

Introdução ao Processamento de Línguas Naturais

SCC5908 Introdução ao Processamento de Língua Natural

Thiago A. S. Pardo

1

Recapitulando...

- Abordagens superficiais vs. profundas
- Simbolismo vs. estatística
- Racionalismo (gerativismo) vs. empirismo (estruturalismo, distribucionalismo)
 - Dominância atual do empirismo, trabalhos com base em córpus e em evidência linguística
 - Análises e modelos estatísticos, frequências de fenômenos textuais

2

Abordagens: PLN

- Exemplo: livros de Tom Sawyer (de Mark Twain)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Tokens = 71.370

Types = 8.018 (poucas
para um texto tão grande)
→ para crianças

Taxa type/token = 0,11
(11%)

Em geral, quanto maior
o cópus, menor a taxa

3

Abordagens: PLN

- Distribuição de palavras
 - Lei de **Zipf**
 - George Kingsley Zipf*
 - Baseada em trabalho de Estoup (1916)
 - Proveniente do “Princípio do Mínimo Esforço”, publicado no livro *Human Behavior and the Principle of Least Effort* (1949)

4

Abordagens: PLN

- Distribuição de palavras
 - Lei de Zipf
 - Contam-se quantas vezes cada palavra ocorre em um corpus grande, montando-se um **ranque** em função da **freqüência** delas
 - Há uma relação entre a freqüência e a posição da palavra no ranque
 - **Freqüência x posição no ranque = constante k**
 - Palavra na posição 50 deve ocorrer 3 vezes mais do que palavra na posição 150

5

Abordagens: PLN

- Exemplo: livros de Tom Sawyer
 - Há distorções, comuns na lei de Zipf

Word	Freq. (f)	Rank (r)	f · r
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440
two	104	100	10400

Word	Freq. (f)	Rank (r)	f · r
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

6

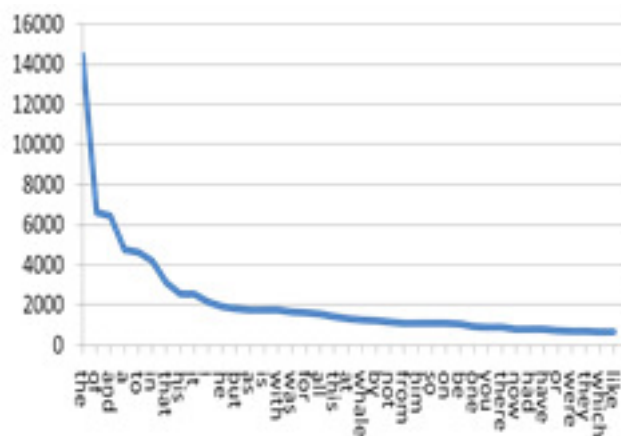
[Abordagens: PLN]

- Distribuição de palavras
 - Lei de Zipf
 - Poucas palavras muito freqüentes
 - Número significativo de palavras de freqüência média
 - Muitas palavras de freqüência baixa
 - É possível plotar um gráfico

7

[Abordagens: PLN]

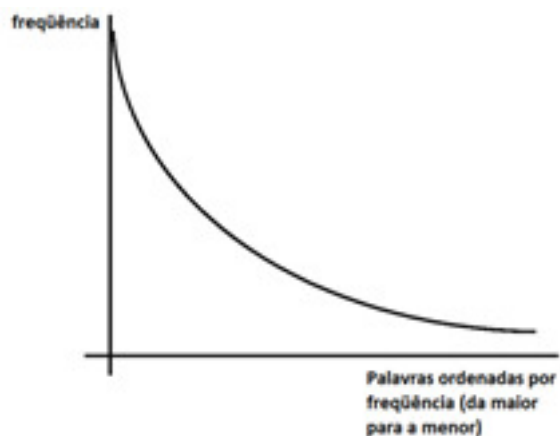
- Exemplo: parte inicial da curva de Zipf para Moby Dick



8

[Abordagens: PLN]

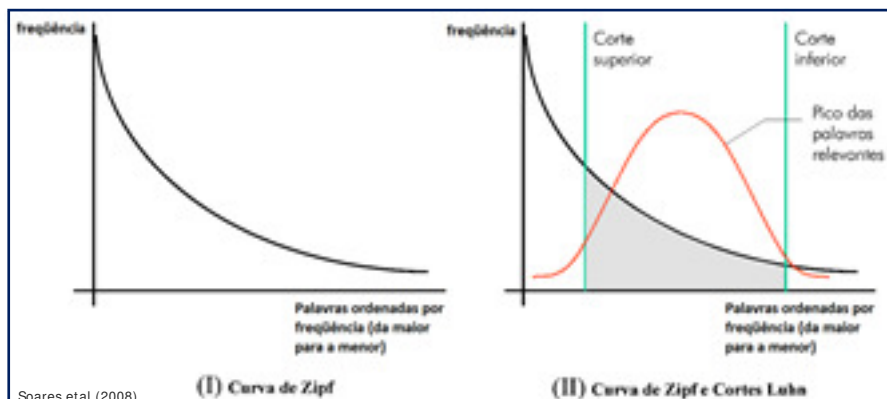
- Curva de Zipf



9

[Abordagens: PLN]

- Distribuição de palavras
 - Curva de Zipf e corte de Luhn (1958)
 - Busca por termos importantes



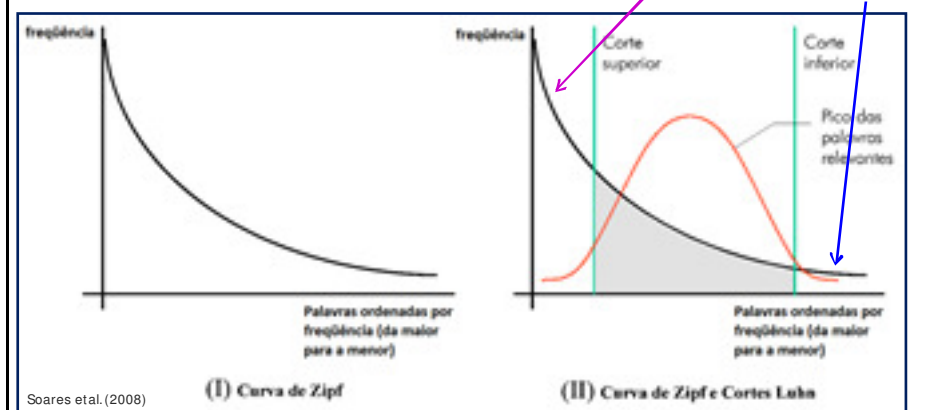
Soares et al. (2008)

Abordagens: PLN

- Distribuição de palavras
 - Curva de Zipf e corte de Luhn (1958)
 - Busca por termos importantes

preposições,
conjunções, etc.

termos raros

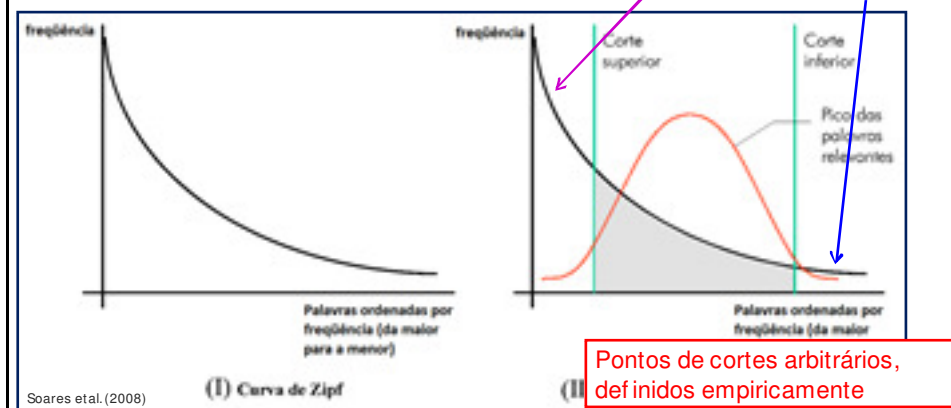


Abordagens: PLN

- Distribuição de palavras
 - Curva de Zipf e corte de Luhn (1958)
 - Busca por termos importantes

preposições,
conjunções, etc.

termos raros



[Abordagens: PLN]

- Distribuição de palavras
 - Outra lei de Zipf
 - O número de significados de uma palavra é correlacionado com sua frequência
 - Palavra com 10.000 ocorrências → 2.1 significados
 - Palavra com 5.000 ocorrências → 3 significados
 - Palavra com 2.000 ocorrências → 4.6 significados

13

[Abordagens: PLN]

- Distribuição de palavras
 - Ainda outras leis de Zipf
 - Uma palavra de conteúdo tende a ocorrer próxima a outra ocorrência sua
 - A frequência de uma palavra é inversamente proporcional ao seu tamanho
 - Quanto maior a frequência de uma palavra, mais “permutações” há (em seus componentes morfológicos)

14

[Abordagens: PLN]

- Leis de Zipf
 - Exageradamente valorizadas
 - Não deveriam ser “leis”, mas “observações” aproximadas
 - Até alguns eventos aleatórios obedecem essas leis
 - Forma de gerar os dados, de construir a curva

15

[PLN]

- Classificação
 - Recursos
 - Ferramentas
 - Aplicações

[Recursos]

- **Cópus**
 - Anotação: humana e/ou automática
 - XML, XCES, TEI, etc.
 - Paralelo, comparável, alinhado, etc.

- **Dicionários monolíngües e bilíngües**
 - *Machine readable vs. machine tractable*

- **Léxicos**
 - Vários paradigmas

17

[Ferramentas]

- Segmentadores textuais: palavras (*tokenizador*), sentenças, parágrafos, tópicos
- Stemmers, lematizadores, nominalizadores
- Etiquetadores morfossintáticos (*taggers*)
- Analisadores sintáticos *shallow (chunkers)* e *deep (parsers)*
- Analisadores semânticos e discursivos
- Alinhadores textuais: lexicais, sentenciais, etc.
- Concordanceadores, *word counting*, etc.

18

[Aplicações]

- Tradutores automáticos
- Revisores ortográficos e gramaticais
- Ferramentas de auxílio à escrita
- Sumarizadores automáticos
- Simplificadores textuais

19

[Recursos, ferramentas e aplicações]

- Atenção
 - Classificação difusa, às vezes
 - Dependente do uso
 - Sumarizador como passo intermediário para recuperação da informação → ferramenta
 - Dicionário eletrônico para consulta → aplicação

20

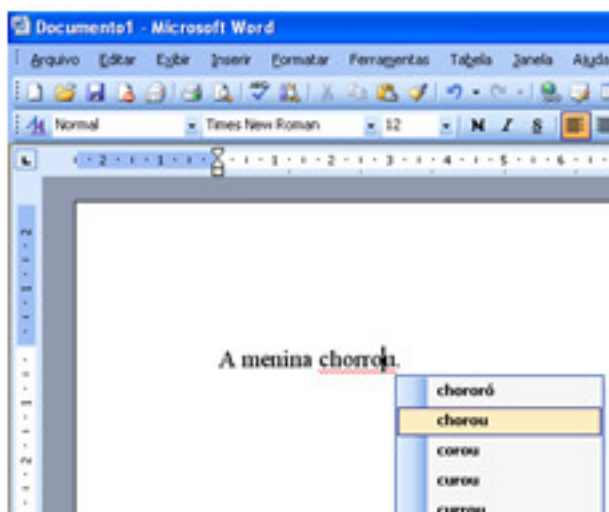
[PLN e áreas correlatas]

- 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
 - Linguística de córpus

21

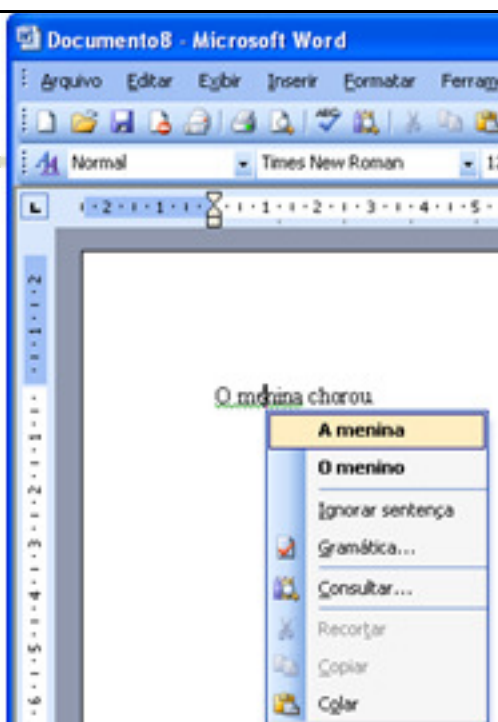
[Exemplos]

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



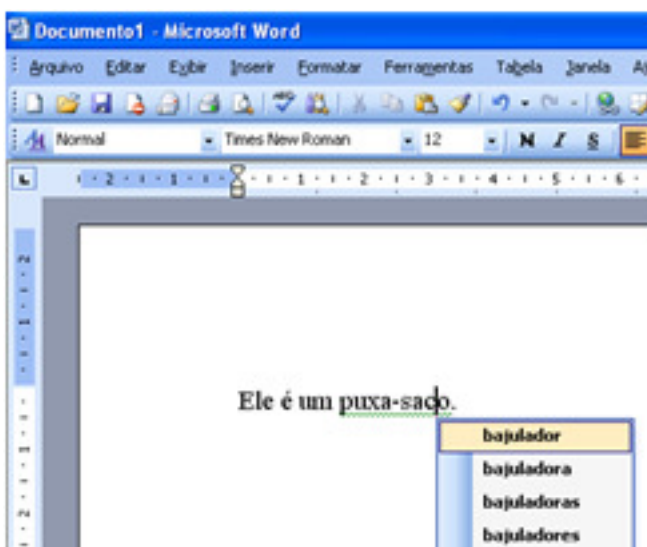
Exemplos

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



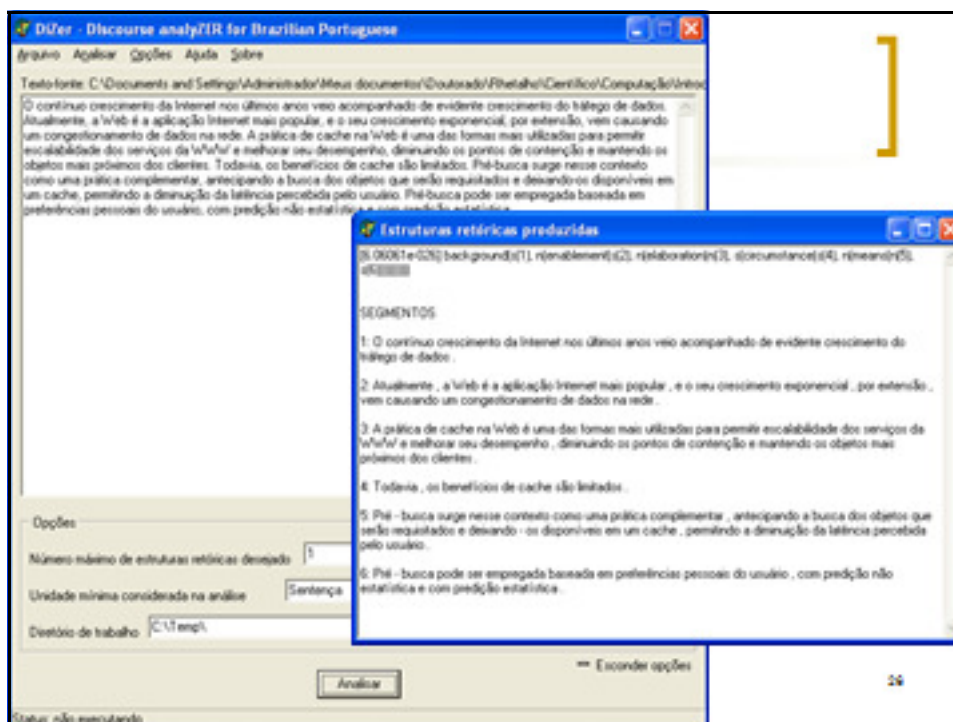
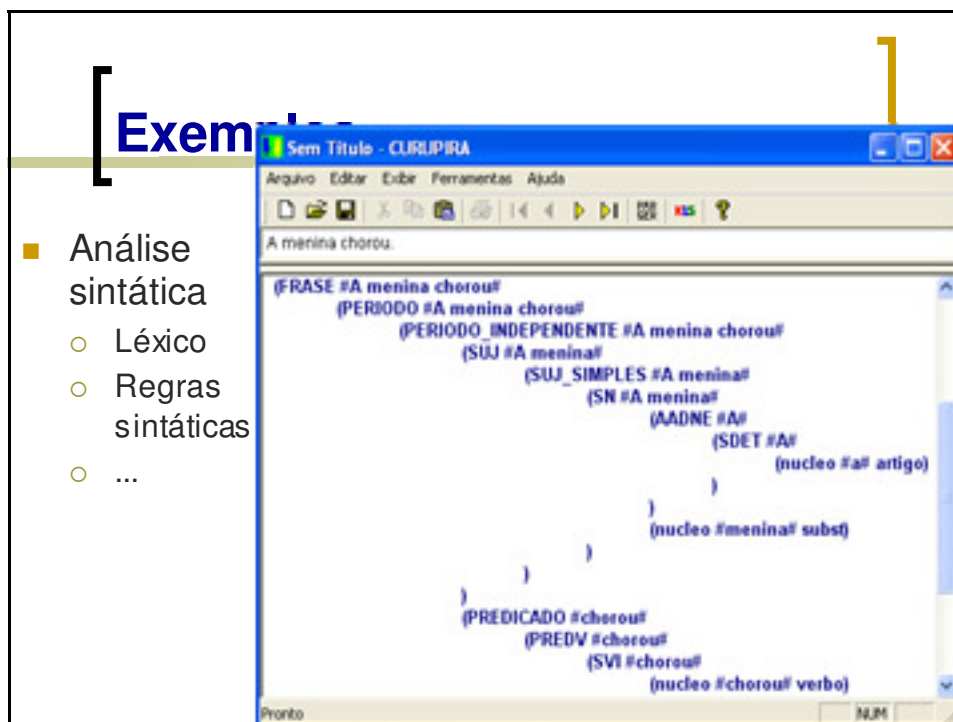
Exemplos

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



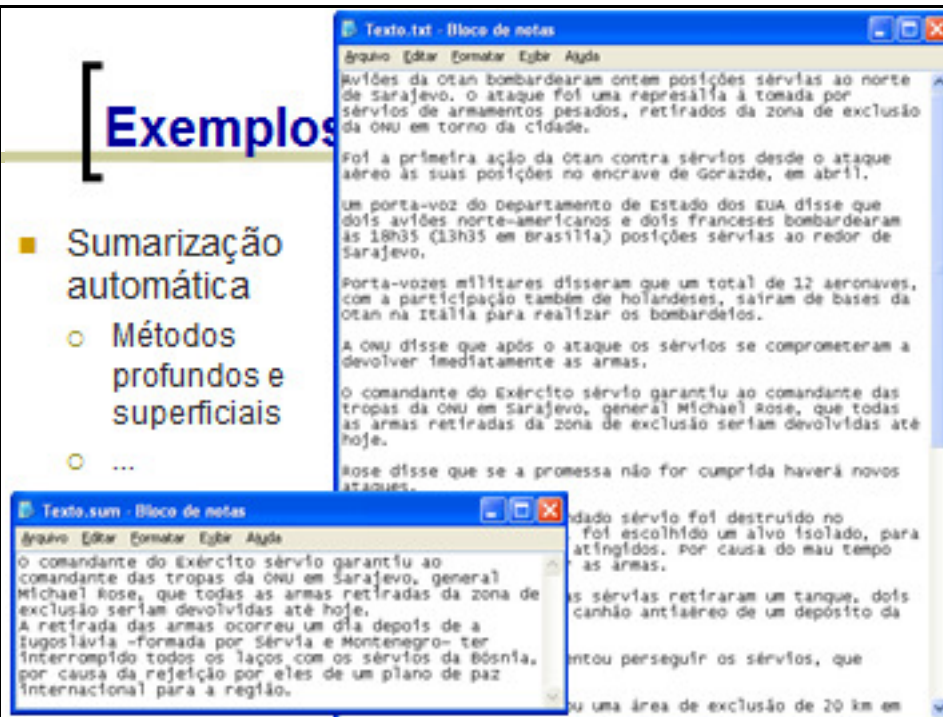
Exemplos

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



Exemplos

- Sumarização automática
 - Métodos profundos e superficiais
 - ...



The screenshot shows two windows of a text editor. The top window, titled 'Texto.txt - Bloco de notas', contains a news article in Portuguese. The bottom window, titled 'Texto.sum - Bloco de notas', contains a summary of the article. The article text is as follows:

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 enclave 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.

The summary text in the bottom window is:

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 Jugoslávia - 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

28

SciPo - Microsoft Internet Explorer

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Endereço: http://www.nlc.icmc.usp.br/~scipo/

SciPo

Resumos | Introduções

Redação: Recrear estrutura | Crítica automática
 Suporte: Exemplos de estratégias | Exemplos de resumo | Marcadores discursivos

Ajuda
Página inicial

Resumo - Seleção da estrutura

Sobre Desce Excluir Reinicia

Contexto

- Declarar proeminência do tópico
- Familiarizar termos e conceitos
- Introduzir a pesquisa a partir da grande área

Lacuna

- Citar problemas/dificuldades
- Citar necessidades/requisitos
- 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

Contexto: Introduzir a pesquisa a partir da grande área
 Lacuna: Citar problemas/dificuldades
 Propósito: Indicar o propósito principal
 Resultado: Apresentar resultados
 Conclusão: Apresentar conclusões

Internet

SciPo - Microsoft Internet Explorer

http://www.nlc.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

Citar necessidades/requisitos
 Citar a ausência ou pouca pesquisa anterior

Propósito

- Indicar o propósito principal
- Detalhar/Especificar o propósito
- Introduzir mais propósitos

Metodologia

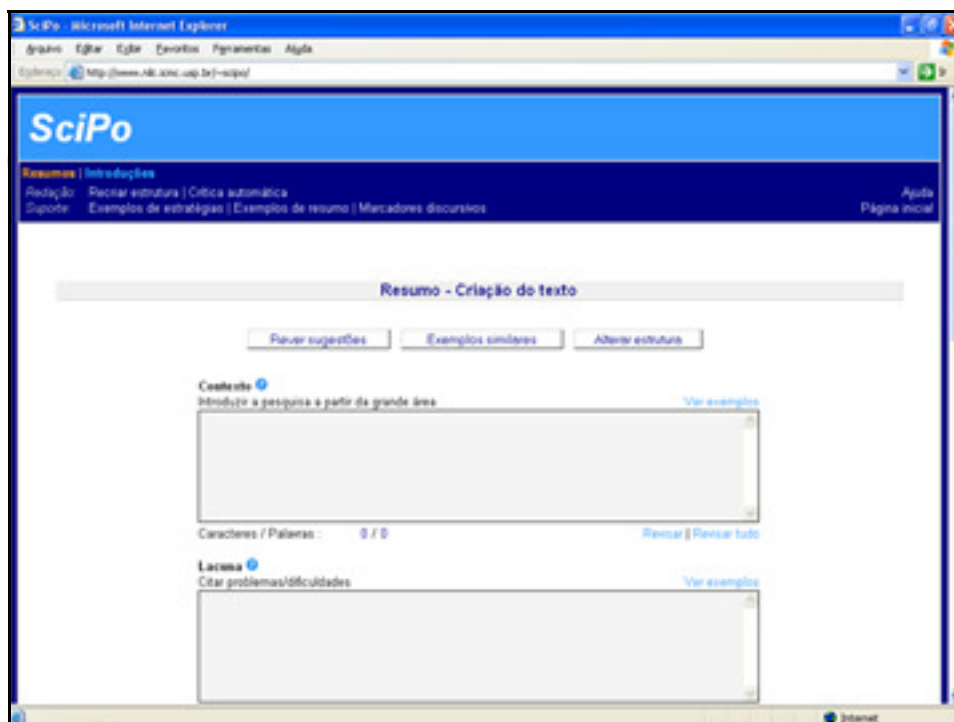
- Listar critérios ou condições
- Citar/Descrever materiais e métodos

Ajuda
Página inicial

Excluir Reinicia

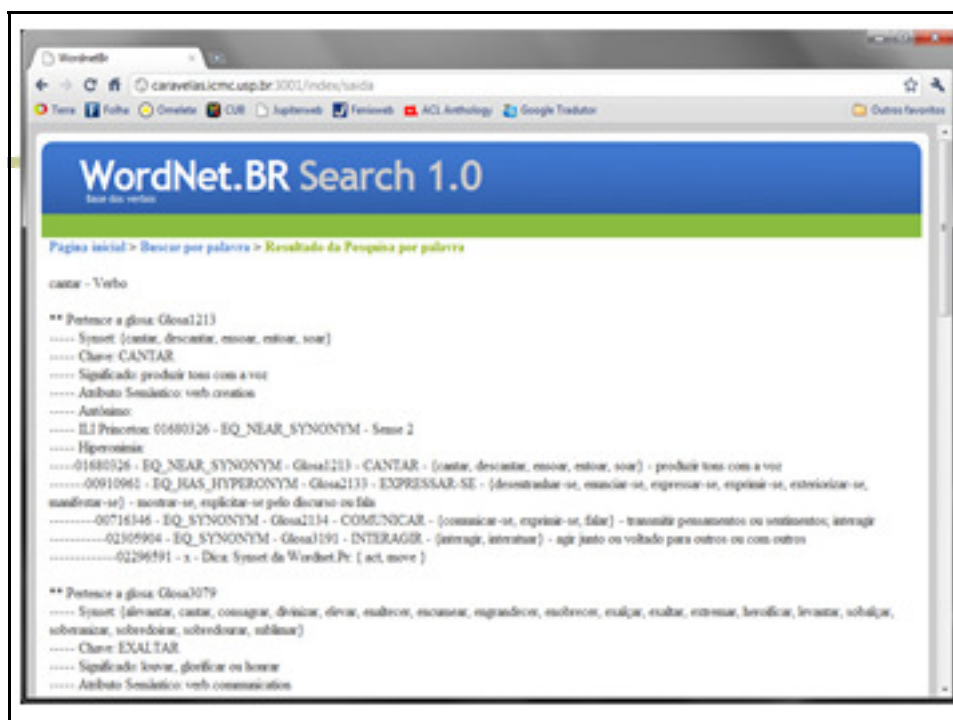
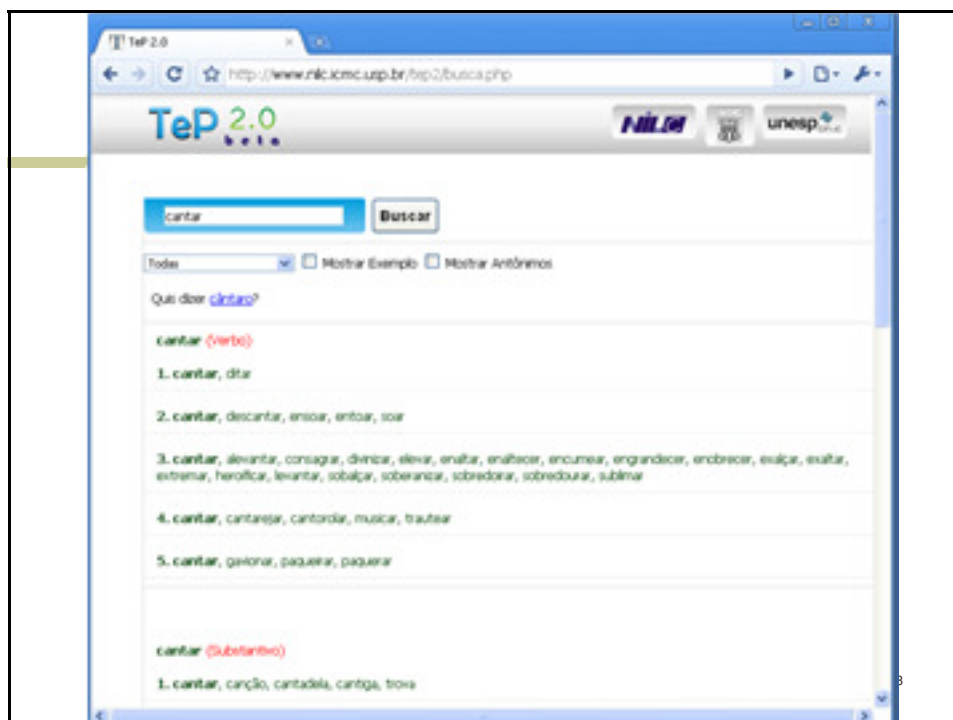
pesquisa a partir da grande área
 ros/dificuldades
 propósito principal
 er resultados
 ar conclusões

Internet



Exemplos

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



[PLN]

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

35

The screenshot shows the Lácio-Web website interface. At the top, the browser window title is 'Lácio-Web - Microsoft Internet Explorer'. The address bar shows the URL 'http://www.nilc.ufsc.br/lacioweb/index.htm'. The page content includes:

- Header:** 'Lácio-Web' logo with a globe icon, and a subtitle 'Compilação de Corpus do Português do Brasil e Implementação de Ferramentas para Análises Lingüísticas'.
- Navigation:** A left sidebar menu with options like 'Conteúdo', 'Descrição', 'Corpus', 'Ferramentas', 'Lançamentos', 'Mensais', 'Downloads', 'Publicações', 'Por que se cadastrar?', 'Colaboradores', 'Como contribuir?', 'FAQ', 'Equipe', 'Após', and 'Contato'. There are also 'Página Principal' and 'Desenvolvimento' links.
- Main Content:** A quote from 'Última flor do Lácio' by Carlos Drummond de Andrade (1934): 'Última flor do Lácio, inculta e bela, / É, a um tempo, esplêndida e sepulta. / Que saibas, que na língua imposta / A fruta viva entre os cascalhos volta.' Below the quote is a large 'Bem-vindo ao Lácio-Web!' message.
- Logos:** Logos for NILC (Núcleo de Informática em Língua Portuguesa) and FFLCH (Faculdade de Filosofia, Letras e Ciências Humanas) are visible on the right side.
- Footer:** A login section titled 'CADASTRE-SE E ACESSO OS CÓRPUIS' with fields for 'Usuário:' and 'Senha:' and an 'Entrar' button.

Conhecimento de mundo



Senso comum



38

[PLN no Brasil]

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

39

[Recentemente]

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

40

[Comissão Especial de PLN]

- Composição
 - Thiago A. S. Pardo (USP)
 - 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

41

Comissão Especial de Processamento de Linguagem Natural - Windows Internet Explorer

http://www.nlc.usp.br/cepln/

Comissão Especial de Processamento de Linguagem Natural

Principal
Comissão
Regimento
Eventos
Periódicos
Fóruns
Novidades

A criação da Comissão Especial de Processamento de Linguagem Natural (CE-PLN) foi aprovada durante o XXXVII Congresso da Sociedade Brasileira de Computação (realizado no Rio de Janeiro-RJ em Junho/Julho de 2007) por pedido dos Profs. Dras. Maria das Graças V. Nunes (da Universidade de São Paulo - USP/São Carlos), Renata Vieira (da Pontifícia Universidade Católica do Rio Grande do Sul - PUC-RS) e Vera L. Strube de Lima (da Pontifícia Universidade Católica do Rio Grande do Sul - PUC-RS), que representam a comunidade de PLN. A comissão reúne associados com interesses comuns na área de PLN.

A área de Processamento da Linguagem Natural (PLN), também denominada Linguística Computacional ou, ainda, Processamento de Língua Natural, lida com problemas relacionados à automação da interpretação e da geração da língua humana em aplicações como Tradução Automática, Sumarização Automática de Textos, Ferramentas de Apoio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação, entre muitos outros, além das tarefas relacionadas de criação e disponibilização de dicionários léxicos e corpora eletrônicos, desenvolvimento de taxonomias e ontologias, investigações em linguística de corpus, desenvolvimento de esquemas de marcação e anotação de conhecimento linguístico-computacional, resolução anafórica, análise morfosintática automática, análise semântico-discursiva automática, etc.

Em seus processos, e no desenvolvimento de recursos, ferramentas e aplicações, a área tem uma forte interação interdisciplinar, principalmente com as áreas de Linguística e Ciência da Informação, e no Brasil tem suas raízes na área de Inteligência Artificial.

O cenário gerado com a Internet e a demanda por serviços e produtos de Tecnologia da Informação tem ampliado ainda mais o campo de atuação do pesquisador desta área e impulsionado o mercado de trabalho.

O objetivo da CE-PLN é promover e representar a área de PLN no Brasil, apoiando e realizando eventos científicos, propondo e organizando meios de publicação e divulgação para a área e gerencando listas e listas de discussão, dentre outras medidas.

Internet | Modo Protegido Ativado 100%

STIL 2011, 26-28 October ...
www.ufmt.br/stil2011/

STIL 2011
The 8th Brazilian Symposium in Information and Human Language Technology

The 8th Brazilian Symposium in Information and Human Language Technology
October 24-26 | Mato Grosso, Brazil

Welcome to STIL 2011!

STIL (formerly known as TIL - Workshop on Information and Human Language Technology) is the bi-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.stil.ufmt.br/stil

In 2011 it will take place at Cuiabá, Mato Grosso, 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

The STIL 2011 welcomes research work in human language technology in general (and not only Portuguese) in various fields. Topics of interest include, but are not limited to:

- * Natural Language Processing Applications.
- * Natural Language Resources & Tools.
- * User Studies and Evaluation Methods.
- * Corpus Linguistics.
- * Phonology/Morphology, Tagging and Chunking, Word Segmentation.
- * Terminology, Lexicology and Lexicography.
- * Lexical Semantics - Grammatical Formalisms, Syntax and Parsing.
- * Semantics, Semantic Representations and Semantic Parsing.

THE EVENT

COLLOCATED EVENTS

LOCAL INFORMATION

Mapa para o Hotel Holiday Inn Cuiabá

A: UFMT
B: Holiday Inn Cuiabá

Propor 2010 | International ...
www.inf.pucrs.br/~propor2010/

propor 2010
International Conference on Computational Processing of the Portuguese Language
April 27-30, 2010, Porto Alegre - RS, Brazil

ABOUT THE EVENT

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.

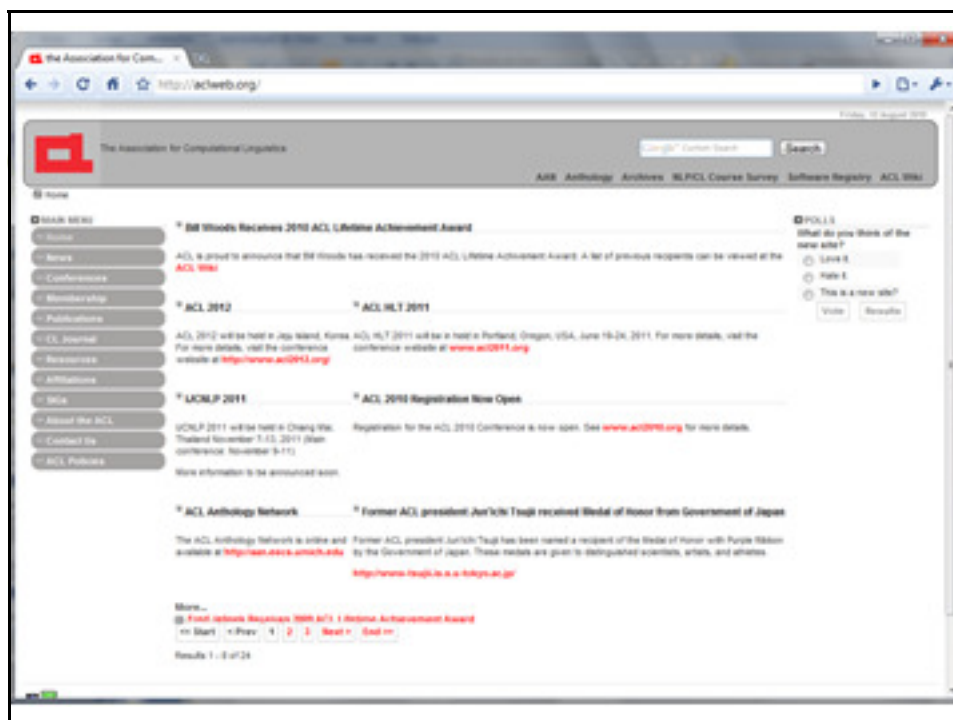
PROPOR, a bi- or tri-annual event, is hosted in Brazil and in Portugal. The meetings have been held in Lisbon (PT, 2003), Curitiba (BR, 2005), Porto Alegre (BR, 2007), Coimbrã (PT, 2009), Alameda (BR, 2010), Funchal (PT, 2011), Guimarães (BR, 2012) and Aveiro (PT, 2013).

NEWS

- April 1, 2010
[Book of Abstracts](#)
- March 26, 2010
[Instructions for Authors](#)
- January 22, 2010
[Preliminary Program](#)
- December 27, 2009
[Registration is now open](#)

SUPPORTS

about the event
call for papers
important dates
instructions for authors
program
invited speaker
accepted papers
subfields
phd and msc context
dissert.
springer proceedings
extended proceedings
social events
photos
registration
local funding
venue
accommodation
about Porto Alegre
organizing committee
steering committee (sc)
contact



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 Linguística de Córpus
 - Workshop de Descrição do Português
 - Junto ao STIL
- Toolkits
 - GATE, NLTK, Giza++ e Moses, AntMover, etc.

[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

47

[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

48

[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*

49

[Dilemas no Brasil]

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

50

[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”

51

[Tendências no mundo]

- E-mails e mensagens instantâneas

- Blogs e microblogs

- Redes sociais

- Análise de opiniões
 - *Sentiment analysis*

52

[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*

53

[PLN: onde encontrar]

- **De âmbito internacional**
 - ACL, NAACL, EACL, HLT, COLING, EMNLP, Interspeech, PROPOR, CÍCLING, CoNLL, EAMT, IJCNLP, LAW, LREC, RANLP, Corpus Linguistics, ...
 - *Computational Linguistics, Natural Language Engineering, Machine Translation, Linguamática, ...*

- **De âmbito nacional**
 - STIL, JDP, ELC, ...

54

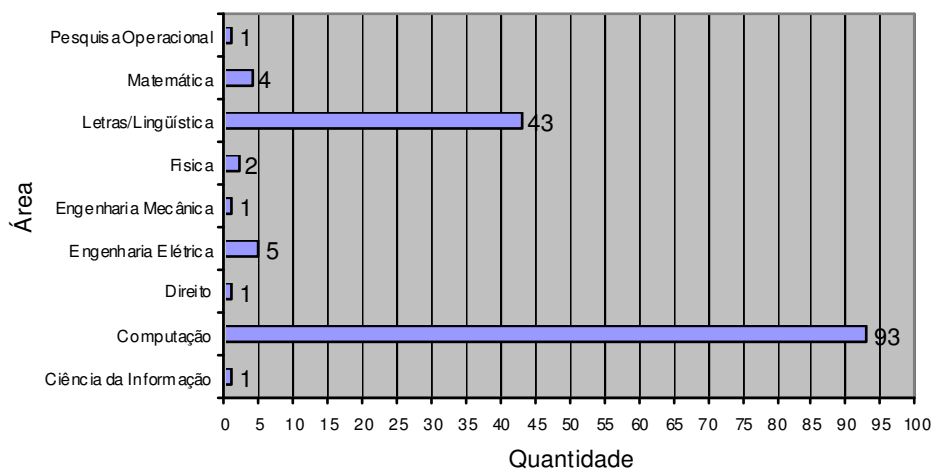
[PLN no Brasil]

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

55

[PLN no Brasil Pardo et al. (2009)]

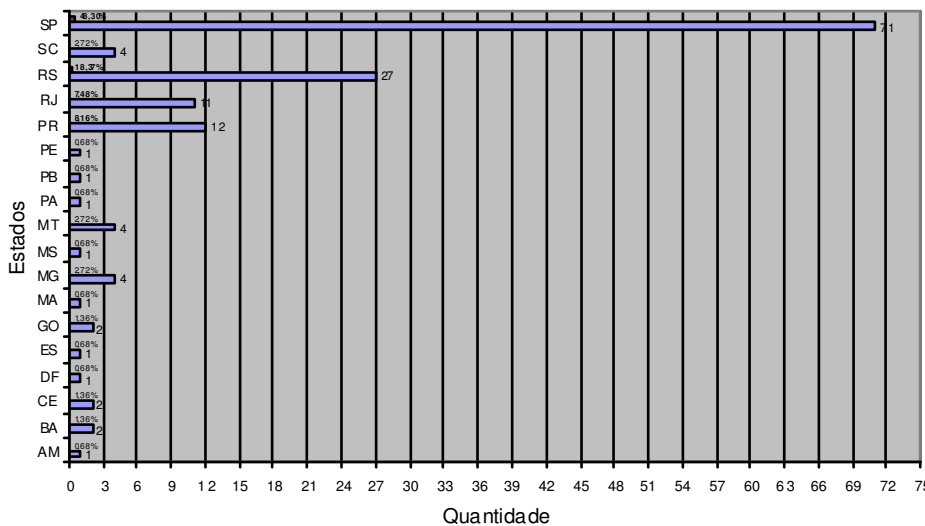
Área de formação



PLN no Brasil

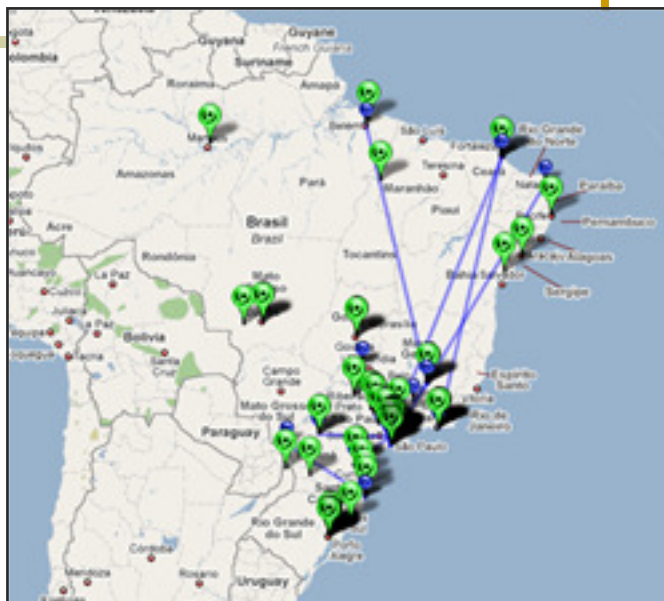
Pardo et al. (2009)

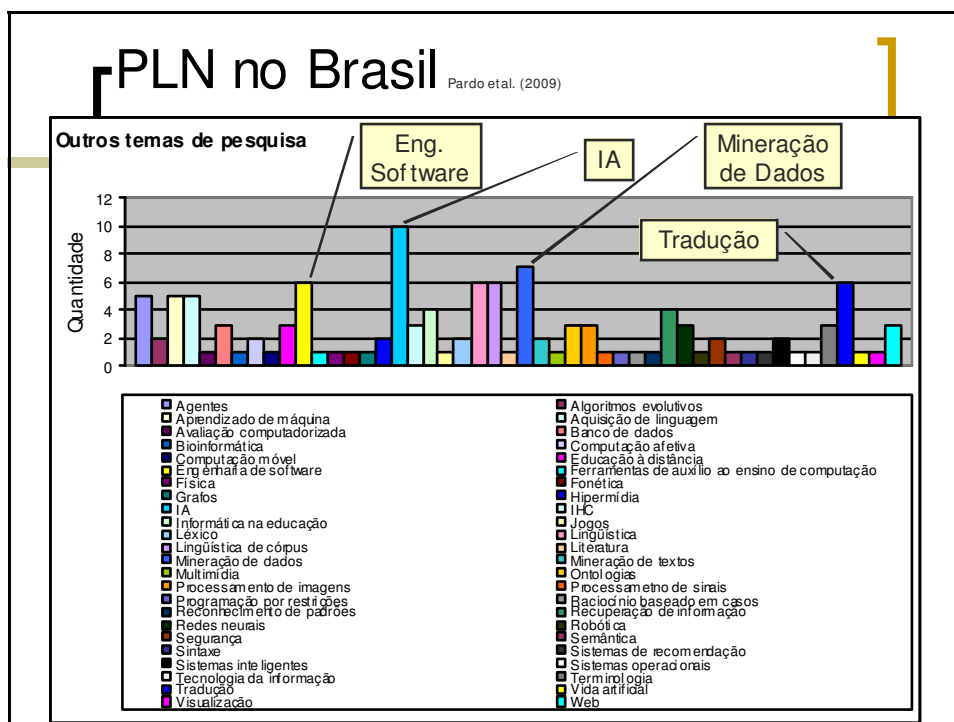
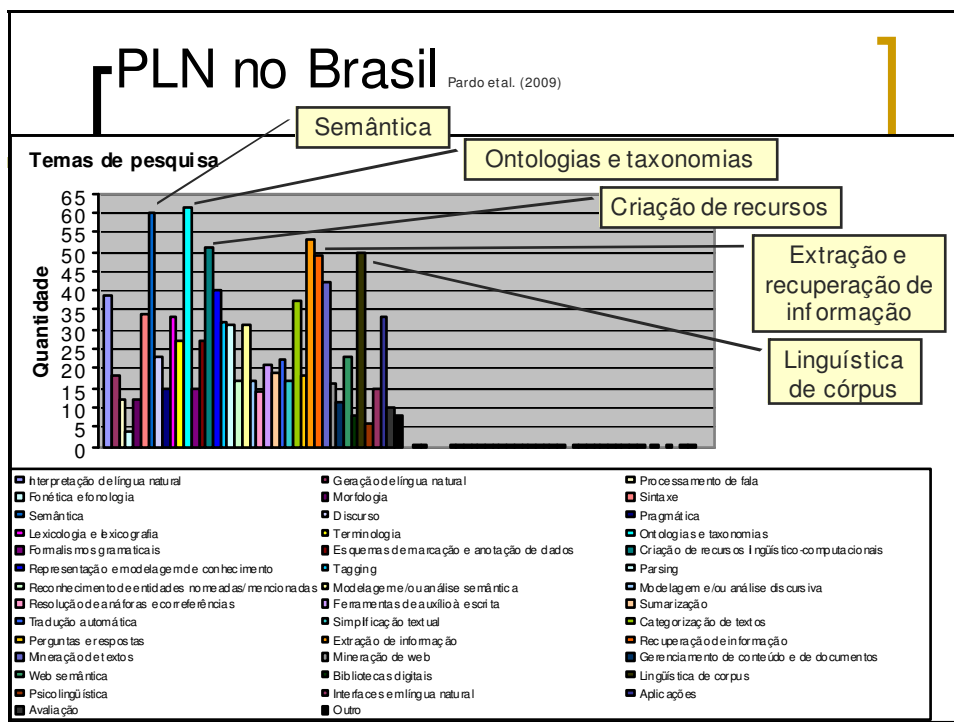
Distribuição de pesquisadores por estado



PLN no Brasil

Pardo et al. (2009)

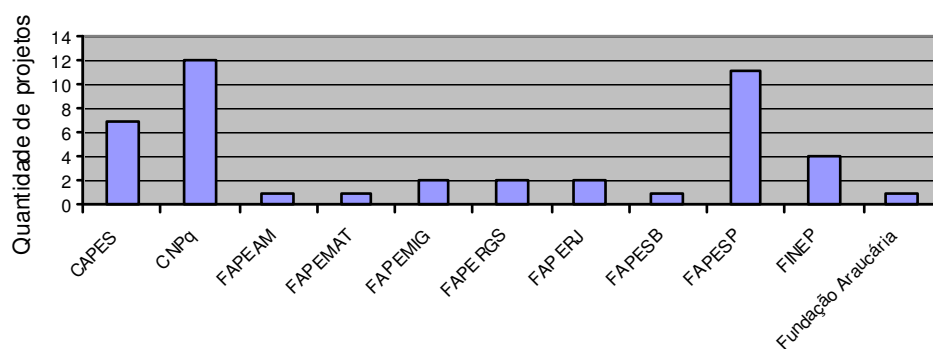




PLN no Brasil

Pardo et al. (2009)

Fontes de financiamento



61

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