (Interinstitutional Center for Research and Development in Computational Linguistics) USP, UFSCar and UNESP

NILC's view 2?

- □ Portuguese has got state of the art tools (as POS taggers and syntactic parsers) and comprehensive corpora of contemporary written language BUT
- □ still needs resources for particular applications
- □ Portuguese syntactic parsers (which are considered basic NLP tools) and wordnet-like resources are still too limited.
- □ CHARLA 2008 ?
- □ contests/conferences such as TAC33, Senseval/ SemEval34, and TREC35, among others, might make Portuguese/Spanish datasets available, as CLEF36 **has done.**

Some detail on CPDOC data

- Brazilian Biographic-Historic Dictionary (Dicionario Historico-Biografico Brasileiro DHBB)
- First edition (printed) 1984:
  - 4.400 entries / 4 volumes.
- Second edition (printed and cd-rom) 2001:
  - 6.626 entries / 5 volumes or cd-rom.
- Third edition (online) 2010:
  - 7.553 entries / online

CPDOC Online Portal

- Since 2009, search tool to access CPDOC's archives
- Documents of personal archives, photos, interviews from the Oral History Project and entries of the DHBB.

- http://www.fgv.br/cpdoc/busca

User Profile

Undergrads	53,68%
Grad studies (lato sensu)	13,18%
Masters	11,50%
High school	11,07%
Doctoral cand	5,72%

User Profile/Goal of research

Artigo acadêmico	17,97%
Curiosidade histórica	14,61%
Monografia (graduação)	11,74%
Dissertação (mestrado)	11,39%
Pesquisa escolar	8,52%
Tese (doutorado)	7,57%

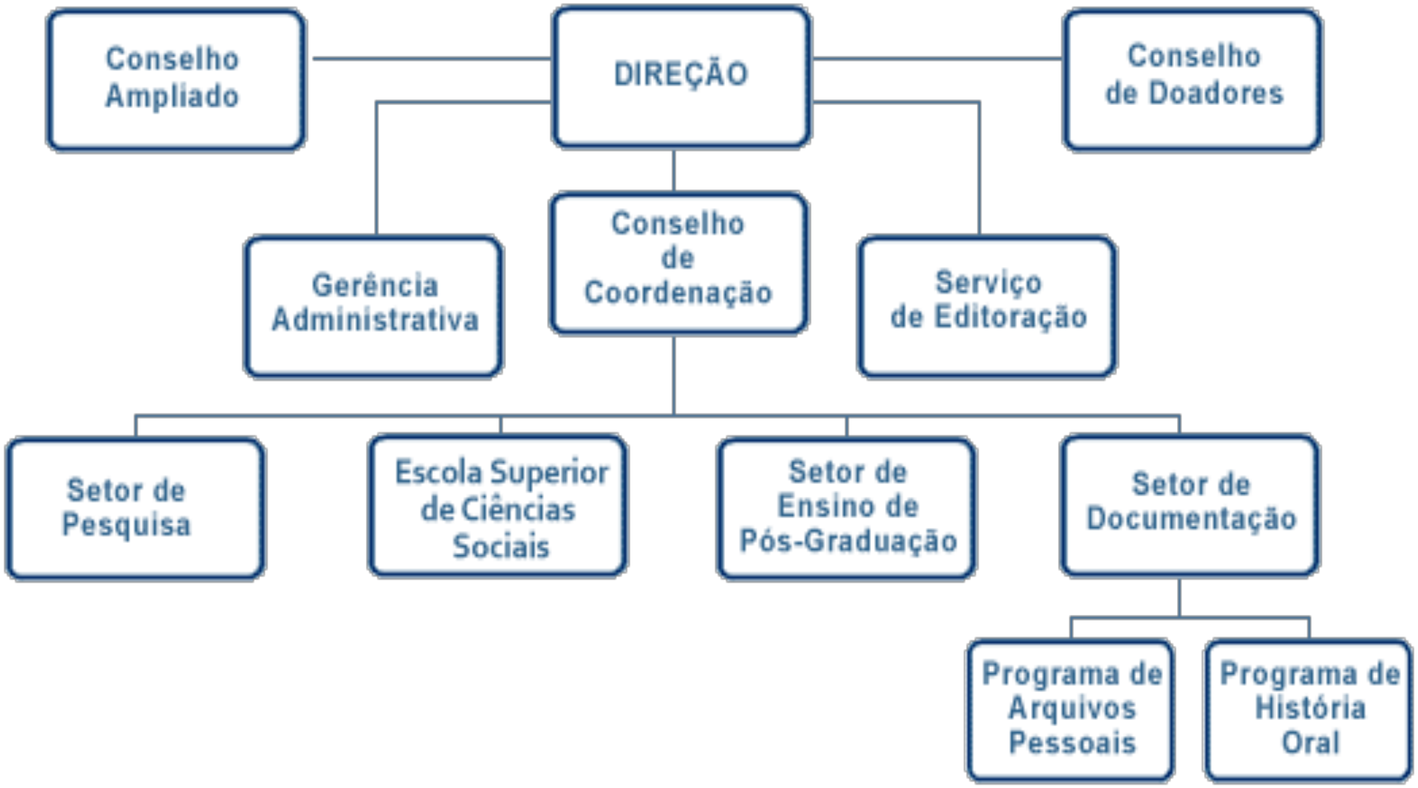
Filme/vídeo (3,13 %) aparecem em penúltimo lugar, e matéria de imprensa (2,29%) em último.

User Profile - media

- Kind of media accessed:

audiovisual	38,21%
textual	33,51%
printed matter	16,33%
all of the above	11,95%

Organizational chart



Thanks!

Revista Estudos Históricos

- Since 1988, half yearly
- Multidisciplinar scientific magazine
- 45 editions so far 2010
- All articles available online
- http://virtualbib.fgv.br/ojs/index.php/reh/index

Thanks!