## Relação entre informática e linguística, ou … entre informáticos e linguistas

Diana Santos

Linguateca
www.linguateca.pt



## Interdisciplinaridade

- Uns vêm de Letras
- Outros de informática
- Mas o caminho a seguir é o mesmo, na área do processamento computacional da língua
- Não interessa onde começaram
- Professores dos 2 lados

Primeira Escola de Verão da Linguateca



## Onde estamos?

- No princípio!
- Na primeira Escola de Verão
- Lançando as primeiras sementes para fora
- Começando a colher os primeiros frutos

- Alguma visibilidade a nível internacional
- Uso generalizado dos nossos recursos
- Mas ainda não chegámos ao cidadão comum

Primeira Escola de Verão da Linguateca



## Outras visões

- Dias da Silva, 1998: "Bridging the gap between linguistic theory and natural language processing"
  - Mining
  - Molding
  - Assembling

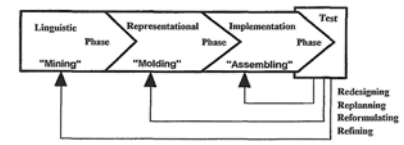


Fig. 2 The three-domain approach to NLP.



## Outras visões: Robert France

- Robert B. France "Model-driven Development of Complex Software", presentation at SINTEF, May 28, 2009
- Em engenharia informática, duas escolas para
  - *Tackling the abstraction challenge:*
  1. *The extensible general-purpose school*
  2. *The domain-specific language school*
  - Cada uma com os seus problemas!
  1. *Difficulty of identifying a small base set of modeling concepts that can be used to express a broad range of abstractions*
  2. *Creating and evolving languages and their toolsets*
- Isto lembra-vos alguma coisa?



## De uma chamada recente ☺ na corpora-list

- For many decades, **NLP has suffered from low software engineering standards** causing a limited degree of re-usability of code and interoperability of different modules within larger NLP systems. While this did not really hamper success in limited task areas (such as implementing a parser), it caused serious problems for the emerging field of language technology where the focus is on **building complex integrated software systems**, e.g., for information extraction or machine translation. This lack of integration has led to duplicated software development, work-arounds for programs written in different (versions of) programming languages, and ad-hoc tweaking of interfaces between modules developed at different sites.

## Outras visões (Maia et al., 2008)

- *Cooperation should not be difficult if the linguist has good theory to test – and the engineer can devise some way to test it*
- *What is needed is good will and serious attempts by both sides to understand each other's point of view*



## Outras visões: Robert Binnick

- Porque é que há uma tendência para não se ler / reconhecer / saber o que já foi escrito sobre o assunto?
- Em IA nos anos 90, em MDA nos anos 2000 ...
- Em PLN nos anos 60, em SW nos anos 2000 ...
- Em filosofia no séc. V a.C, em IA nos anos 60 ...
- Em filosofia nos anos 50, em política nos anos 2000...



## A diferença é entre ciência e engenharia...

- O que pode correr mal neste algoritmo?
- *Pegar em 86 pares de páginas da rede candidatas a serem textos paralelos, pôr pessoas a detectar se sim se não, criar uma colecção dourada para depois avaliar*

(Resnik & Smith, 2003)

- O que está mal neste sistema?
- *Um sistema que em 90% dos casos dá uma resposta certa*

(Artstein & Poesio, 2008)



## A diferença é entre ciência e engenharia...

- Se se comparar as entradas de um dicionário língua 1 -> inglês e língua 2 -> inglês para obter um dicionário l1->l2 ou l2-11, o que pode correr mal?

(Sjöbergh, 2005)

- Se se comparar dois documentos com base nos nomes próprios que partilham, o que pode correr mal?

(Friburg & Maurel, 2002)

- Se se tentar analisar páginas da Wikipédia através das categorias a que pertencem, o que pode correr mal?

(Cardoso, 2009, em conversa)



## Categorias a mais?

- http://en.wikipedia.org/wiki/**Empire_State_Building**
- Categories: 1931 architecture | **Accidents involving fog** | Art Deco buildings in New York City | Fifth Avenue (Manhattan) | Former world's tallest buildings |National Historic Landmarks in New York City | Office buildings in New York City | National Register of Historic Places in Manhattan | Skyscrapers in New York City | Skyscrapers over 350 meters | Visitor attractions in New York City | Tallest buildings in their city
- http://en.wikipedia.org/wiki/**Rembrandt**
- Categories: Rembrandt | Dutch painters | Dutch engravers | Dutch Golden Age painters | Baroque painters | Portrait artists | People from Leiden | People from Amsterdam | **Leiden University** | 1669 deaths | 1606 births



## Referências


- Artstein, Ron & Massimo Poesio. "Inter-Coder Agreement for Computational Linguistics". *Computational Linguistics* **34**, 4, 2008, pp. 555-596.
- Binnick, R.J. *Time and the Verb*, Basil Blackwell, 1991.
- Dias-da-Silva, Bento Carlos. "Bridging the gap between linguistic theory and natural language processing". In *Proceedings of the 16th International Congress of Linguists*, *ICL 16*, 1997, Paris. Oxford: Elsevier-Pergamon, 1998.
- France, Robert B. "Model-driven Development of Complex Software", apresentação no SINTEF, Oslo, 28 de Maio de 2009

## Referências (cont.)


- Friburger, N. & D. Maurel. “Textual Similarity Based on Proper Names”, in *Proceeedings of Mathematical/Formal Methods in Information Retrieval, SIGIR 2002* (Tampere, 15 August 2002), pp. 155-167.
- Maia, Belinda, Rui Sousa Silva, Anabela Barreiro & Cecília Fróis. "N-grams in search of theories". In Barbara Lewandowska-Tomaszczyk (ed.) *Corpus Linguistics, Computer Tools, and Applications – State-of-the Art.* Peter Lang, 2008, pp. 71-84.
- Resnik, Philip & Noah A. Smith. "The Web as a parallel corpus". *Computational Linguistics* **29**, 3, September 2003, pp. 349-380.
- Sjöbergh, Jonas. "Creating a free digital Japanese-Swedish dictionary", *PACLING 2005*.