

Introdução ao Processamento de Línguas Naturais

SCC5869 Tópicos em Processamento de Língua Natural

Thiago A. S. Pardo

1

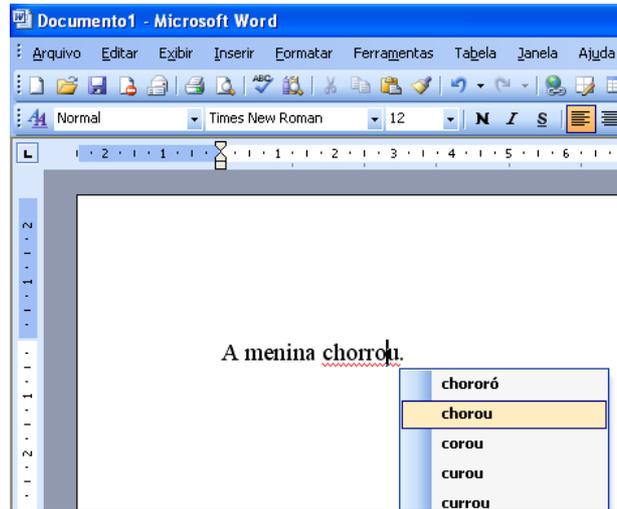
PLN e áreas correlatas

- 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

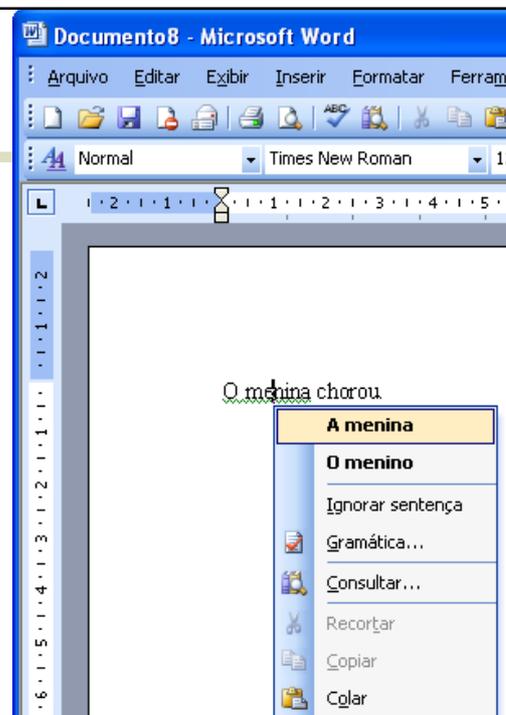
[Exemplos]

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



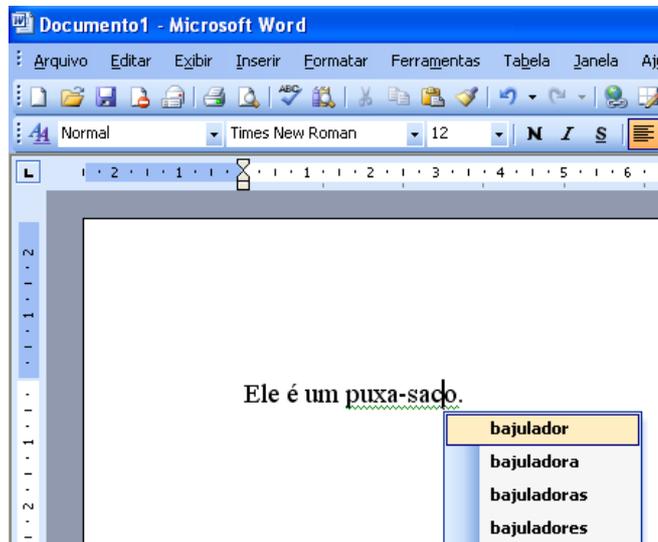
[Exemplos]

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



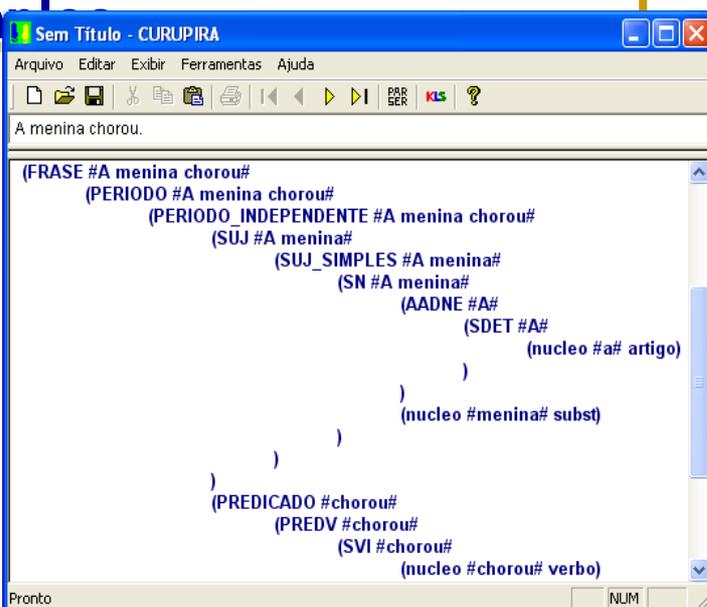
Exemplos

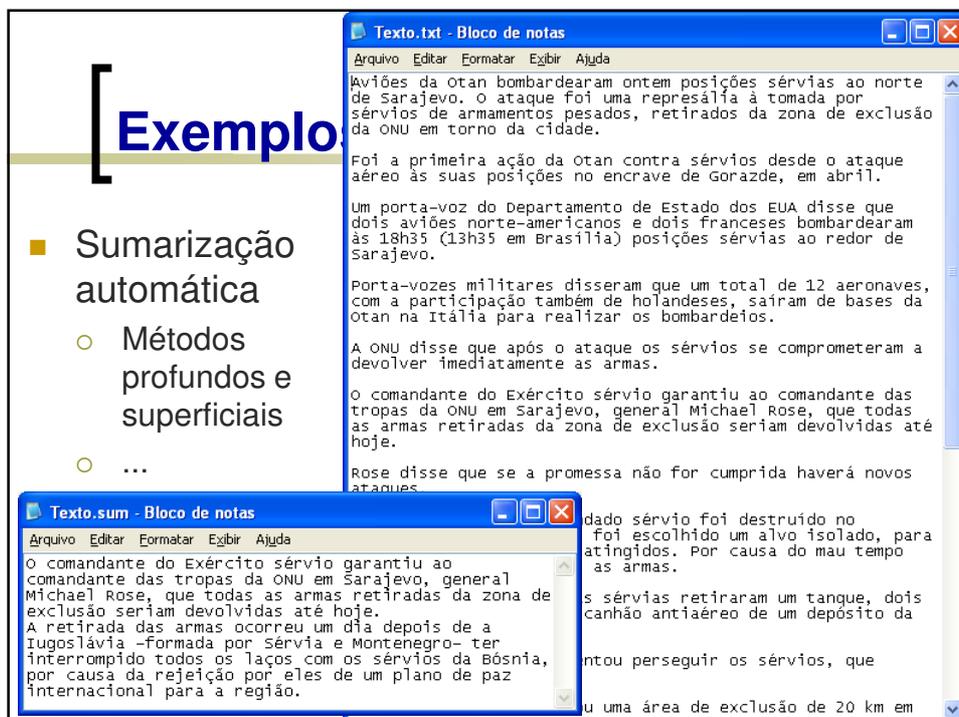
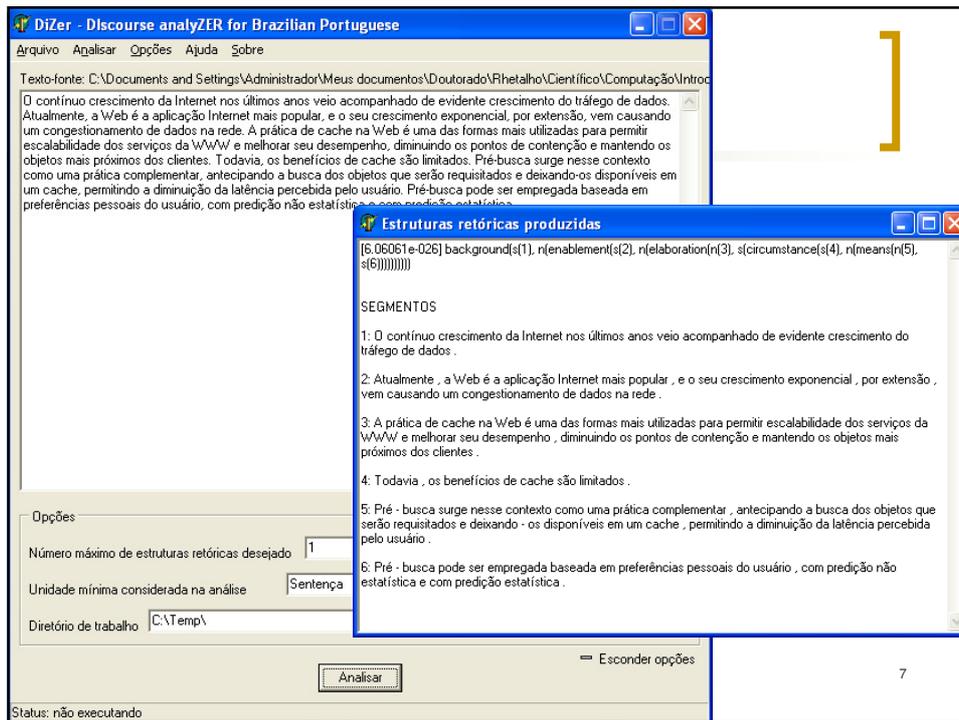
- Revisão estilística
 - Tokenizador
 - Regras estilísticas
 - ...



Exemplos

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





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

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Endereço http://www.nilc.icmc.usp.br/~scipo/ Ir

SciPo

Resumos | Introduções

Redação: Recriar estrutura | Crítica automática Ajuda
 Suporte: Exemplos de estratégias | Exemplos de resumo | Marcadores discursivos Página inicial

Resumo - Seleção da estrutura

Sobe Desce Exclui 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

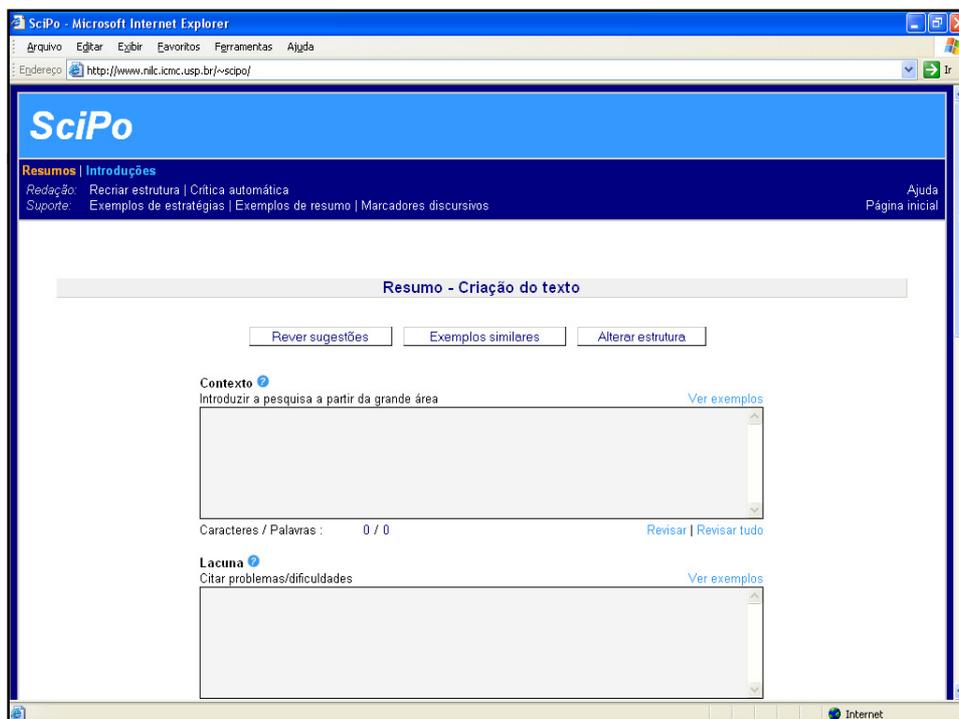
- Indicar o propósito principal
- Detalhar/Especificar o propósito
- Introduzir mais propósitos

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



Exemplos

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

13

The screenshot displays the TeP 2.0 beta web interface. The browser address bar shows the URL <http://www.nlc.icmc.usp.br/tep2/busca.php>. The page features a search bar with the word "cantar" entered and a "Buscar" button. Below the search bar, there are options for "Todas" (selected), "Mostrar Exemplo", and "Mostrar Antônimos". The search results are organized into sections for "cantar (Verbo)" and "cantar (Substantivo)".

cantar (Verbo)

1. cantar, ditar
2. cantar, descantar, ensoar, entoar, soar
3. cantar, levantar, consagrar, divinizar, elevar, enaltar, enaltecer, encumear, engrandecer, enobrecer, exaltar, extremar, heroificar, levantar, sobalçar, soberanizar, sobredoirar, sobredourar, sublimar
4. cantar, cantarejar, cantorolar, musicar, trautear
5. cantar, gavionar, paqueirar, paquerar

cantar (Substantivo)

1. cantar, canção, cantadela, cantiga, trova

[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

Lácio-Web
Compilação de Corpus do Português do Brasil e Implementação de Ferramentas para Análises Linguísticas

Última Flor do Lácio
Olavo Bilac, 1914

Última flor do Lácio, inculta e bela,
És, a um tempo, esplendor e sepultura:
Ouro nativo, que na ganga impura
A bruta mina entre os cascalhos vela...

Amo-te assim, desconhecida e obscura,
Autor de um clarear, d'um sorriso,
Es, o ardo da saudade e da ternura!

Amo o teu viço agreste e o teu aroma
De virgens selvagens e de oceano largo!
Amo-te, ô rude e doloroso idioma,
em que da voz materna ouví "meu filho!",
E em que Camões chorou, no exílio amargo,
O gênio sem ventura e o amor sem brilho!

Bem-vindo ao Lácio-Web!

CADASTRE-SE E ACESSE OS CÓRPUS!

Usuário: Senha: Entrar

Conhecimento de mundo

Página principal - Wikipédia, a enciclopédia livre - Windows Internet Explorer

http://pt.wikipedia.org/wiki/P%C3%A1gina_principal

Google

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Google wikipedia.pt OK 42 bloqueado Verificar Enviar para Configurações

W Página principal - Wikipédia, a enciclopédia livre

Entrar / criar conta

artigo discussão ver fonte história

Boas-vindas Guia Perguntas frequentes Comunidade Políticas da Wikipédia Doações Contatos

Bem-vindo(a) à Wikipédia,
a enciclopédia livre que todos podem editar.

13h29min (UTC); Quarta-feira, 7 de Maio de 2008
376 479 artigos · Portais · Índice geral

☆ Artigo em destaque

Paracetamol (DCI), ou **acetaminofeno**, é um fármaco com propriedades analgésicas, mas sem propriedades antiinflamatórias clinicamente significativas. Atua por inibição da síntese das prostaglandinas, mediadores celulares responsáveis pelo aparecimento da dor. Esta substância tem também efeitos antipiréticos. É utilizado nas seguintes formas de apresentação: cápsulas, comprimidos, gotas, xaropes e injetáveis. Atualmente é um dos analgésicos mais utilizados por ser bastante seguro e não interagir com a maioria dos medicamentos.

Faz parte da composição de uma série de medicamentos usados

Representação da molécula do paracetamol.

Eventos recentes

- Mianmar contabiliza oficialmente mais de 22.460 mortos e 41.000 desaparecidos pela passagem do **Ciclone Nargis** (imagem) no país.
- China inaugura a maior ponte marítima do mundo, chamada **Ponte da Baía de Hangzhou**, com 36 mil metros de comprimento.
- Arqueólogos datam pinturas a óleo em **Bamiyan** (Afeganistão) no século VII, antecedendo em seis séculos a mais antiga descoberta conhecida do gênero.
- Tremor de magnitude 5,2 graus na escala de Richter, com epicentro no oceano, é sentido na cidade de **São Paulo** e em mais

Internet | Modo Protegido: Ativado 100%

Senso comum

OpenMind Commonsense no Brasil - Windows Internet Explorer

http://www.sensocomum.ufscar.br/2080/omcs/login_pt_BR.jsp

Google

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Google open mind common se 42 bloqueado Verificar Enviar para open Configurações

OpenMind Commonsense no Brasil

OPEN MIND
common sense no Brasil
Ensinando ao computador as coisas que todos nós sabemos

Resultado do Desafio OMCS Natal 2007 ^{NOVO!}
O que é Open Mind Common Sense?
Junte-se a nossa comunidade no Orkut

Se você já é cadastrado, por favor, identifique-se e comece a ensinar!

Login: Senha:

(Esqueceu a senha?)

[Clique aqui para cadastrar-se no Open Mind Common Sense!](#)

O Open Mind Commonsense no Brasil (OMCS-Br) é um projeto do Laboratório de Interação Avançada (LIA) da UFSCar, em colaboração com o MediaLab do Massachusetts Institute of Technology (MIT), para a coleta de conhecimento das pessoas, chamado de senso comum, visando desenvolver aplicações computacionais que possam usar tal conhecimento e se tornar mais úteis às pessoas. Este trabalho é destinado exclusivamente à pesquisa e não tem fins lucrativos. Convidamos todos a participar do projeto.

Concluído

Internet | Modo Protegido: Ativado 100%

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
 - 
 - 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

Comissão Especial de Processamento de Linguagem Natural - Windows Internet Explorer

http://www.nil.icmc.usp.br/cepln/

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Google C Google

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 [XXVII Congresso da Sociedade Brasileira de Computação](#) (realizado no Rio de Janeiro-RJ em Junho/Julho de 2007) por pedido das 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 representavam 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 Línguas Naturais, 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 Auxílio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação, entre muitas outras, além das tarefas relacionadas de criação e disponibilização de dicionários léxicos e corpús eletrônicos, desenvolvimento de taxonomias e ontologias, investigações em linguística de corpús, desenvolvimento de esquemas de marcação e anotações 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 gerenciando listas e fóruns de discussão, dentre outras medidas.

Internet | Modo Protegido: Ativado 100%

STIL 2009, 8-11 Setembro... x

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 \(WTI\)\)](#)
- [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.nllc.icmc.usp.br/tli

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/

Arquivo Editar Exibir Favoritos Ferramentas Ajuda

Google OK 42 bloqueado Verificar Enviar para Configurações

propor2008.org - International Conference on Co...



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-annual 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).

Internet | Modo Protegido: Ativado 100%

the Association for Com... x
http://aclweb.org/

Friday, 13 August 2010

The Association for Computational Linguistics

Google Custom Search Search

Home

MAIN MENU

- Home
- News
- Conferences
- Membership
- Publications
- CL Journal
- Resources
- Affiliations
- SIGs
- About the ACL
- Contact Us
- ACL Policies

Bill Woods Receives 2010 ACL Lifetime Achievement Award

ACL is proud to announce that Bill Woods has received the 2010 ACL Lifetime Achievement Award. A list of previous recipients can be viewed at the [ACL Wiki](#).

ACL 2012 **ACL HLT 2011**

ACL 2012 will be held in Jeju Island, Korea. ACL HLT 2011 will be held in Portland, Oregon, USA, June 19-24, 2011. For more details, visit the conference website at <http://www.acl2012.org/>

ACL HLT 2011 will be held in Portland, Oregon, USA, June 19-24, 2011. For more details, visit the conference website at www.acl2011.org

IJCNLP 2011 **ACL 2010 Registration Now Open**

IJCNLP 2011 will be held in Chiang Mai, Thailand November 7-13, 2011 (Main conference: November 9-11)

Registration for the ACL 2010 Conference is now open. See www.acl2010.org for more details.

More information to be announced soon.

ACL Anthology Network **Former ACL president Jun'ichi Tsujii received Medal of Honor from Government of Japan**

The ACL Anthology Network is online and available at <http://aan.eecs.umich.edu>

Former ACL president Jun'ichi Tsujii has been named a recipient of the Medal of Honor with Purple Ribbon by the Government of Japan. These medals are given to distinguished scientists, artists, and athletes.

<http://www-tsujii.is.s.u-tokyo.ac.jp/>

More...

Fred Jelinek Receives 2009 ACL Lifetime Achievement Award

<< Start < Prev 1 2 3 Next > End >>

Results 1 - 8 of 24

POLLS

What do you think of the new site?

Love it

Hate it

This is a new site?

Vote Results

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.

[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

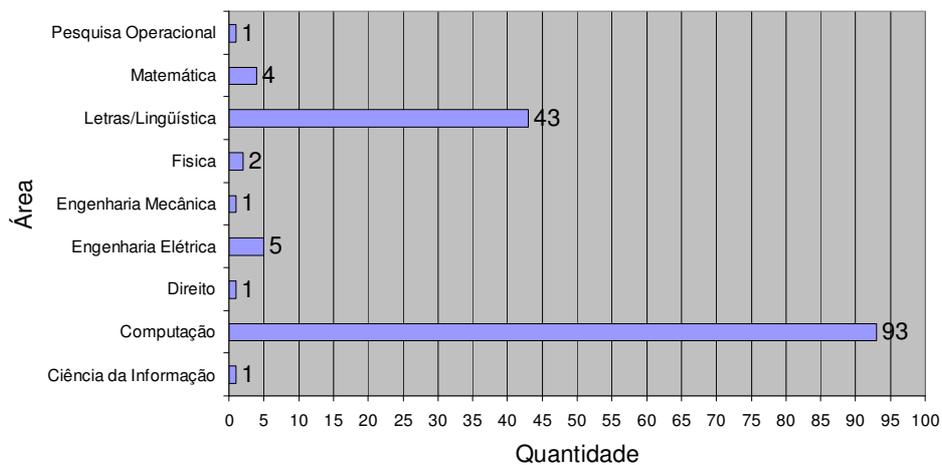
[PLN no Brasil]

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

33

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

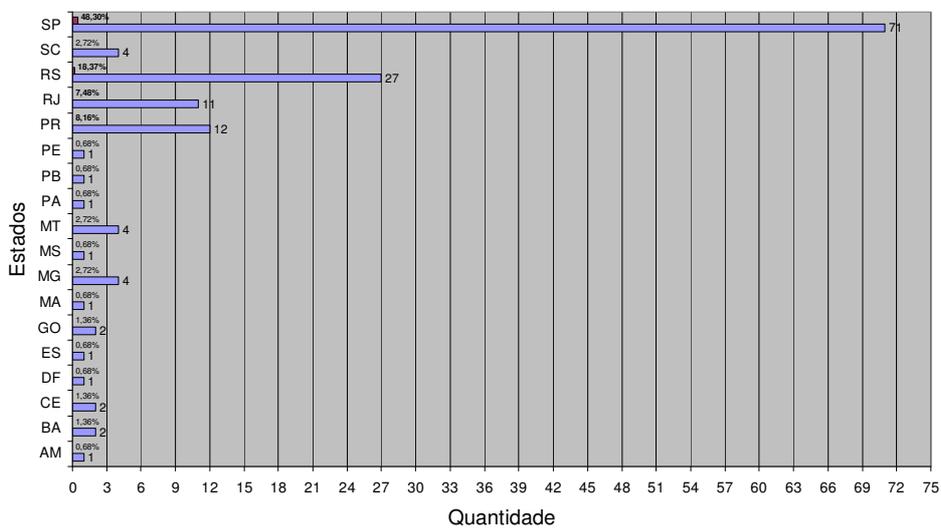
Área de formação



PLN no Brasil

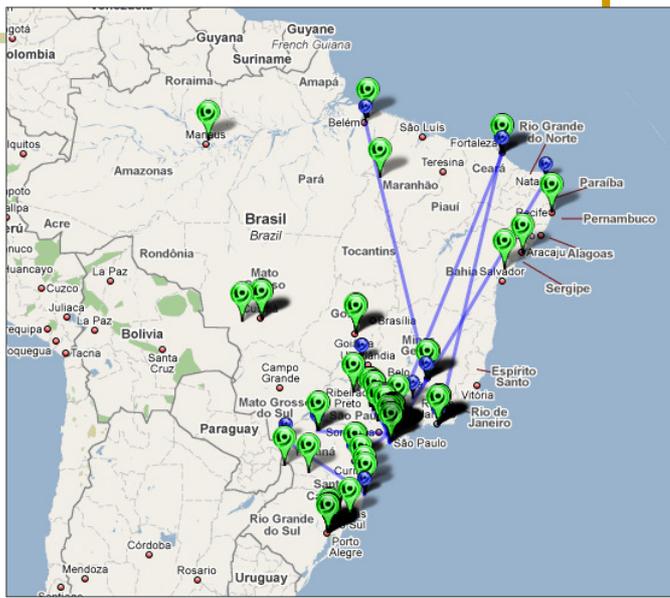
Pardo et al. (2009)

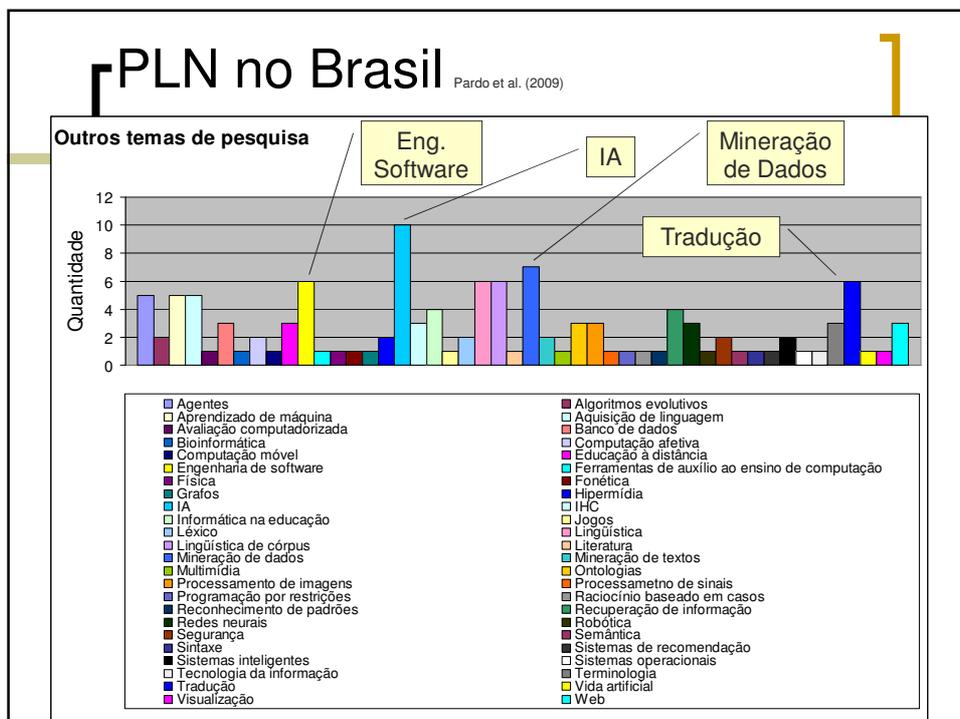
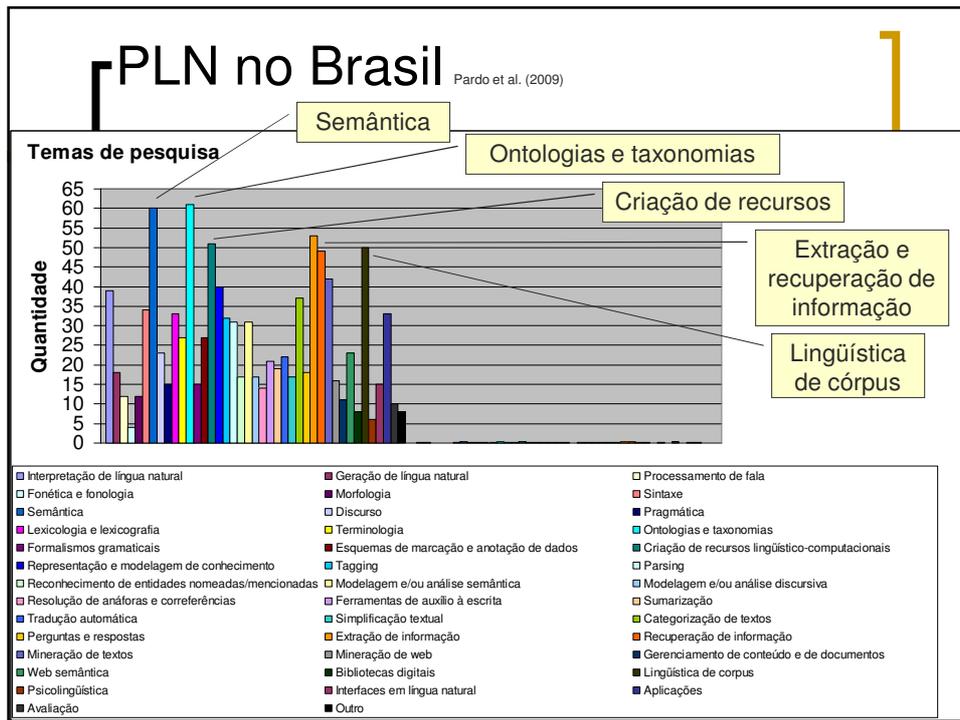
Distribuição de pesquisadores por estado



PLN no Brasil

Pardo et al. (2009)

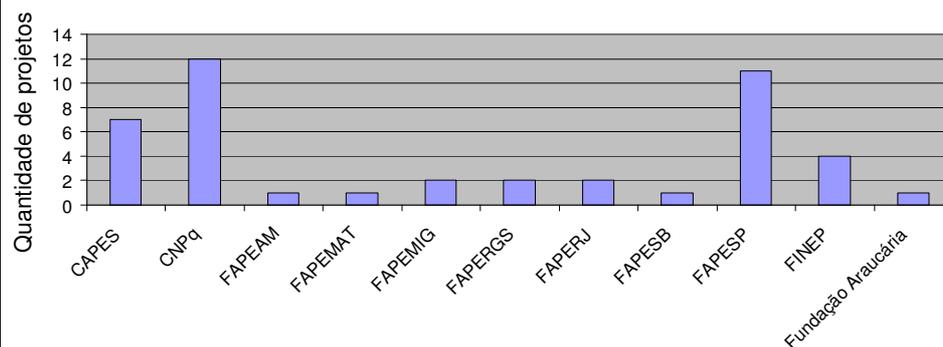




PLN no Brasil

Pardo et al. (2009)

Fontes de financiamento



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 (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 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 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 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
<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 em anais</i>	1603	6264	0,26
<i>Resumos expandidos em anais</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