Introdução ao Processamento de Línguas Naturais

SCC5869 Tópicos em Processamento de Língua Natural

Thiago A. S. Pardo

1

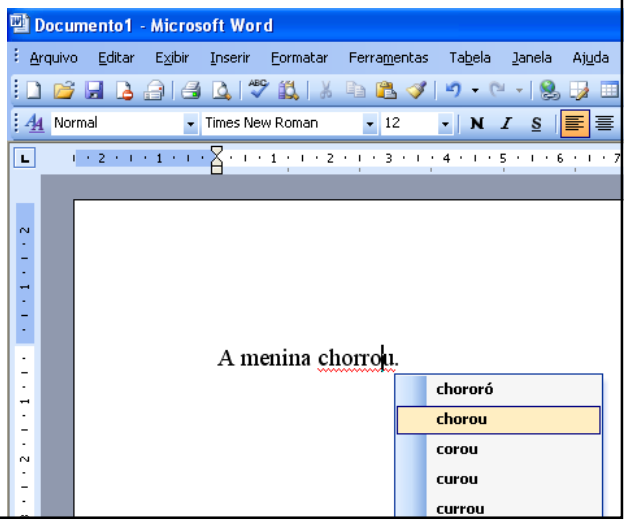
PLN e áreas correlatas

- Limites entre PLN e outras áreas: como percebem isso?
  - Recuperação de informação
  - Extração de informação
  - Inteligência artificial
  - Banco de dados
  - Interação humano-computador
  - Tradução automática
  - Tradução
  - Mineração de textos
  - Lingüística de córpus

2

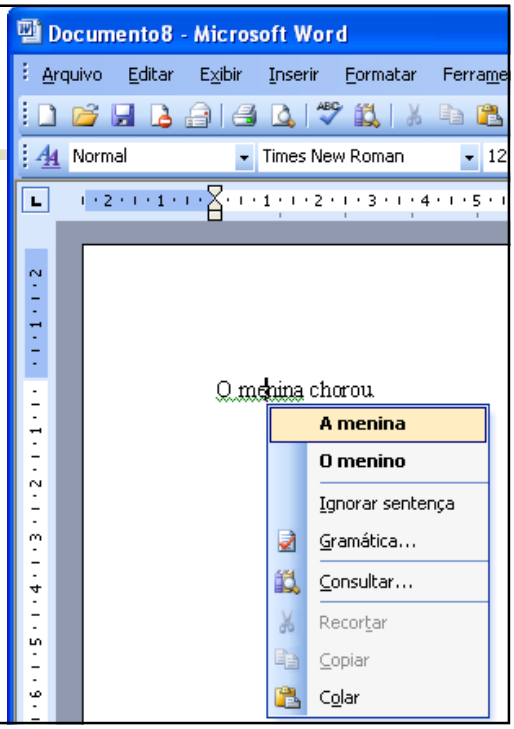
Exemplos

- Revisão ortográfica
  - Tokenizador
  - Léxico
  - Regras para ordenar sugestões



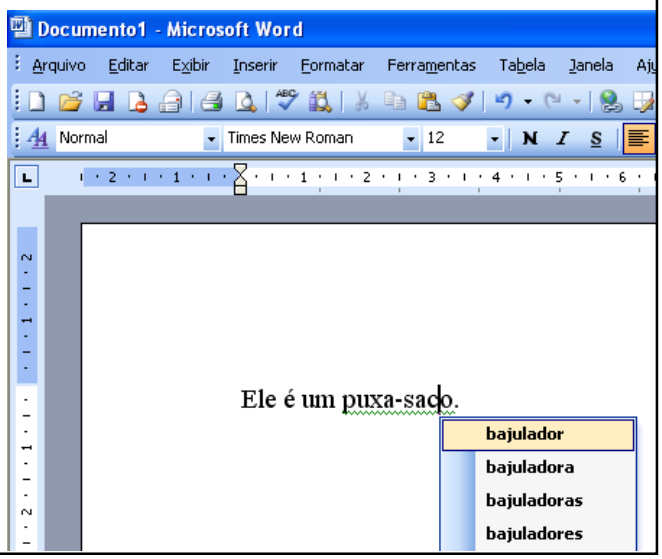
Exemplos

- Revisão gramatical
  - Tokenizador
  - Segmentador sentencial
  - Etiquetador morfossintático
  - Analisador sintático
  - Léxico
  - Regras gramaticais



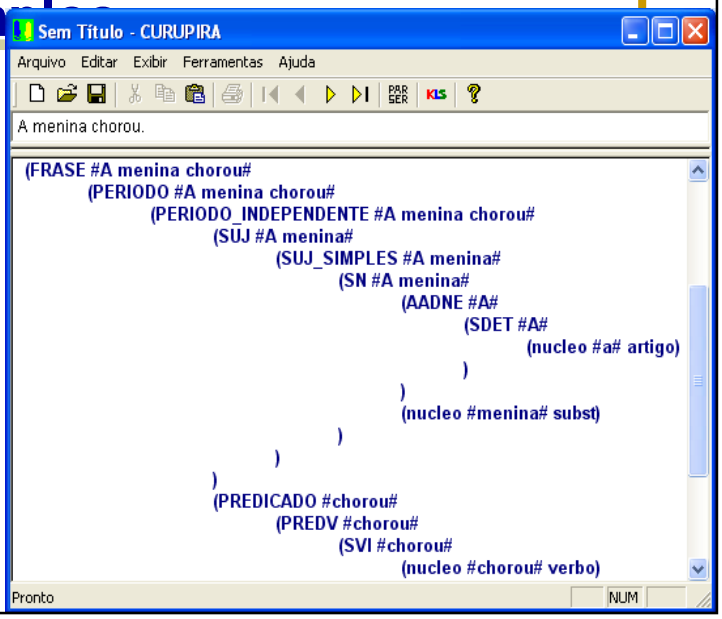
Exemplos

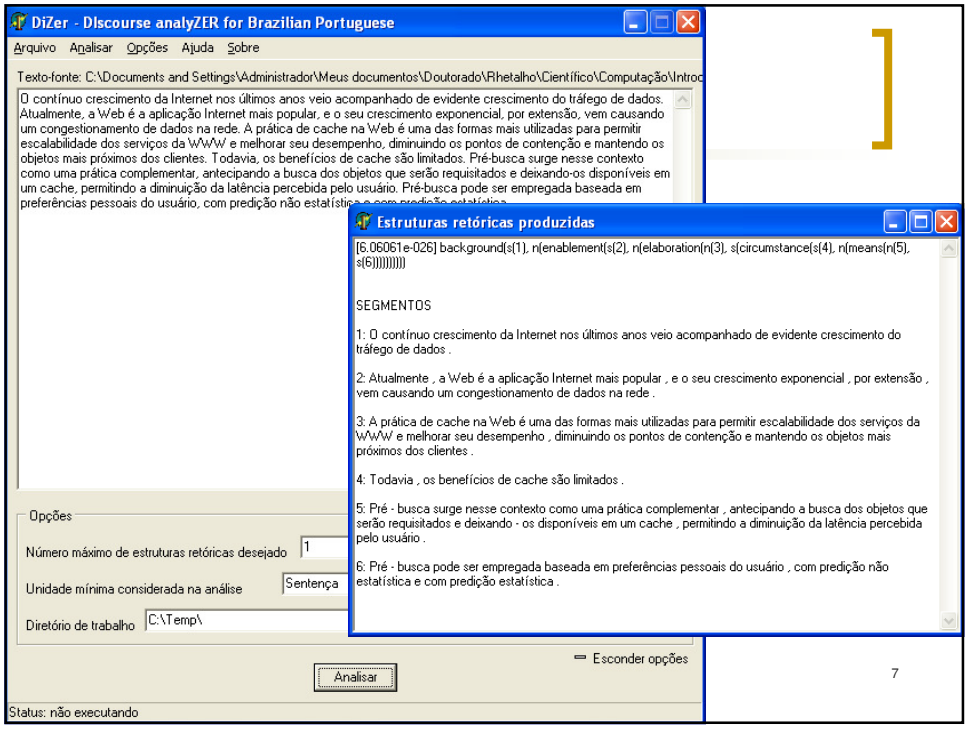
- Revisão estilística
  - Tokenizador
  - Regras estilísticas
  - ...

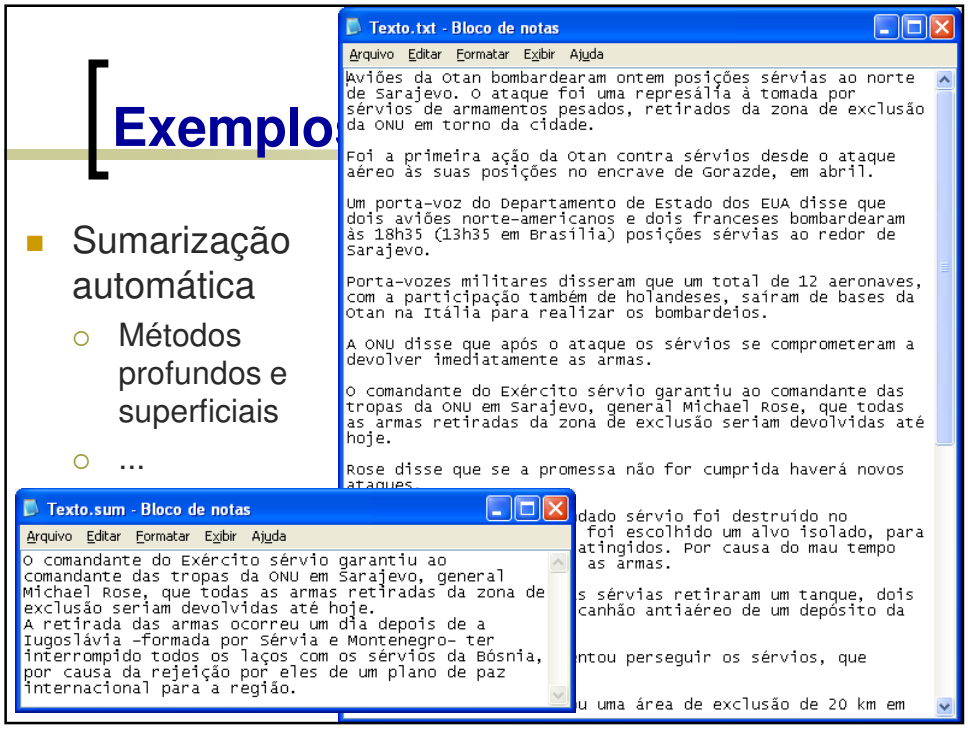


Exemplos

- Análise sintática
  - Léxico
  - Regras sintáticas
  - ...



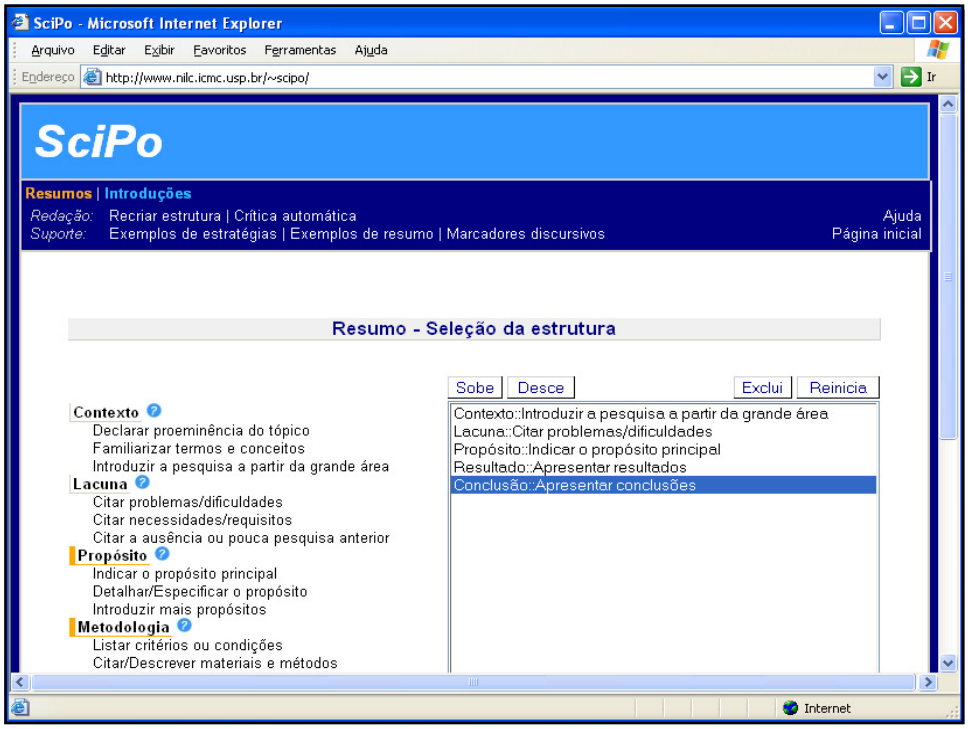

DiZer - DIscourse analyZER for Brazilian Portuguese
Arquivo Analisar Opções Ajuda Sobre
Texto-fonte: C:\Documents and Settings\Administrador\Meus documentos\Doutorado\Rhetalho\Científico\Computação\Introc
O contínuo crescimento da Internet nos últimos anos veio acompanhado de evidente crescimento do tráfego de dados. Atualmente, a Web é a aplicação Internet mais popular, e o seu crescimento exponencial, por extensão, vem causando um congestionamento de dados na rede. A prática de cache na Web é uma das formas mais utilizadas para permitir escalabilidade dos serviços da WWW e melhorar seu desempenho, diminuindo os pontos de contenção e mantendo os objetos mais próximos dos clientes. Todavia, os benefícios de cache são limitados. Pré-busca surge nesse contexto como uma prática complementar, antecipando a busca dos objetos que serão requisitados e deixando-os disponíveis em um cache, permitindo a diminuição da latência percebida pelo usuário. Pré-busca pode ser empregada baseada em preferências pessoais do usuário, com predição não estatística
Opções
Número máximo de estruturas retóricas desejado 1
Unidade mínima considerada na análise Sentença
Diretório de trabalho C:\Temp\
Esconder opções
Analisar
Status: não executando
Estruturas retóricas produzidas
[6.06061e-026] background(s(1), n(enablement(s(2), n(elaboration(n(3), s(circumstance(s(4), n(means(n(5), s(6)))))))))
SEGMENTOS
1: O contínuo crescimento da Internet nos últimos anos veio acompanhado de evidente crescimento do tráfego de dados .
2: Atualmente , a Web é a aplicação Internet mais popular , e o seu crescimento exponencial , por extensão , vem causando um congestionamento de dados na rede .
3: A prática de cache na Web é uma das formas mais utilizadas para permitir escalabilidade dos serviços da WWW e melhorar seu desempenho , diminuindo os pontos de contenção e mantendo os objetos mais próximos dos clientes .
4: Todavia , os benefícios de cache são limitados .
5: Pré - busca surge nesse contexto como uma prática complementar , antecipando a busca dos objetos que serão requisitados e deixando - os disponíveis em um cache , permitindo a diminuição da latência percebida pelo usuário .
6: Pré - busca pode ser empregada baseada em preferências pessoais do usuário , com predição não estatística e com predição estatística .
7

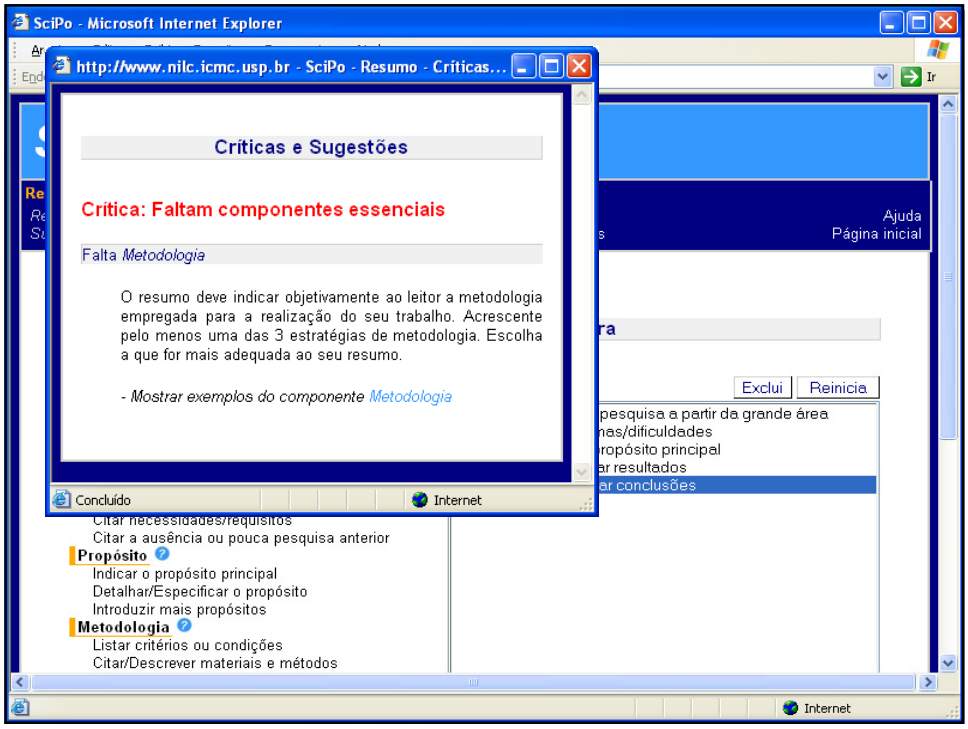

Exemplo
Sumarização automática
Métodos profundos e superficiais
...
Texto.txt - Bloco de notas
Arquivo Editar Formatar Exibir Ajuda
Aviões da Otan bombardearam ontem posições sérvias ao norte de Sarajevo. O ataque foi uma represália à tomada por sérvios de armamentos pesados, retirados da zona de exclusão da ONU em torno da cidade.
Foi a primeira ação da Otan contra sérvios desde o ataque aéreo às suas posições no encrave de Gorazde, em abril.
Um porta-voz do Departamento de Estado dos EUA disse que dois aviões norte-americanos e dois franceses bombardearam às 18h35 (13h35 em Brasília) posições sérvias ao redor de Sarajevo.
Porta-vozes militares disseram que um total de 12 aeronaves, com a participação também de holandeses, saíram de bases da Otan na Itália para realizar os bombardeios.
A ONU disse que após o ataque os sérvios se comprometeram a devolver imediatamente as armas.
O comandante do Exército sérvio garantiu ao comandante das tropas da ONU em Sarajevo, general Michael Rose, que todas as armas retiradas da zona de exclusão seriam devolvidas até hoje.
Rose disse que se a promessa não for cumprida haverá novos ataques
Texto.sum - Bloco de notas
Arquivo Editar Formatar Exibir Ajuda
O comandante do Exército sérvio garantiu ao comandante das tropas da ONU em Sarajevo, general Michael Rose, que todas as armas retiradas da zona de exclusão seriam devolvidas até hoje.
A retirada das armas ocorreu um dia depois de a Iugoslávia -formada por Sérvia e Montenegro- ter interrompido todos os laços com os sérvios da Bósnia, por causa da rejeição por eles de um plano de paz internacional para a região.

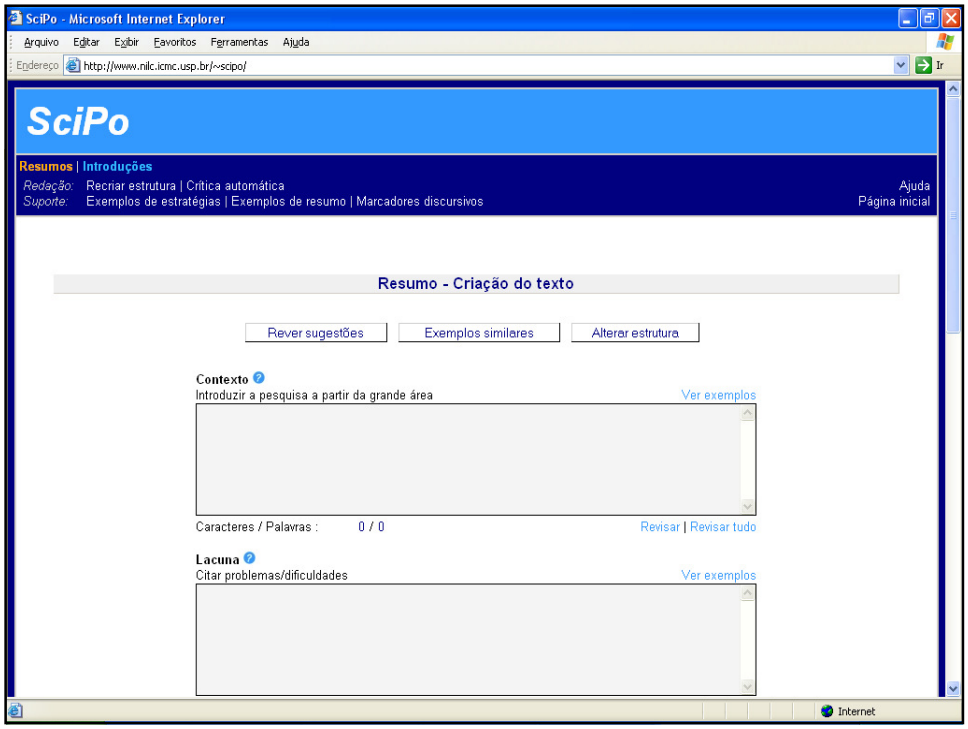
Exemplos

- Auxílio à escrita de textos científicos
  - Regras de estruturação textual
  - Exemplos da estruturas de outros textos
  - Crítica de cada parte do texto

9




SciPo - Microsoft Internet Explorer
http://www.nilc.icmc.usp.br - SciPo - Resumo - Críticas...
Críticas e Sugestões
Crítica: Faltam componentes essenciais
Falta Metodologia
O resumo deve indicar objetivamente ao leitor a metodologia empregada para a realização do seu trabalho. Acrescente pelo menos uma das 3 estratégias de metodologia. Escolha a que for mais adequada ao seu resumo.
- Mostrar exemplos do componente Metodologia
Concluído
Internet
Ajuda
Página inicial
Exclui
Reinicia
Citar a ausência ou pouca pesquisa anterior
Propósito
Indicar o propósito principal
Detalhar/Especificar o propósito
Introduzir mais propósitos
Metodologia
Listar critérios ou condições
Citar/Descrever materiais e métodos
Internet


SciPo - Microsoft Internet Explorer
Arquivo Editar Exibir Favoritos Ferramentas Ajuda
Endereço http://www.nilc.icmc.usp.br/~scipo/
SciPo
Resumos | Introduções
Redação: Recriar estrutura | Crítica automática
Suporte: Exemplos de estratégias | Exemplos de resumo | Marcadores discursivos
Ajuda
Página inicial
Resumo - Criação do texto
Rever sugestões
Exemplos similares
Alterar estrutura
Contexto
Introduzir a pesquisa a partir da grande área
Ver exemplos
Caracteres / Palavras : 0 / 0
Revisar | Revisar tudo
Lacuna
Citar problemas/dificuldades
Ver exemplos
Internet

Exemplos

- WordNet
  - Base de dados lexicais e conceituais
  - Relações entre palavras
    - Sinonímia
    - Antonímia
    - Acarretamento
    - Etc.
  - Relações ontológicas

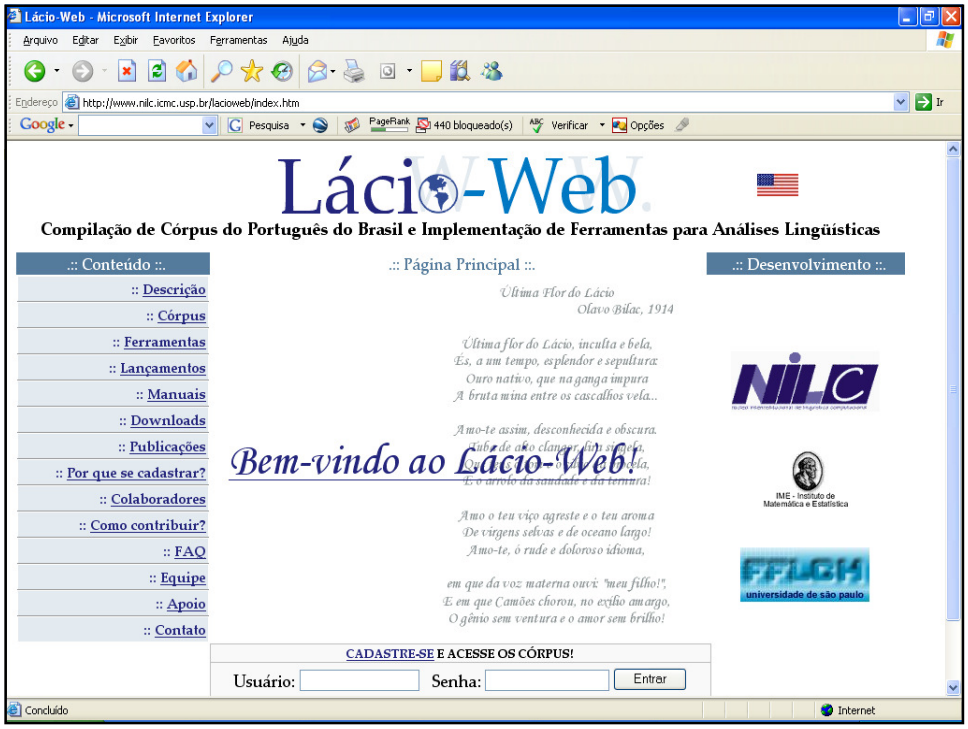
13

14

PLN

- Conhecimento lingüístico é a base para muitos sistemas que manipulam língua natural
  - Extração de conhecimento de córpus
    - Regras gramaticais, sintáticas e discursivas
    - Estrutura textual
    - Regras de tradução
    - Critérios para resumir

15



Conhecimento de mundo

Senso comum

18

PLN no Brasil

- Poucos grupos de pesquisa no país
  - São Carlos
  - Porto Alegre
  - Rio de Janeiro
  - Outros?

19

Recentemente

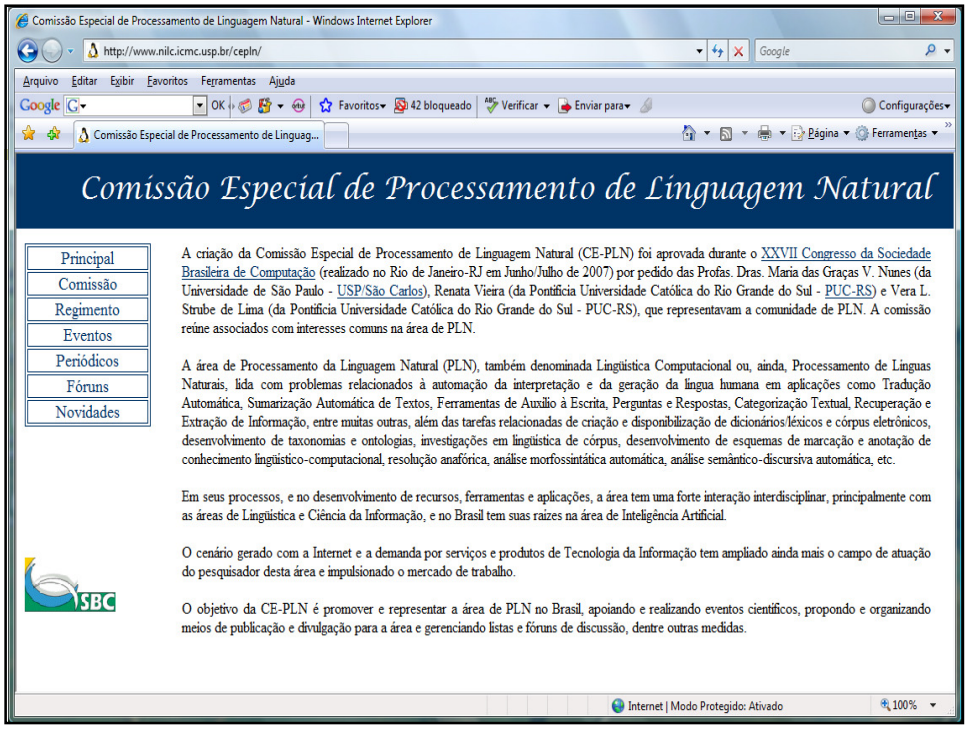
- A área de PLN tem crescido no Brasil
  - Tecnologia da Informação
  - Google Brasil
  - Comissão especial da SBC
  - Eventos científicos próprios melhores e maiores a cada ano
    - Além dos eventos típicos de IA
  - Nascimento de uma revista nacional
  - Iniciativas internacionais importantes

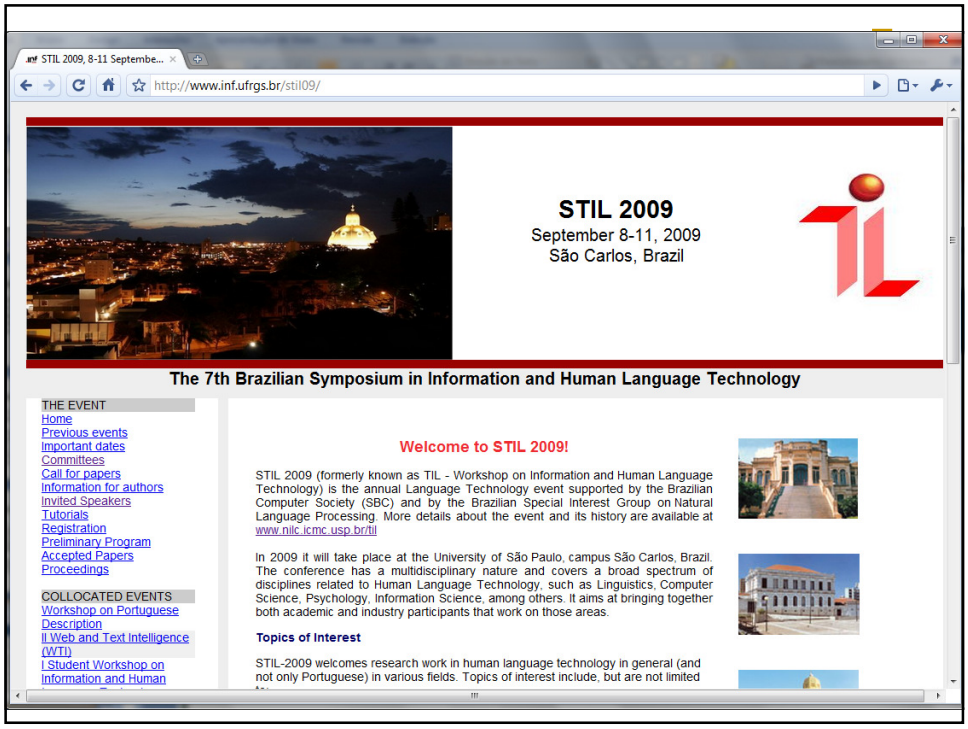
20

Comissão Especial de PLN

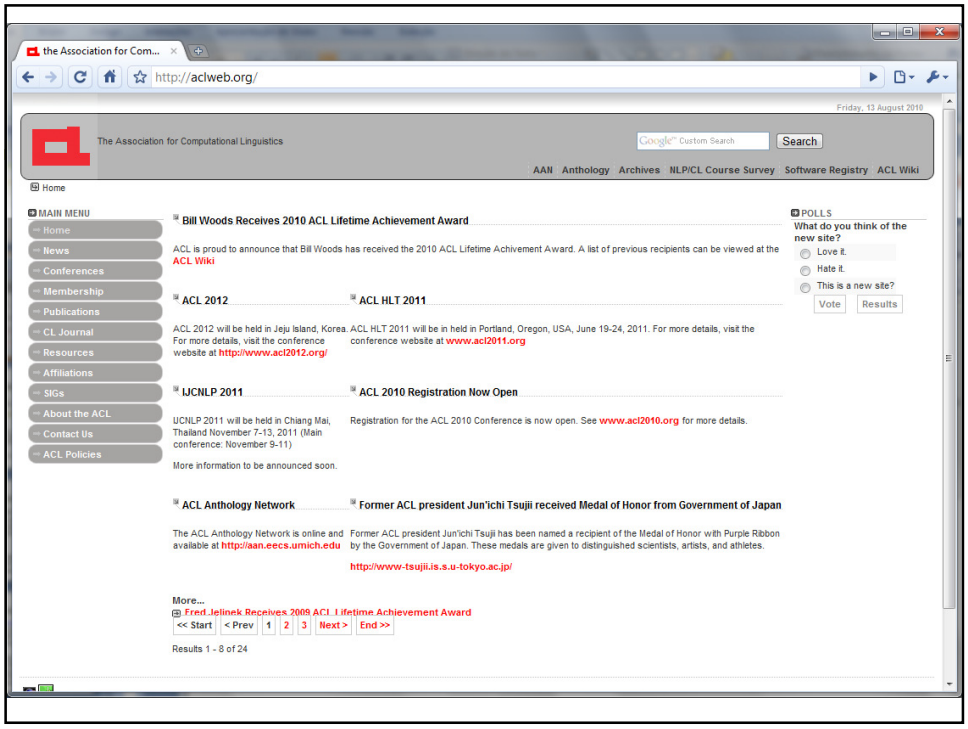
- Composição
  - Thiago A. S. Pardo (USP) - presidente
  - Renata Vieira (PUC-RS)
  - Helena Caseli (UFSCar)
  - Aline Villavicencio (UFRGS)
  - Caroline Gasperin
- www.sbc.org.br/ce-pln
  - Aproximadamente 200 membros na lista de discussão
  - Não precisa ser membro da SBC

21




STIL 2009, 8-11 Septembe...
http://www.inf.ufrgs.br/stil09/
STIL 2009
September 8-11, 2009
São Carlos, Brazil
The 7th Brazilian Symposium in Information and Human Language Technology
THE EVENT
Home
Previous events
Important dates
Committees
Call for papers
Information for authors
Invited Speakers
Tutorials
Registration
Preliminary Program
Accepted Papers
Proceedings
COLLOCATED EVENTS
Workshop on Portuguese Description
II Web and Text Intelligence (WTI)
I Student Workshop on Information and Human
Welcome to STIL 2009!
STIL 2009 (formerly known as TIL - Workshop on Information and Human Language Technology) is the annual Language Technology event supported by the Brazilian Computer Society (SBC) and by the Brazilian Special Interest Group on Natural Language Processing. More details about the event and its history are available at www.nilc.icmc.usp.br/til
In 2009 it will take place at the University of São Paulo, campus São Carlos, Brazil. The conference has a multidisciplinary nature and covers a broad spectrum of disciplines related to Human Language Technology, such as Linguistics, Computer Science, Psychology, Information Science, among others. It aims at bringing together both academic and industry participants that work on those areas.
Topics of Interest
STIL-2009 welcomes research work in human language technology in general (and not only Portuguese) in various fields. Topics of interest include, but are not limited


propor2008.org - International Conference on Computational Processing of Portuguese Language - Windows Internet Explorer
http://www.propor2008.org/
propor 2008
International Conference on Computational Processing of Portuguese Language
8 to 10 of September, Aveiro, Portugal
Final Version and Registration
Final versions and authors early registration are due 19th May.
Publication by Springer
2nd Satellite Event
1st Satellite Event
Support
Home
Organization
Call for papers
Important Dates
Accepted Papers
Registrations
Program Committee
Support
The Venue
Grants
Satellite Events
The International Conference on Computational Processing of Portuguese, former Workshop on Computational Processing of the Portuguese Language - PROPOR - is the main event in the area of Natural Language Processing that is focused on Portuguese and the theoretical and technological issues related to this specific language.
The meeting has been a very rich forum for the interchange of ideas and partnerships for the research communities dedicated to the automated processing of the Portuguese language. PROPOR brings together research groups in the area, promoting the development of methodologies, linguistic resources and projects that can be shared among all researchers and practitioners in the field.
PROPOR, a tri- or bi-anual event, is hosted in Brazil and in Portugal . The meetings have been held in Lisbon , PT (1993), Curitiba, BR (1996), Porto Alegre, BR (1998), Évora, PT (1999), Atibaia, BR (2000), Faro, PT (2003) and Itatiaia, BR (2006).
This coming 8th edition will take place in Aveiro and in nearby Curia, Portugal, being the 1st adopting the International Conference label. The papers selected for oral presentation will be published by Springer (LNCS/LNAI).



Outras iniciativas

- ACL (aclweb.org)
  - *ACL anthology*, listas de discussão, wiki
  - *Registry of Latin American Researchers in Natural Language Processing and Computational Linguistics*
- Linguateca (www.linguateca.pt)
  - Oficialmente finalizado
- forum-lp
- Eventos correlatos
  - Encontro de Lingüística de Córpus
  - Workshop de Descrição do Português
    - Junto ao STIL
- Toolkits
  - GATE, NLTK, Giza++ e Moses, AntMover, etc.

26

Dilemas no Brasil

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

27

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

28

Dilemas no Brasil

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

29

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”

30

Tendências no mundo

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

31

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*, *Natural Language Engineering*, *Machine Translation*, Linguamática, ...
- De âmbito nacional
  - STIL, ELC, ...
  - *Intelligent Computing*, ...

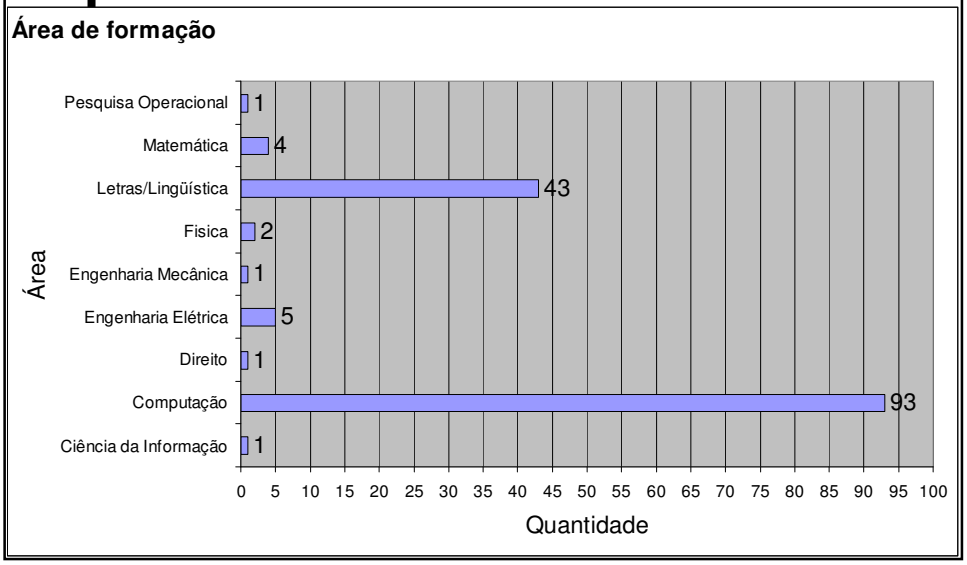
32

PLN no Brasil

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

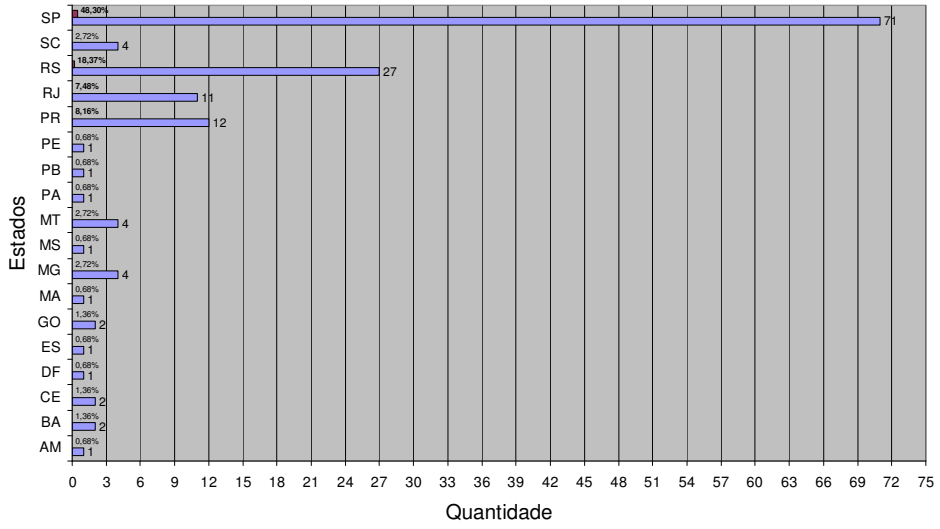
33

PLN no Brasil Pardo et al. (2009)

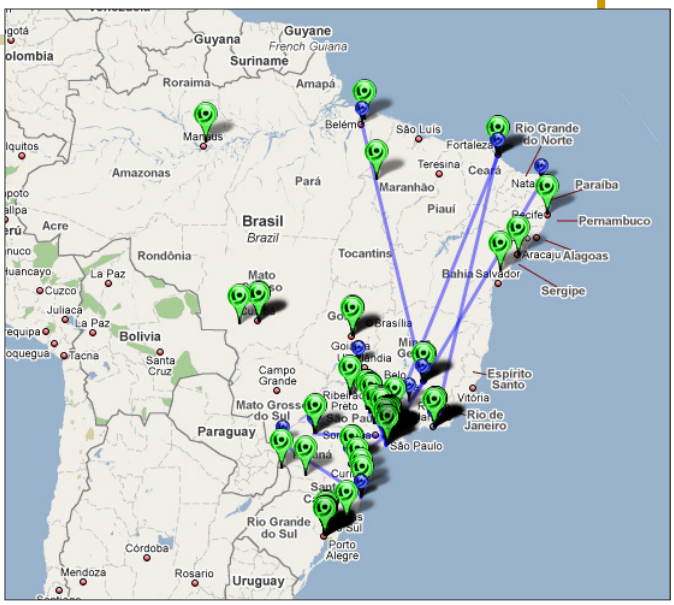


PLN no Brasil Pardo et al. (2009)

Distribuição de pesquisadores por estado

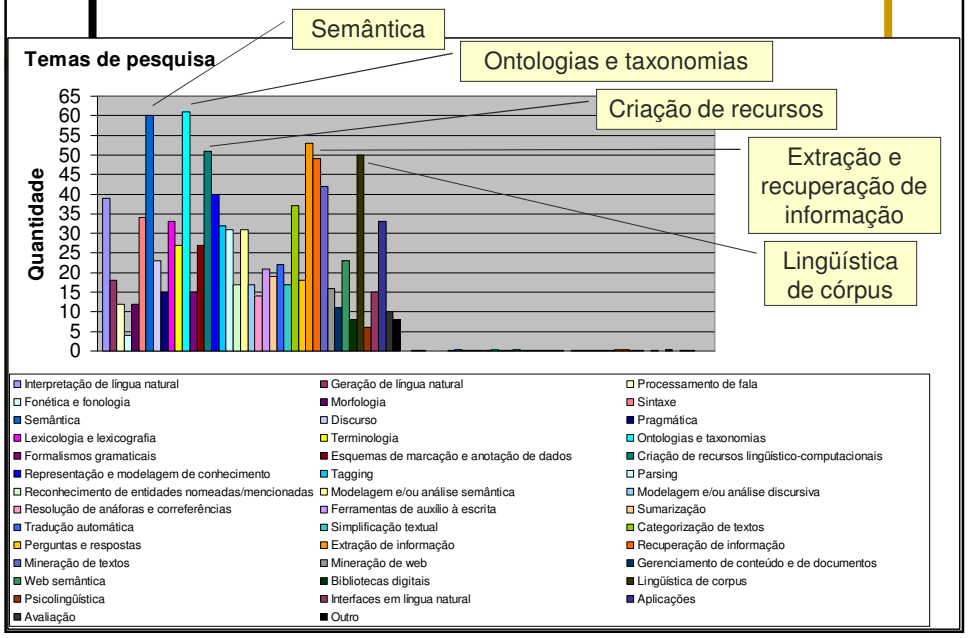


PLN no Brasil Pardo et al. (2009)



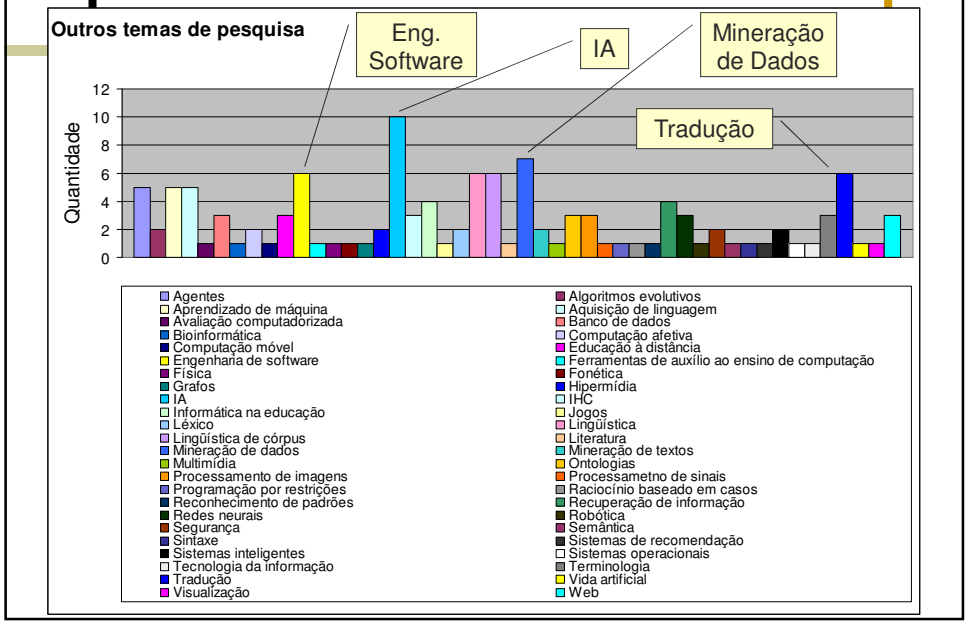
PLN no Brasil

Pardo et al. (2009)



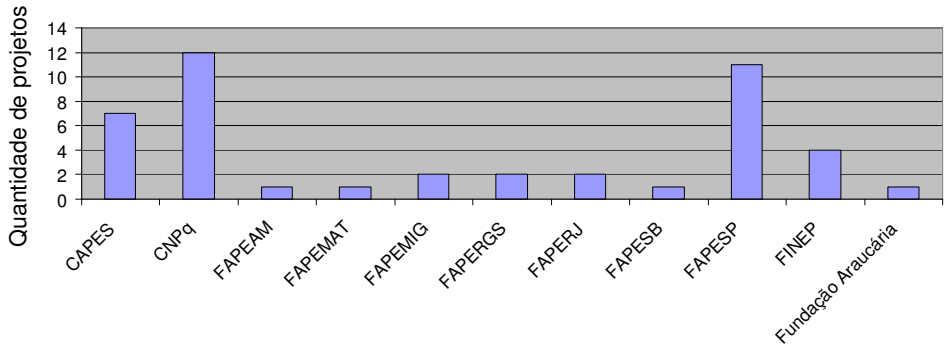
PLN no Brasil

Pardo et al. (2009)



PLN no Brasil Pardo et al. (2009)

Fontes de financiamento



39

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