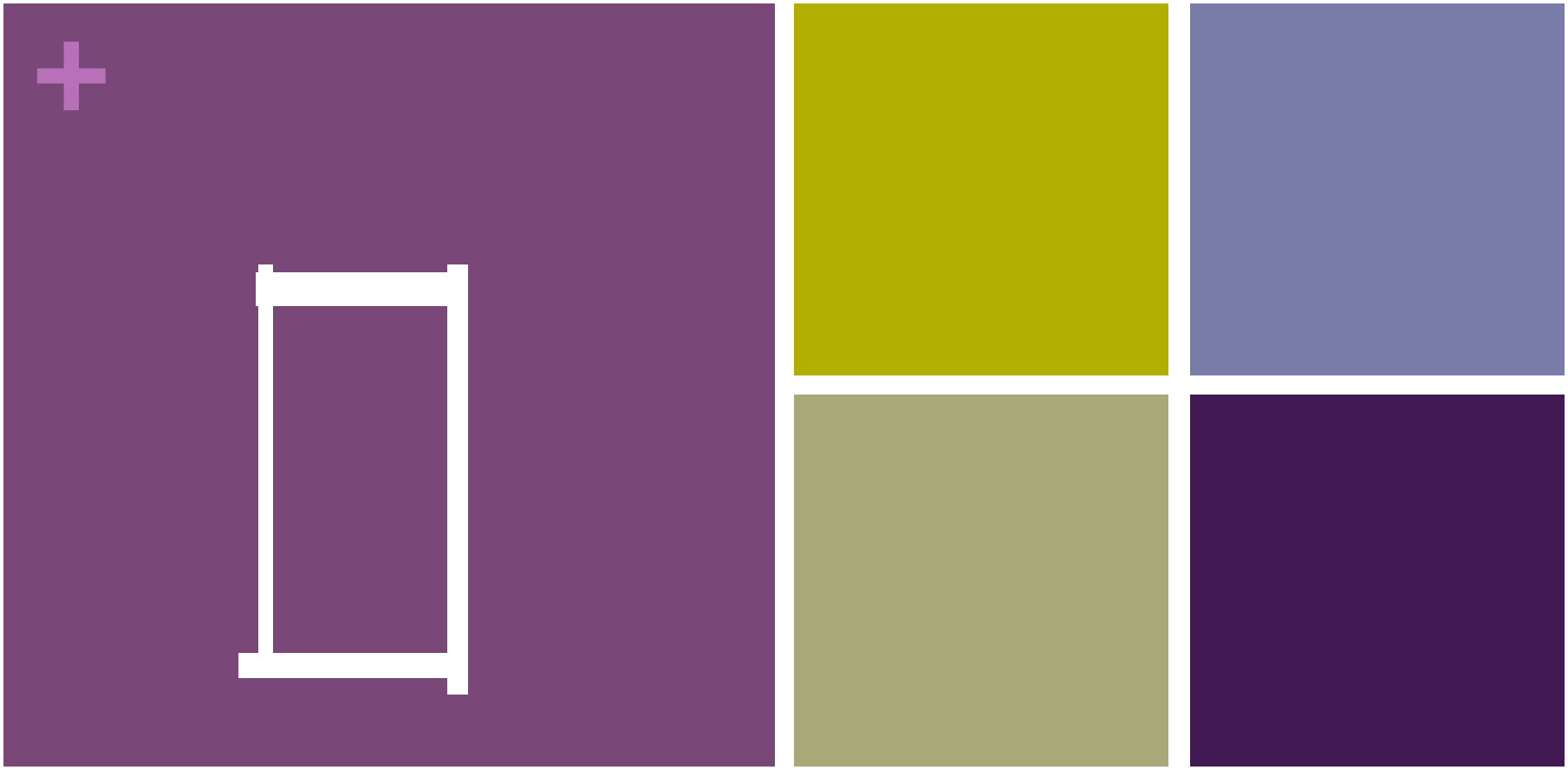


FUNDAÇÃO GETULIO VARGAS
EMAp
Escola de Matemática Aplicada

Connecting OpenWordNet-PT and SUMO

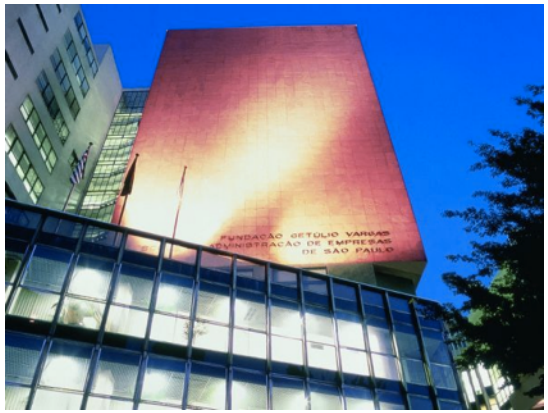
Alexandre Rademaker, EMAp,FGV- Rio
Valeria de Paiva, Rearden Commerce, CA
Gerard de Melo, Berkeley
Rafael Hausler, EMAp/FGV

Global Wordnet Conference 2012
Matsue, Japan

Fundação Getulio Vargas (FGV)

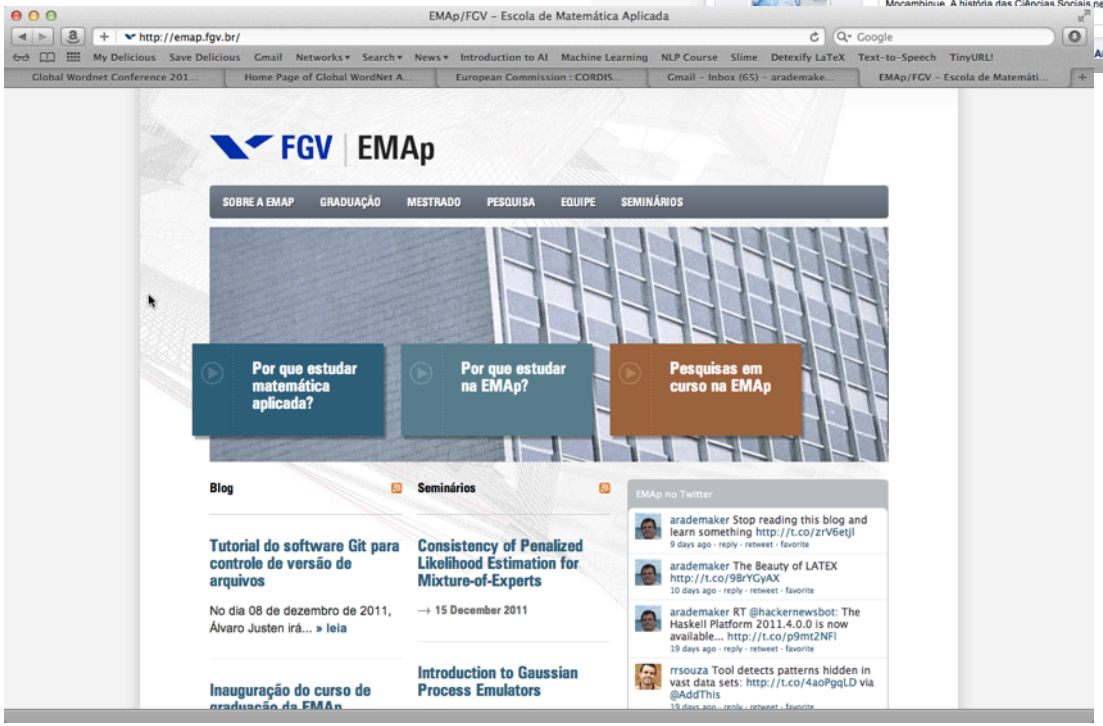
http://www.fgv.br

"**Fundação Getulio Vargas** (*FGV*) is a Brazilian higher education and research institution founded in December 20, 1944. It offers regular courses of Economics, Business Administration, Law, Social Sciences and Applied Mathematics. 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

EMAp



We are starting a project (part of MIST), joint work of CPDOC and EMAp, where we want, in the long run, to use formal logical tools to reason about knowledge obtained from text in Portuguese. We want to improve the structure and search in the CPDOC databases and files.

CPDOC: Center of Brazilian Contemporary History (http://cpdoc.fgv.br)

- CPDOC is a major center for teaching and researching in the Social Sciences and Contemporary History located in Rio de Janeiro.

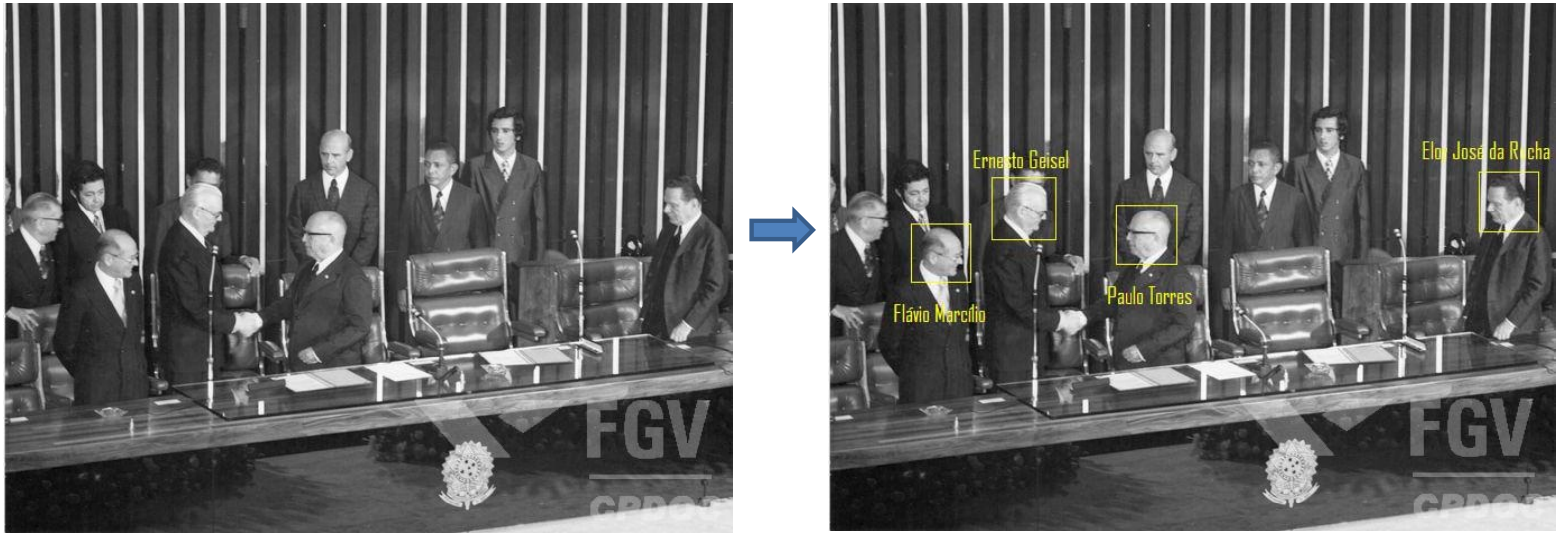
- CPDOC is the leading historical research institute in the country. It holds a major collection of personal archives, oral histories and audiovisual sources pertaining to Brazilian contemporary history.
  - Personal Archives: About 200 archival funds, summing up to 1,8 million documents, among text, images and videos.
  - Oral History Program: A huge set of testimonies (in audio and video) consisting of more than 1.000 interviews, which correspond to up to 5 thousand hours of recordings.
  - Brazilian Historical Biographic Dictionary (DHBB): in the current version, it comprehends 7.553 entries, of which 6.584 are of biographical nature and 969 related to institutions, events and concepts of interest for the Brazilian history after 1930.

EMAp: School of Applied Mathematics (http://emap.fgv.br)

- Created to develop expertise in Mathematics applied to science an technology and help advance FGV's own mission.
- Core team of highly creative and competent mathematicians experts in image and signal/sound processing. Not much in text processing.
- Huge demand for mathematical and computational tools to model the recent social changes in Brazil
- Active partnerships with other schools at FGV and other institutions like Light (power supplier company of RJ) , Petrobras etc.
- Undergraduate and graduate courses (Master)
- Some projects include: Mathematical Epidemiology, Facial Recognition, Modeling the Judiciary, Modeling Legal Conflicts and Natural Language Processing

MIST Project: images
Asla Sá

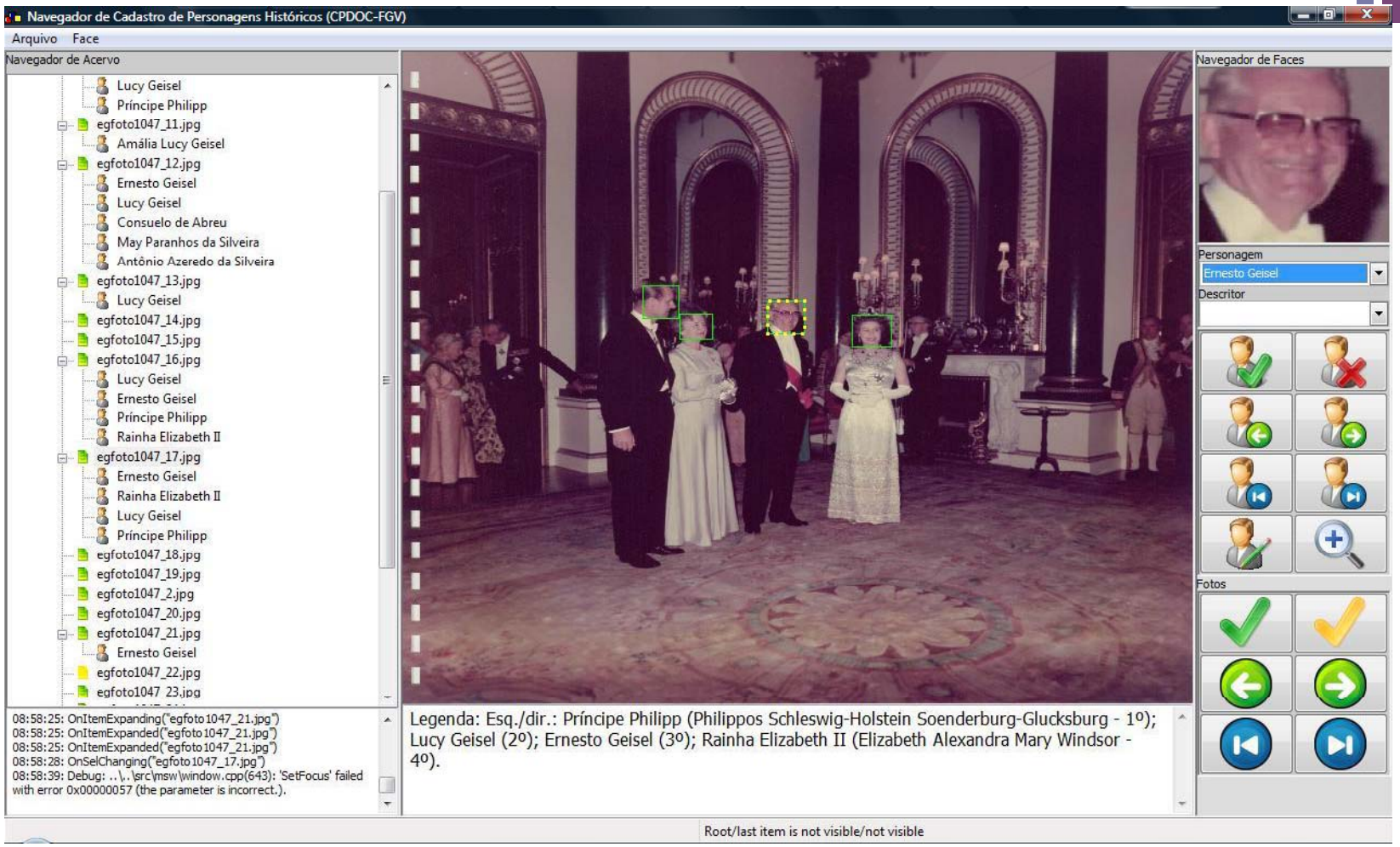
- Original Problem



Legend: Esq./dir.: (1o plano) Flávio Marcílio (1o); Ernesto Geisel (2o); Paulo Torres (3o); Eloy José da Rocha (4o). (2o plano) Adalberto Pereira dos Santos (1o). Foto: Agência Nacional (Estúdio/Agência).

MIST Project: images

Very Important Faces, developed by EMAp team



MIST Project: audio files

Moacyr Silva

MIST Project

Aligning text and sound

ENTREVISTA COM ALEX PERISCINOTO
PROJETO A PROPAGANDA BRASILEIRA – TRAJETÓRIAS E EXPERIÊNCIAS DOS PUBLICITÁRIOS E DAS INSTITUIÇÕES DE PROPAGANDA
São Paulo
Entrevistadoras: Luciana Heymann e Ilana Strozenberg
Transcritor: Oswaldo Moelhmann Cordeiro de Farias
Data da Transcrição: 30.09.2004

Entrevista: 13.07.2004

(00:00:13,6) L.H – Bom, então vamos começar. O seu currículo é bastante sucinto, são dados gerais...

(00:00:21,2) A.P – Ah, mas pode cortar.

(00:00:23,1) L.H – Não, a gente não quer cortar, a gente quer acrescentar. A gente queria começar do começo: quando o senhor nasceu, onde, como era a família…

(00:00:34,7) A.P. – Ah, vem de lá?

(00:00:35,8) L.H – Vem de lá, *from the begining*.

(00:00:39,2) A.P – Família de imigrantes, todos nós acho que somos.

(00:00:45,5) L.H – Imigrantes italianos?

(00:00:46,4) A.P. – Meu pai saiu da Itália, tinha dois navios no inverno na Itália. A Itália é muito fria. Em Veneza, onde ele morava, é mais frio ainda por causa da umidade. E os dois navios que vinham para a América, um era preto e o outro era cinza. O navio preto tinha uma fila

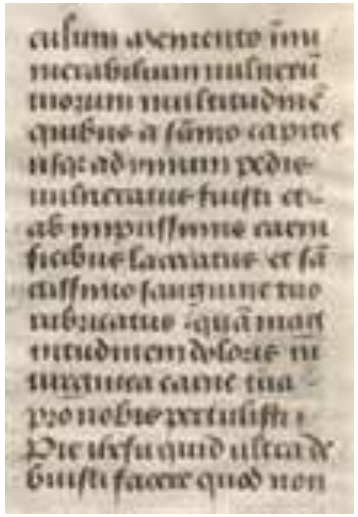
MIST Project: NLP and ontology engineering

Alexandre Rademaker and Renato Rocha

- Conversion of the current authorized subject headings into a history thesaurus: people, processes, events, places etc. These will be afterward converted to domain ontologies and incorporated in the **Semantic Portal**.
- Unify access to the CPDOC Systems; Enhanced visibility to search engines with unification of concepts terminology;
- Integration with the Linked Open Data (LOD) via RDF triplification;
- Integration with the Learning Objects Databases and the FGV Digital Library;
- NLP to extract more relations and knowledge from texts (first DHBB)

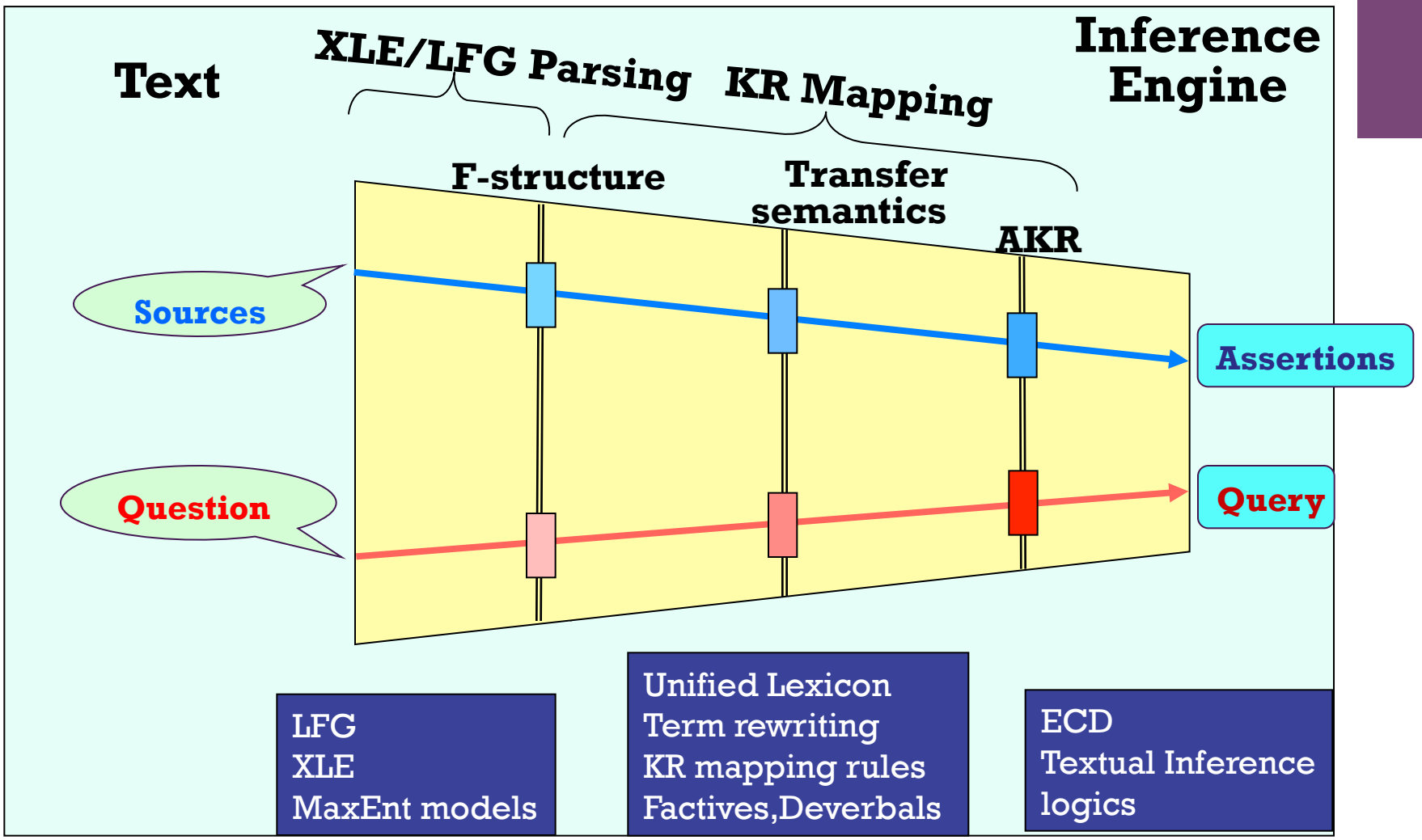
OpenWordnet-PT? (aren't all wordnets open?)

There are some attempts: WordNet.PT and WordNet.PT global (Lisboa), MultiWordNet.PT and Brazilian WordNet by Bento Dias.



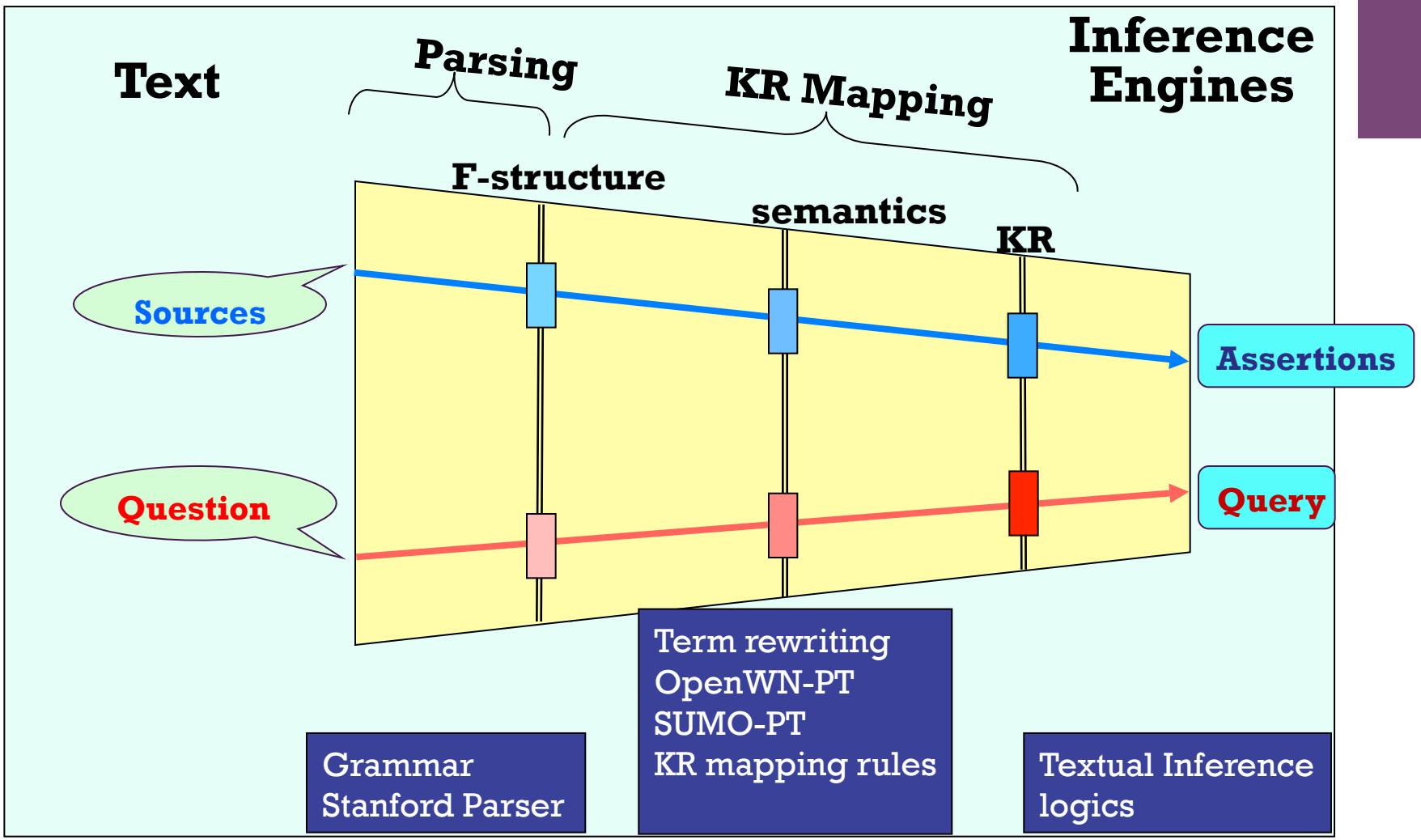
We need a Portuguese Wordnet for our work, but none of the previous projects are openly available.

+ Inspiration: PARC's Bridge Architecture



Basic idea: canonicalization of meanings

Simplifying the PARC's Bridge Architecture



Idea: Simplify and reproduce components in PORTUGUESE

Language/KR (mis?)alignments:

- Language
  - Generalizations come from the structure of the language
  - Representations compositionally derived from sentence structure
- Knowledge representation
  - Generalizations come from the structure of the world
  - Representations to support reasoning
  - Maintain multiple interpretations
- Layered bridge helps with the different constraints
- FIRST STEP of simplified architecture:

  WORDNET for PORTUGUESE

OpenWN-PT: How?

- Leverage EuroWordNet, Global WordNet experiece
- Leverage YAGO, UWN experience…
- **Recruited Gerard de Melo for project**
- **Gerard's work: UWN/MENTA** A large-scale multilingual lexical knowledge base built using statistical methods, transforming WordNet into a massively multilingual resource (over 1 million words and several million named entities in a single large multilingual taxonomy)
- Let us look at Portuguese-projection of UWN/Menta. This is an automated version of a Portuguese WordNet, publicly available.

https://github.com/arademaker/wordnet-br

OpenWN-PT: is it done?

- Universal WordNet (UWN) experience: **Towards a Universal Wordnet by Learning from Combined Evidence (**de Melo, Weikum, *(CIKM 2009)*)

- A methodology for the automatic construction of a large-scale multilingual lexical database where words of many languages are hierarchically organized in terms of their meanings and their semantic relations to other words.

- Bootstrapped from WordNet, extends it with around 1.5 million meaning links for 800,000 words in over 200 languages, drawing on evidence extracted from a variety of resources including existing (monolingual) wordnets, (mostly bilingual) translation dictionaries, and parallel corpora.

  Graph-based scoring functions and statistical learning techniques are used to iteratively integrate this information and build an output graph.

- Experiments show high level of precision and coverage more than 86%. Approx 24K terms in Portuguese

- Is it good enough? Depends on application…

OpenWN-PT: How we started?

The file was generated by combining the following data:

Princeton WordNet 3.0 was used to obtain English glosses and English terms for synset IDs.

The unreleased 2010-12 version UWN and MENTA provided candidate terms in Portuguese, candidate glosses in Portuguese (from Wikipedia), and candidate terms in Spanish.

The EuroWordNet base concept list (5000_bc.xml) provides the base concept numbers. The original file was mapped from WordNet 2.0 to 3.0 using the mappings from WN-Map. When multiple mappings for a WordNet 2.0 synset existed, all possible WordNet 3.0 synsets were kept. Hence, there may be multiple entries with the same base concept number.

http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=57

https://github.com/arademaker/wordnet-br

OpenWN-PT: what does it look like?

- Typical good entry with minor manual improvements.
- Automatic produces candidate Portuguese words for each of some of WN3.0 synsets.
- Check suggested words and add Portuguese gloss and examples.

```
<row>
  <BC>255</BC>
  <WN-3.0-Synset>v233335</WN-3.0-Synset>
  <PT-Words-Man>limitar, restringir</PT-Words-Man>
  <PT-Word-Cand>limitar, restringir</PT-Word-Cand>
  <EN-Gloss>place limits on (extent or access); "restrict the use of
  this parking lot"; "limit the time you can spend with your
  friends"</EN-Gloss>
  <EN-Words>bound, confine, limit, restrain, restrict, throttle,
  trammel</EN-Words>
  <PT-Gloss>estabelecer limites em (alcance ou acesso); "vamos
  restringir o uso desse estacionamento"; "Limite o tempo que voce
  pode gastar com seus amigos"</PT-Gloss>
  <PT-Gloss-Sug />
  <SPA-Words-Sug>limitar</SPA-Words-Sug>
  <Comments />
</row>
```

OpenWN-PT: what does it look like?

```
<row>
  <BC>989</BC>
  <WN-3.0-Synset>v1059564</WN-3.0-Synset>
  <PT-Words-Man />
  <PT-Word-Cand />
  <EN-Gloss>refuse to acknowledge; "She cut him dead at the meeting"</EN-Gloss>
  <EN-Words>cut, disregard, ignore, snub</EN-Words>
  <PT-Gloss />
  <PT-Gloss-Sug />
  <SPA-Words-Sug />
  <Comments />
</row>
```

Not very useful

```
<row>
  <BC>997</BC>
  <WN-3.0-Synset>n1096245</WN-3.0-Synset>
  <PT-Words-Man />
  <PT-Word-Cand>comércio, negócio</PT-Word-Cand>
  <EN-Gloss>the volume of commercial activity; "business is good today"; "show me where the business was today"</EN-Gloss>
  <EN-Words>business</EN-Words>
  <PT-Gloss />
  <PT-Gloss-Sug />
  <SPA-Words-Sug>comercio</SPA-Words-Sug>
  <Comments />
</row>
```

Good automatically suggestion

OpenWN-PT: lexical gaps

```
<row>
  <BC>9</BC>
  <WN-3.0-Synset>v9147</WN-3.0-Synset>
  <PT-Words-Man lexicalGap="True">mudar, depenar, desplumar, trocar</PT-Words-Man>
  <PT-Words-Man></PT-Words-Man>
  <PT-Words-Cand />
  <EN-Gloss>cast off hair, skin, horn, or feathers; "our dog sheds every Spring"</EN-Gloss>
  <EN-Words>exuviate, molt, moult, shed, slough</EN-Words>
  <PT-Gloss>arrematar o cabelo, pele, chifre ou penas; "nosso
  cachorro muda de pelo toda primavera"</PT-Gloss>
  <PT-Gloss-Sug />
  <Spa-Words-Sug>mudar, pelechar</Spa-Words-Sug>
  <Comments by="rafael">esse synset eu não sei se intendi corretamente. Intendi
  que o Gloss diz respeito a mudança de característica física (tanto
  animal quanto de homen), estou em duvida quanto aos PT-Words-Man,
  por favor corrija se estiver errado.</Comments>
  <Comments by="vcvp">Rafael esse synset não é necessário, ele não
  existe em português, eu acho. Em ingles eles teem verbos pra mudar
  de pele/pelo.</Comments>
</row>
```

```
<row>
  <BC>10</BC>
  <WN-3.0-Synset>v10054</WN-3.0-Synset>
  <PT-Words-Man lexicalGap="True">mover_reflexivamente, mover_involuntariamente</PT-Words-Man>
  <PT-Words-Man></PT-Words-Man>
  <PT-Words-Cand />
  <EN-Gloss>move in an uncontrolled manner</EN-Gloss>
  <EN-Words>move involuntarily, move reflexively</EN-Words>
  <PT-Gloss>mover de forma incontrolada</PT-Gloss>
  <PT-Gloss-Sug />
  <Spa-Words-Sug />
  <Comments by="vcvp">esse tb nao sei porque existe</Comments>
</row>
```

OpenWN-PT: revisions

```
<row>
  <BC>12</BC>
  <WN-3.0-Synset>v10435</WN-3.0-Synset>
  <PT-Words-Man >agir, atuar, comportar-se, parecer, fazer_papel</PT-Words-Man>
  <PT-Words-Cand>actuar, atuar</PT-Words-Cand>
  <EN-Gloss>behave in a certain manner; show a certain behavior;
  conduct or comport oneself; "You should act like an adult"; "Don't
  behave like a fool"; "What makes her do this way?"; "The dog acts
  ferocious, but he is really afraid of people"</EN-Gloss>
  <EN-Words>act, behave, do</EN-Words>
  <PT-Gloss >agir de uma certa forma; comportar-se ou conduzir-se;
  "você deveria agir como adulto"; "não se comporte como um tolo";
  "o que faz ela agir dessa maneira?"; "o cachorro parece feroz, mas
  ele realmente tem medo de pessoas"</PT-Gloss>
  <PT-Gloss-Sug />
  <Spa-Words-Sug>construir</Spa-Words-Sug>
  <Comments by="vcvp">construir? nao faz sentido</Comments>
</row>
```

```
<row>
  <BC>14</BC>
  <WN-3.0-Synset>v12613</WN-3.0-Synset>
  <PT-Words-Man> paralizar</PT-Words-Man>
  <PT-Words-Cand />
  <EN-Gloss>suddenly behave coldly and formally; "She froze when she
  saw her ex-husband"</EN-Gloss>
  <EN-Words>freeze</EN-Words>
  <PT-Gloss>agir repentinamente de forma fria e formal; "ela
  paralizou quando viu seu ex-marido"</PT-Gloss>
  <PT-Gloss-Sug />
  <Spa-Words-Sug />
  <Comments by="vcvp">acho que esse esta' estranho, dizemos "o seu
  sorriso congelou" mas paralizou na frase do gloss é mais ficar
  sem ação, do que realmente agir de uma maneira fria. enfim estou
  mantendo.</Comments>
</row>
```

We are not using linguistic experts, revision is always necessary!

OpenWN-PT: first step guidelines

- Read the English gloss and the English words.
- Come up with Portuguese words that express the same meaning as the English gloss and have the part-of-speech indicated by the first letter of the WordNet synset identifer and write them into "PT-Words-Man".
- Write a Portuguese gloss into the "PT-Gloss" field. If the gloss contains English example sentences, then only translate them if their translations sound natural in Portuguese and if the translation actually contains the Portuguese words added to the synset.

Done? Not so simple…

- Checking is much easier than starting from scratch..
- But long and tedious work to check even the initial 5k synsets suggested by GWA let alone the 24k synsets already in UWN
- Necessary? YES! Lexical gaps of all sorts
- Evolving guidelines for translators/checkers
- Assumed we'd be done on 5K for this talk, but still working.
- Payoff expected: A huge body of work on data, hopefully reproducible in Portuguese

OpenWN-PT: next step

- Keep following the procedure described as the "expand approach" for the global wordnet grid.
- First translate the synsets in the Princeton WordNet to Portuguese, then take over the relations from Princeton and revise, adding the Portuguese terms that satisfy different relations. Then revise and revise and revise until we can guarantee the consistency of the taxonomy.
- Since we have a first target corpus, the Brazilian Historical Biographic Dictionary, we can also calculate word frequency to prioritize expansion of the OpenWN-PT.

Conclusions

- Took to heart GWA's claim: need OPEN Portuguese WordNet, starting with 5k concepts suggested.
- Have automatically-constructed version obtained from Universal WordNet UWN/ Menta
- We're not where we wanted to be, but things are progressing solidly. Many issues on working at a distance. We had hoped to have 5k synsets done by now. 812 synsets is a good start, considering the Zipfian distribution of WF. Each synset has multiple words, and Francis and Kucera showed that with 1000 words, you can already understand 72% of written text.
- Of the 300 synsets that were **double** inspected/corrected by hand, Gerard methods really seem to be living up to expectations. The data is language, so it's messy, noisy and subject to interpretation, but mostly it seems good quality.
- Need to increase number of people doing it, need to create more checks. We want to experiment crowd sourcing, like http://tagger.thepcf.org.uk/, or game oriented, http://freerice.com/. To volunteers workers, maybe motivated by status upgrade like http://stackoverflow.com/. Try the Asian Wordnet Management System.

Thanks!

References

Towards a Universal Wordnet by Learning from Combined Evidence Gerard de Melo, Gerhard Weikum (2009) *18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, Hong Kong, China.

Bridges from Language to Logic: Concepts, Contexts and Ontologies Valeria de Paiva (2010) Logical and Semantic Frameworks with Applications, LSFA'10, Natal, Brazil, 2010.

`A Basic Logic for Textual inference", AAAI Workshop on Inference for Textual Question Answering, 2005.

``Textual Inference Logic: Take Two", CONTEXT 2007.

``Precision-focused Textual Inference", Workshop on Textual Entailment and Paraphrasing, 2007.

PARC's Bridge and Question Answering System Proceedings of Grammar Engineering Across Frameworks, 2007.