Introdução ao Processamento de Línguas Naturais

SCC5908 Introdução ao Processamento de Língua Natural

Thiago A. S. Pardo

1

Dilemas no Brasil

- Como lidar com a interdisciplinaridade
  - Linda no papel, complicada na prática
    - Carta de Búzios
      - Linguística é área afim da Computação?
- Qualis
  - Relativamente confortável para a Linguística (será?)
  - Árduo para a Computação

2

Dilemas no Brasil

- Como atrair áreas correlatas? Na contramão do que se exige em Computação?
  - Ciência da Informação
- Processamos o português e publicamos em inglês para estrangeiros?
  - Aceitação nem sempre fácil em conferências internacionais
  - Valorização do trabalho com o português

3

Dilemas no Brasil

- Dilema do PROPOR
  - Inglês
    - Língua franca da ciência
    - Internacionalização da pesquisa
  - Mas qual o limite de internacionalização de um evento chamado *International Conference on Computational Processing of Portuguese*

4

Dilemas no Brasil

- Texto vs. fala
  - Comunidades separadas, mas tentando conversar
    - Texto: cientistas da computação, linguistas
    - Fala: engenheiros elétricos

5

Tendências no mundo

- Aplicações *cross-language*
  - Apesar de limitações de PLN
- Robustez, escalabilidade e independência de língua
  - “Deve funcionar para qualquer coisa retornada pelo Google”

6

Tendências no mundo

- E-mails e mensagens instantâneas
- Blogs e microblogs
- Redes sociais
- Análise de opiniões
  - *Sentiment analysis*

7

Tendências no mundo

- Atenção aos minoritários
  - Desafio científico & (ou versus?) trabalho social
- Conferências de avaliação conjunta
  - NIST, TREC, MUC, DUC/TAC, CLEF, HAREM, PÁGICO, etc.
  - *Roadmaps*

8

PLN: onde encontrar

- De âmbito internacional
  - ACL, NAACL, EACL, HLT, COLING, EMNLP, Interspeech, PROPOR, CICLING, CoNLL, EAMT, IJCNLP, LAW, LREC, RANLP, Corpus Linguistics, ...
  - *Computational Linguistics*, *Transactions of the Association for Computational Linguistics*, *Natural Language Engineering*, *Machine Translation*, Linguamática, ...
- De âmbito nacional
  - STIL, JDP, ELC, ...

9

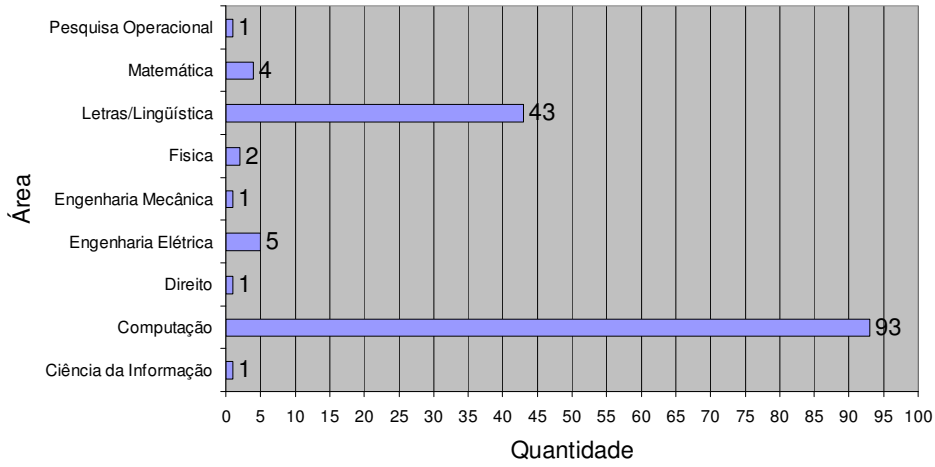
PLN no Brasil

- Como sentem?
  - Vai bem?
  - Principais áreas de pesquisa?

10

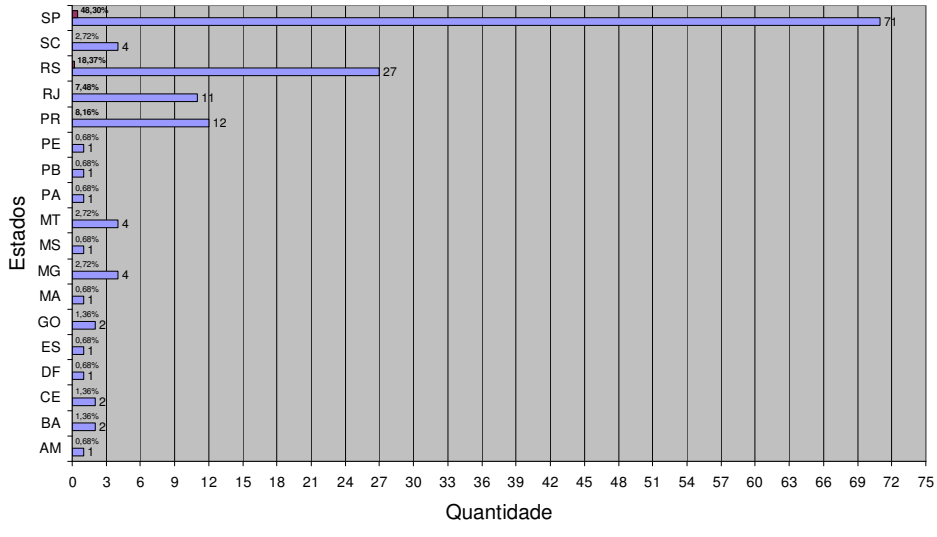
PLN no Brasil Pardo et al. (2009)

Área de formação



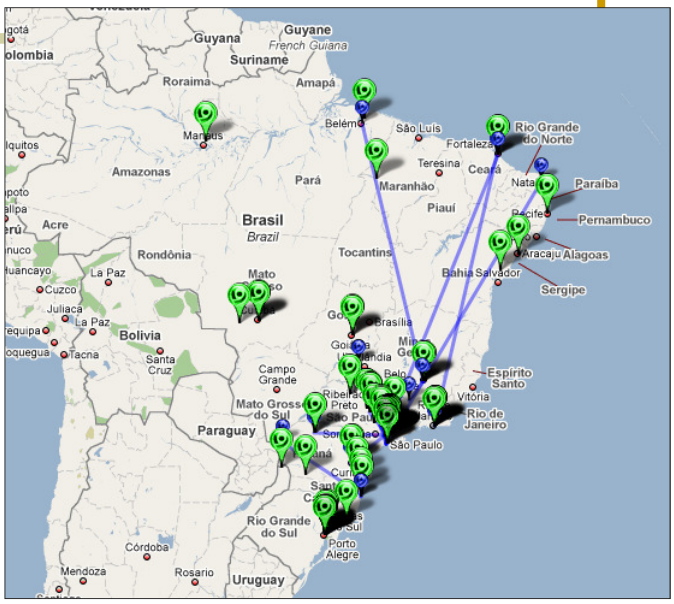
PLN no Brasil Pardo et al. (2009)

Distribuição de pesquisadores por estado



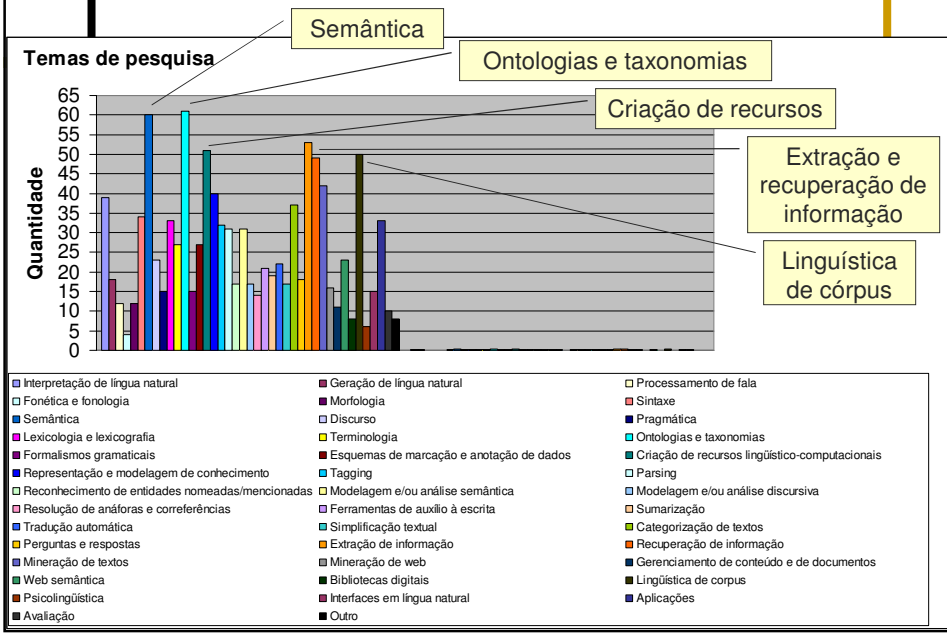
PLN no Brasil

Pardo et al. (2009)



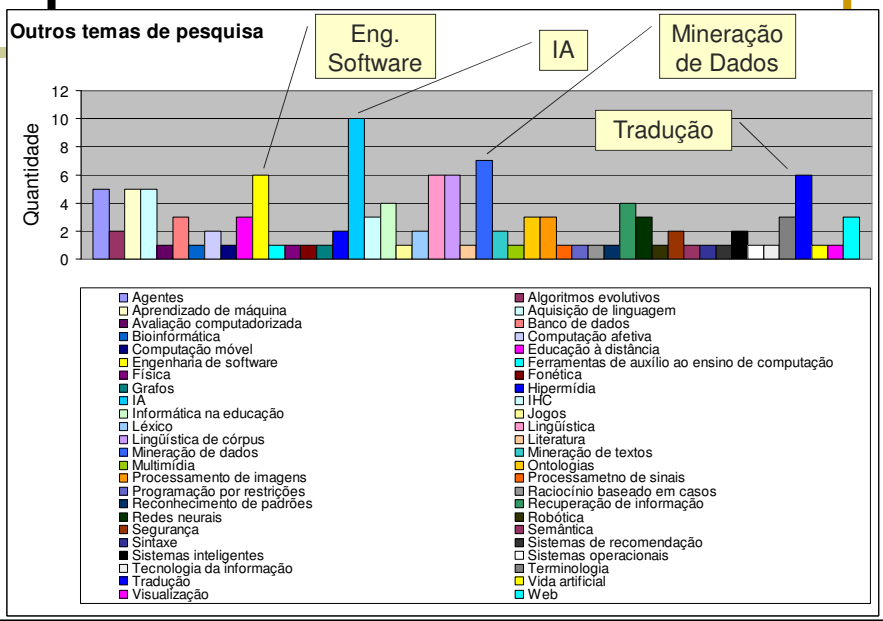
PLN no Brasil

Pardo et al. (2009)



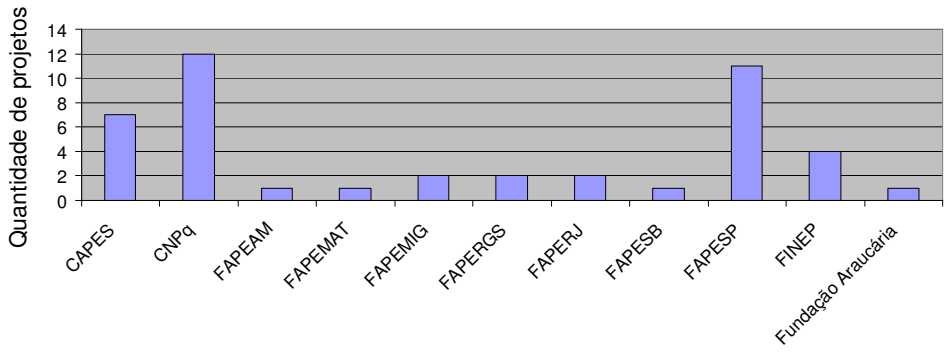
PLN no Brasil Pardo et al. (2009)

Outros temas de pesquisa



PLN no Brasil Pardo et al. (2009)

Fontes de financiamento



16

PLN no Brasil Pardo et al. (2009)

Desafios refinados	%	Nro.
Financiamento de projetos	14,2%	19
Ausência de recursos básicos de qualidade para o português (córpus, um bom parser, WN, REM)	11,9%	16
Dificuldade em atrair e formar alunos e pesquisadores	6,7%	9
Criação e refinamento de modelos de descrição e análise lingüística	5,2%	7
Montagem e coordenação de esforços multidisciplinares	4,5%	6
Pouca interação entre universidade e empresa nessa área de pesquisa	4,5%	6
Criação de ontologias	3,7%	5
Escassez no país de material de pesquisa relevante (por exemplo, livros de autores renomados da área)	3,7%	5
Interação multidisciplinar	3,7%	5
Anotação de córpus	3,0%	4
Certa marginalização da área tanto na Computação quanto na Lingüística	3,0%	4
Falta de formação computacional básica para lingüistas	3,0%	4
Metodologia de avaliação robusta de recursos, ferramentas e aplicações	2,2%	3
Realizar pesquisa em conjunto com as demais atividades que as universidades demandam	2,2%	3
Divulgação da área e das ferramentas criadas	2,2%	3
Sistematização e automatização das práticas da lexicografia e terminologia	1,5%	2
Resultados insatisfatórios na extração automática de termos	1,5%	2
Maior e melhor interface e interatividade dos sistemas de PLN	1,5%	2
Acesso a bases de dados nacionais e internacionais	1,5%	2
Produção de material de pesquisa em português	1,5%	2
Falta de cooperação entre grupos nacionais	1,5%	2

PLN no Brasil Pardo et al. (2009)

Pouca integração entre os grupos de pesquisa nacionais e internacionais	0,7%	1
Desenvolvimento de sistemas para aplicações reais e de alto desempenho	0,7%	1
Falta de ações da SBC para favorecer pesquisas multidisciplinares	0,7%	1
Pulverização da pesquisa em subáreas distintas	0,7%	1
Trabalhar com língua portuguesa e ter inserção internacional	0,7%	1
Falta de modelos de processamento integrado dos vários níveis de conhecimento lingüístico	0,7%	1
Desequilíbrio na distribuição de financiamento (grupos estabelecidos conseguem mais)	0,7%	1
Criação de um glossário eletrônico	0,7%	1
Lacunas lexicais, culturais e pragmáticas entre inglês e português	0,7%	1
Editor que permita armazenar e manipular os resultados de pesquisas lingüísticas	0,7%	1
Busca de padrões em textos criptografados	0,7%	1
Alinhamento semântico entre línguas naturais	0,7%	1
Resultados insatisfatórios em extração de informação	0,7%	1
Incorporar conhecimento da Lingüística Computacional para construção da web semântica	0,7%	1
Direitos autorais para construção de córpus	0,7%	1
Equipamento computacional ultrapassado	0,7%	1
Poucas pesquisas em Geração de Língua Natural	0,7%	1
Resultados insatisfatórios em recuperação de informação	0,7%	1
Criação de recursos que permitam avanços nas pesquisas em tradução automática	0,7%	1
Poucos avanços recentes na área de tradução automática	0,7%	1
Desenvolvimento de técnicas para anotação automática de dados	0,7%	1
Desenvolvimento de sistemas sem a necessidade de dados anotados	0,7%	1
Pouco desenvolvimento da área de pesquisa	0,7%	1

PLN no Brasil Pardo et al. (2009)

- PLN & IA (até 2008)

	PLN	IA	Proporção
Artigos em periódicos	809	1307	0,62
Livros	110	179	0,61
Capítulos de livros	264	473	0,56
Trabalhos em anais	1603	6264	0,26
Resumos expandidos em anais	197	506	0,39
Resumos em anais	975	1695	0,58
Doutorados finalizados	102	225	0,45
Mestrados finalizados	455	1267	0,36
ICs finalizadas	418	983	0,43
Doutorados em andamento	45	143	0,31
Mestrados em andamento	184	335	0,55
ICs em andamento	42	220	0,19