
# Natural Language Processing: Dealing with unstructured data

Renata Vieira and Joaquim Neto

FGV Workshop on Data Science
2019

FACIN PUCRS

# PLN

Pensamos e falamos

Colocamos os pensamentos no mundo

Escrevemos, armazenamos e compartilhamos

Criamos mais sobre o que pensar

E há muito para ler

Pensamos sobre a maneira como pensamos e falamos

E assim construimos maquinas para nos ajudar a comunicar

E mudamos nossa forma de pensar, de falar e de comunicar

FACIN PUCRS

# Roteiro – dia 1

- Parte 1
  - Complexidade da língua natural
  - Conhecimento linguístico
  - Conhecimento de mundo
  - Tarefas de PLN
  - Língua Portuguesa

- Parte 2
  - Modelos estatísticos da língua
  - Word Embeddings

# Roteiro – dia 2

- Parte 1
  - Relações entre ontologias e PLN
  - Pesquisas desenvolvidas nessas áreas

- Parte 2
  - Prática
  - Tarefa: Reconhecimento de entidades nomeadas

# Parte 1

- Complexidade da língua natural
- Conhecimento linguístico
- Conhecimento de mundo
- Tarefas de PLN
- Língua Portuguesa

FACIN PUCRS

# Língua natural

# Língua natural

Estão abertas as candidaturas a **subsídios** para a participação na Wikimania 2008.

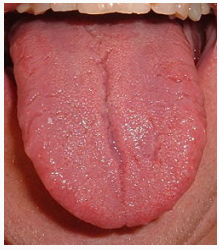
# Língua

Origem: Wikipédia, a enciclopédia livre.

*Esta é uma página de desambiguação, que lista artigos associados a um mesmo título.*
*Se uma ligação interna o conduziu até aqui, sugerimos que a corrija para que aponte diretamente para o artigo adequado.*

- Língua (anatomia) - órgão muscular localizado na parte ventral da boca;
- Língua natural - idioma falada por seres humanos;
- Língua de escolarização - língua em que se aprende a ler e a escrever;
- Linguagem - sistema de signos que serve de meio de comunicação.
- Língua materna - primeira língua que uma criança aprende.

Categoria: Desambiguação




WIKIPÉDIA
A enciclopédia livre

navegação
- Página principal
- Os melhores artigos
- Eventos atuais
- Página aleatória
- Portais

colaboração
- Portal comunitário
- Mudanças recentes
- Ajuda
- Donativos

busca
Ir | Pesquisa

ferramentas
- Artigos afluentes
- Novidades relacionadas
- Carregar arquivo
- Páginas especiais

# A língua natural é ambígua

- Diferentemente das linguagens artificiais
  - Java
  - C
  - HTML
  - Lógica

FACIN PUCRS

# Comunicação humana e ambiguidade

- Mas como a comunicação humana é eficiente?
  - (pelo menos, às vezes)

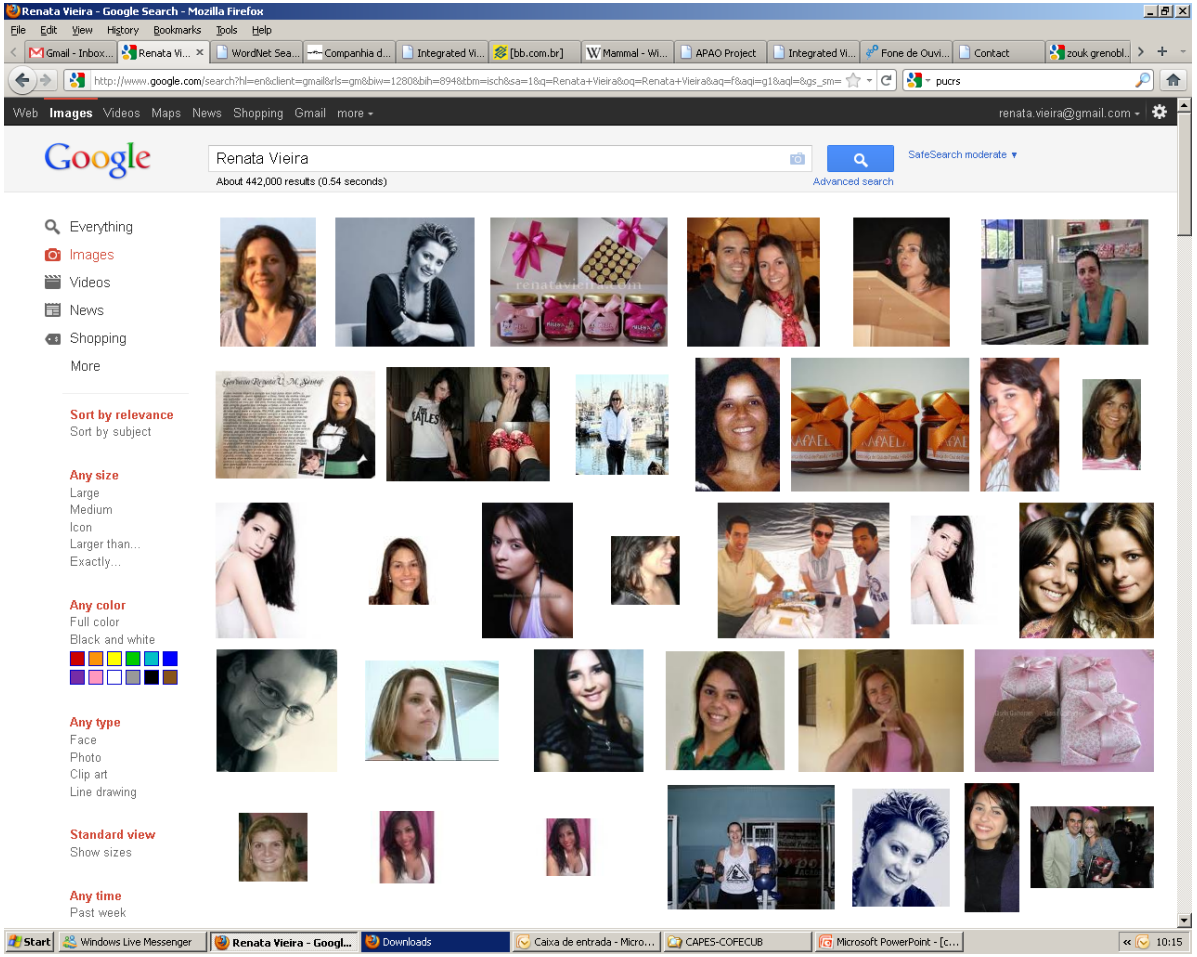
- Associações
  - Fala/texto e visão
  - Fala/texto e conhecimento prévio
  - Sistema altamente contextualizado
    - Dito e o não dito

# Vamos por partes: nomes, termos, conceitos, significados e outras palavras...

# Nomes

- Nomes e seus referentes
  - Entidades nomeadas
    - Brasil
      - o país
      - o terrritório
      - time de futebol
    - Renata Vieira
      - pessoa

# Renata Vieira



# Nomes, **termos**

- Palavras simples e compostas geralmente usadas em contextos específicos
- Termos técnicos usados por exemplo em ciências ou artes especificas
- Terminologia: disciplina lingüística para o estudo científico dos **conceitos** e **termos** usados nas línguas de especialidade".

# Nomes, termos, **conceitos**

- Conceito
  - compreensão que alguém tem de uma palavra; noção, concepção, ideia
  - O termo "**conceito**" tem origem a partir do latim “conceptus” (do verbo concipere) que significa "coisa concebida" ou "**formada na mente**"

# Nomes, termos, conceitos, **significados**

- **Significado de Significado**
  - Definição atribuída a um termo, palavra, frase, texto; aquilo que alguma coisa quer dizer; sentido.
  - Relevância que se dá a algo: sua participação teve muito significado

FACIN PUCRS

# Nomes, termos, conceitos, **significados**

- **Significado de Significado**
  - [Linguística] Significação; forma representativa e mental que se relaciona com a forma linguística; o que o signo quer significar; a parte do signo linguístico definida pelo conceito.

# Nomes, termos, conceitos, significados e **outras palavras**

- Verbos, adjetivos, preposições ...
- **Artigos**

# O artigo definido “the”

- B Russell – Mind, 1905.
  - On denoting

- R Vieira - PhD Thesis, 1998
  - Definite Description Resolution in Unrestricted Texts

FACIN PUCRS

# O artigo definido “the”

- B Abbott - Proceedings of SALT, 2011
  - Support for a Unique Theory of Definiteness
    - There is undoubtedly more to be said on this subject, which seems to continue to excite at least some people today as much as it did Russell at the turn of the century.
- J Cho, International Journal of Bilingualism 2017
    - The acquisition of different types of definite noun phrases in L2-English

FACIN PUCRS

# Em outras palavras ...

- Somos naturalmente aptos pra lidar com palavras, por que ao estudá-las elas começam a parecer tão complicadas?

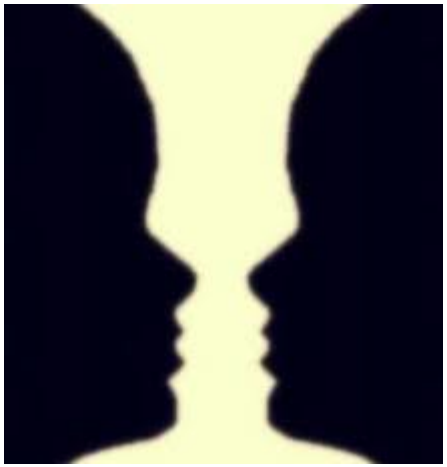
# O Sistema Simbólico

- Linguagem:
  - Convenções, associações
  - Nomes próprios
    - Renata Vieira
  - Designadores descritivos
    - A palestrante da tarde de hoje (categorias e propriedades)

# Mente

- Conceito
  - "coisa concebida" ou "formada na mente"

# A mente e a cunhagem simbólica

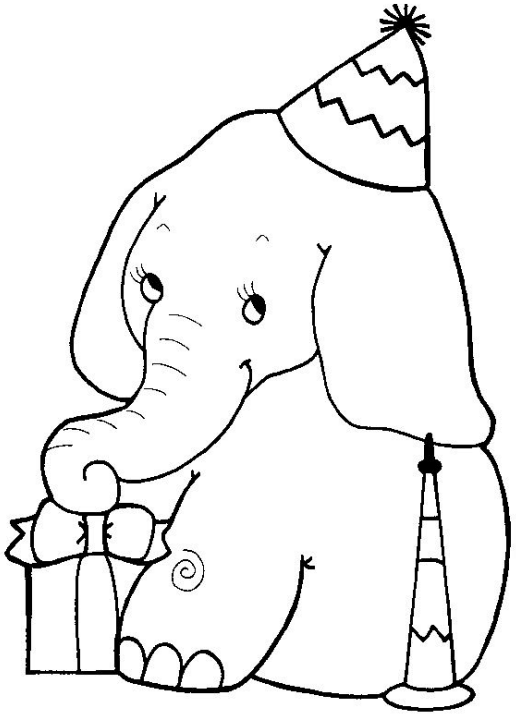


# A mente e a cunhagem simbólica

- El___ant

# A mente e a cunhagem simbólica

- El___ant



# A mente dos outros

- O sucesso da comunicação está na habilidade de reconhecer a intenção comunicativa do falante e inferir sua intenção informativa

# Palavras

- Apenas algumas razões das dificuldades

- Outras razões:
  - Mentes
  - Frases, Textos, Posts
  - Sintaxe, semântica e pragmática

# Além das palavras

- Frases
  - Parsing ou análise sintática
- Identificar os constituintes de uma frase, verificando as suas partes de acordo com a gramática da língua
  - Sujeito verbo objeto
    - A língua é o único músculo voluntário do corpo humano que não fadiga.

FACIN PUCRS

# Tree structure

Enter text to parse:

A língua é o único músculo voluntário do corpo humano que não fadiga.

Parse and Show

Export and Download

Reset

Visualization: Slant (applet) Notational convention VISL - default

Applet is now running in a separate window.

VISL - Visual Interactive Syntax Learning

File Symbols Display Extras Language Settings Tools Help

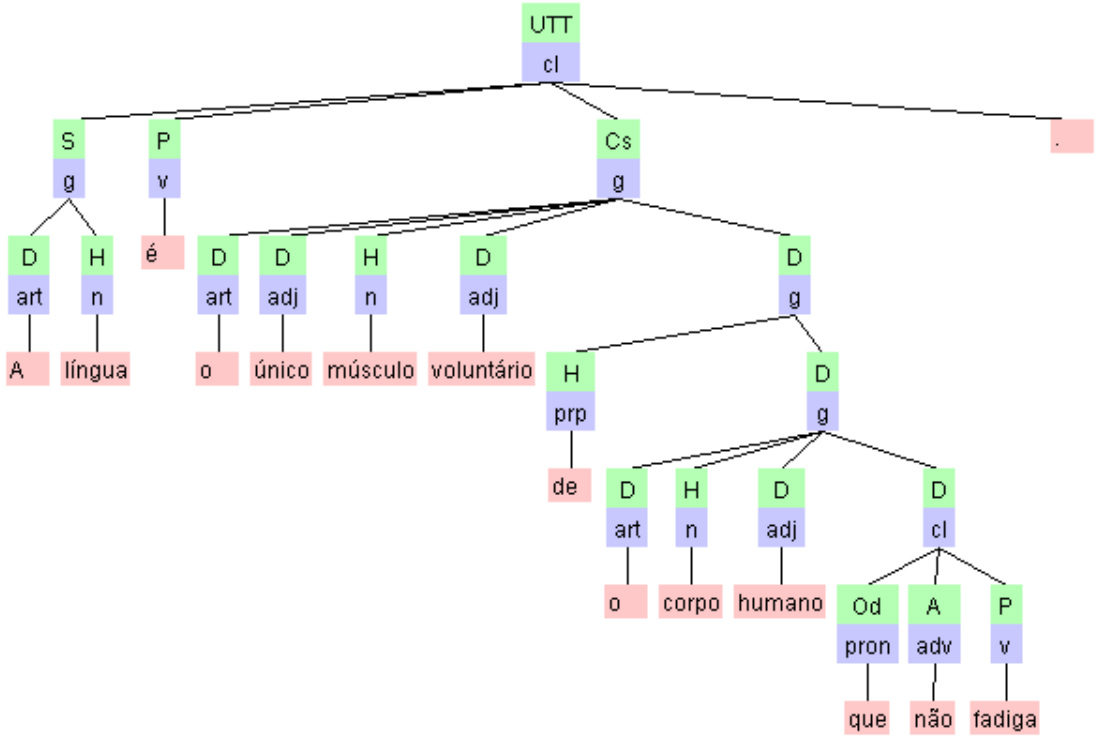
Sentence: A língua é o único músculo voluntário de o corpo humano que não fadiga .

Function: S P Od Oi Op Cs Co fC A SUB CO CJT D H UTT STA QUE COM EXC X

Form: n prop v adj adv art pron prp conj num intj cl par g x

Collapse Tree

Expand Tree




Analysis 1 of 1

Java Applet Window

Iniciar VISL ... 2008 http:/... Erbas... pln-er... apres... Palest... Palest... VISL ... 13:42

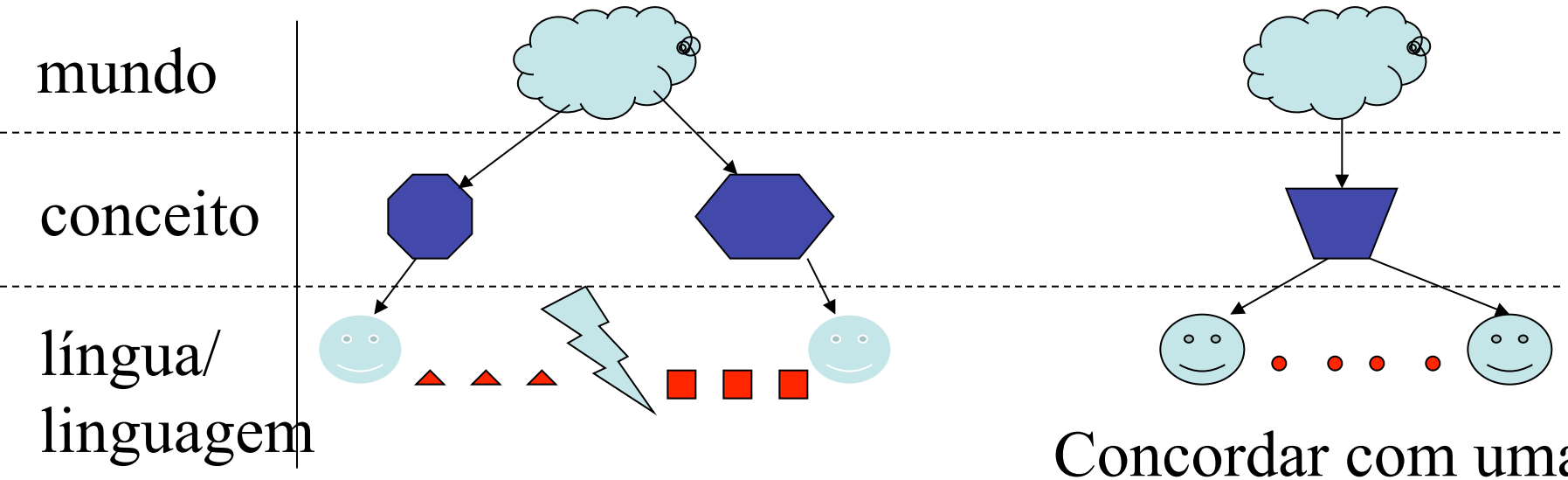
# Sintaxe, Semântica, Pragmática

- Forma
- Conteúdo/significado
- Uso
- Conhecimento de mundo

# Semântica

- Significado
  - Thesauros
  - Dicionários
  - Ontologias

# Ontologias




Desentendimento

Confusão conceitual e terminológica

Concordar com uma conceitualização

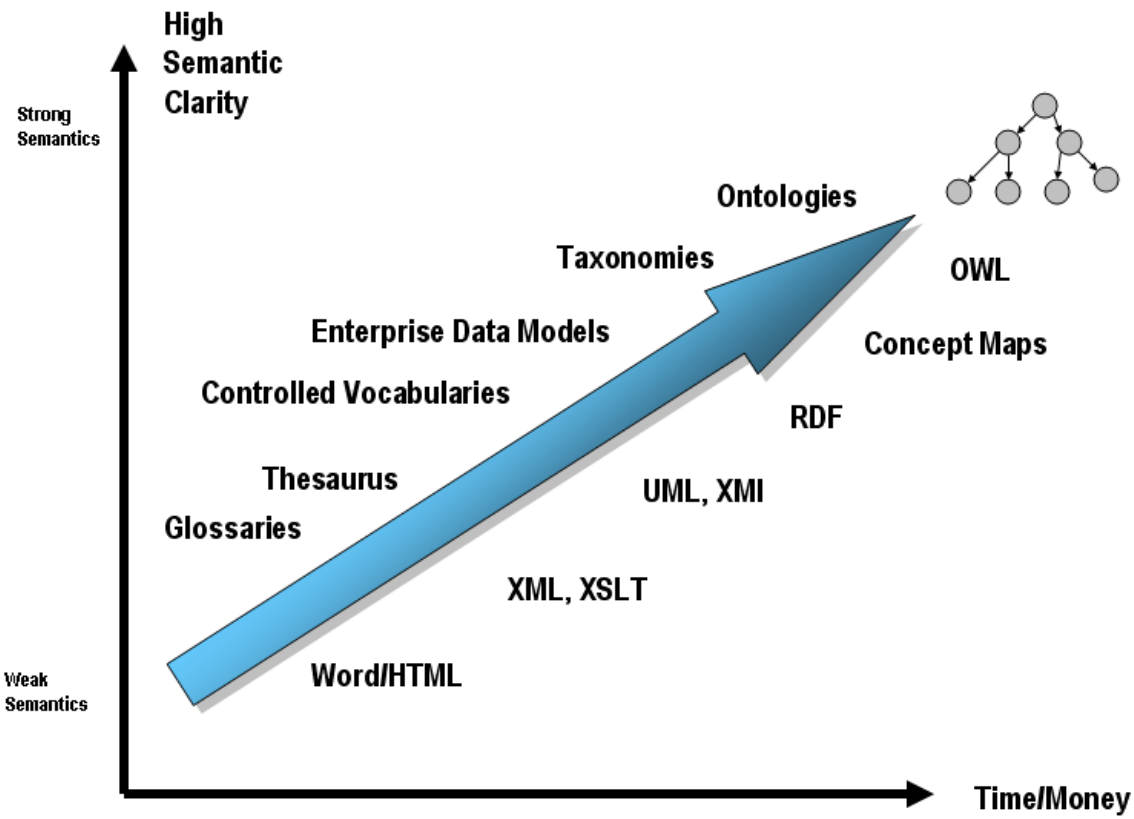
Torná-la explícita em alguma linguagem

**Atores: humanos e máquinas**

FACIN PUCRS

# Ontologias

- Semântica explícita
- Significado estruturado
- Estrutura lógica



# Ontologias

- Conceito (classe)
  - Extensão x Intensão
  - Conjunto de referentes x conjunto de propriedades
  - Hierarquia de conceitos
  - Relacionamentos

# Ontologias

- Relacionamentos
  - É_um (hierarquia)
  - Parte de
  - Same_as

# Ontologia da sala

- Conceitos
- cadeira, mesa, porta, auditório, lustres, janelas, paredes
- Hierarquia
- Auditório é um tipo de sala
- Mesa é um tipo de móvel

# Ontologia da sala

- Conceitos
  - cadeira, mesa, porta, auditório, lustres, janelas, paredes, teto e piso
- Relacionamentos
  - janelas situam-se nas paredes
  - cadeiras ficam sobre o piso
- Axiomas
  - cadeiras tem 4 pés

FACIN PUCRS

# Ontologias de topo

- Objetos físicos
- Objetos abstratos

- Continuantes
- Perdurantes

# Outros recursos semânticos

- Bases semânticas
  - WordNet
  - ConceptNet
  - BabelNet

# WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: language Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

## Noun

- S: (n) **language**, linguistic communication (a systematic means of communicating by the use of sounds or conventional symbols) *"he taught foreign languages"; "the language introduced is standard throughout the text"; "the speed with which a program can be executed depends on the language in which it is written"*
- S: (n) speech, speech communication, spoken communication, spoken language, **language**, voice communication, oral communication ((language) communication by word of mouth) *"his speech was garbled"; "he uttered harsh language"; "he recorded the spoken language of the streets"*
- S: (n) lyric, words, **language** (the text of a popular song or musical-comedy number) *"his compositions always started with the lyrics"; "he wrote both words and music"; "the song uses colloquial language"*
- S: (n) linguistic process, **language** (the cognitive processes involved in producing and understanding linguistic communication) *"he didn't have the language to express his feelings"*
- S: (n) **language**, speech (the mental faculty or power of vocal communication) *"language sets homo sapiens apart from all other animals"*
- S: (n) terminology, nomenclature, **language** (a system of words used to name things in a particular discipline) *"legal terminology"; "biological nomenclature"; "the language of sociology"*

WordNet home page

Visuwords™ online graphical dictionary and thesaurus - Mozilla Firefox

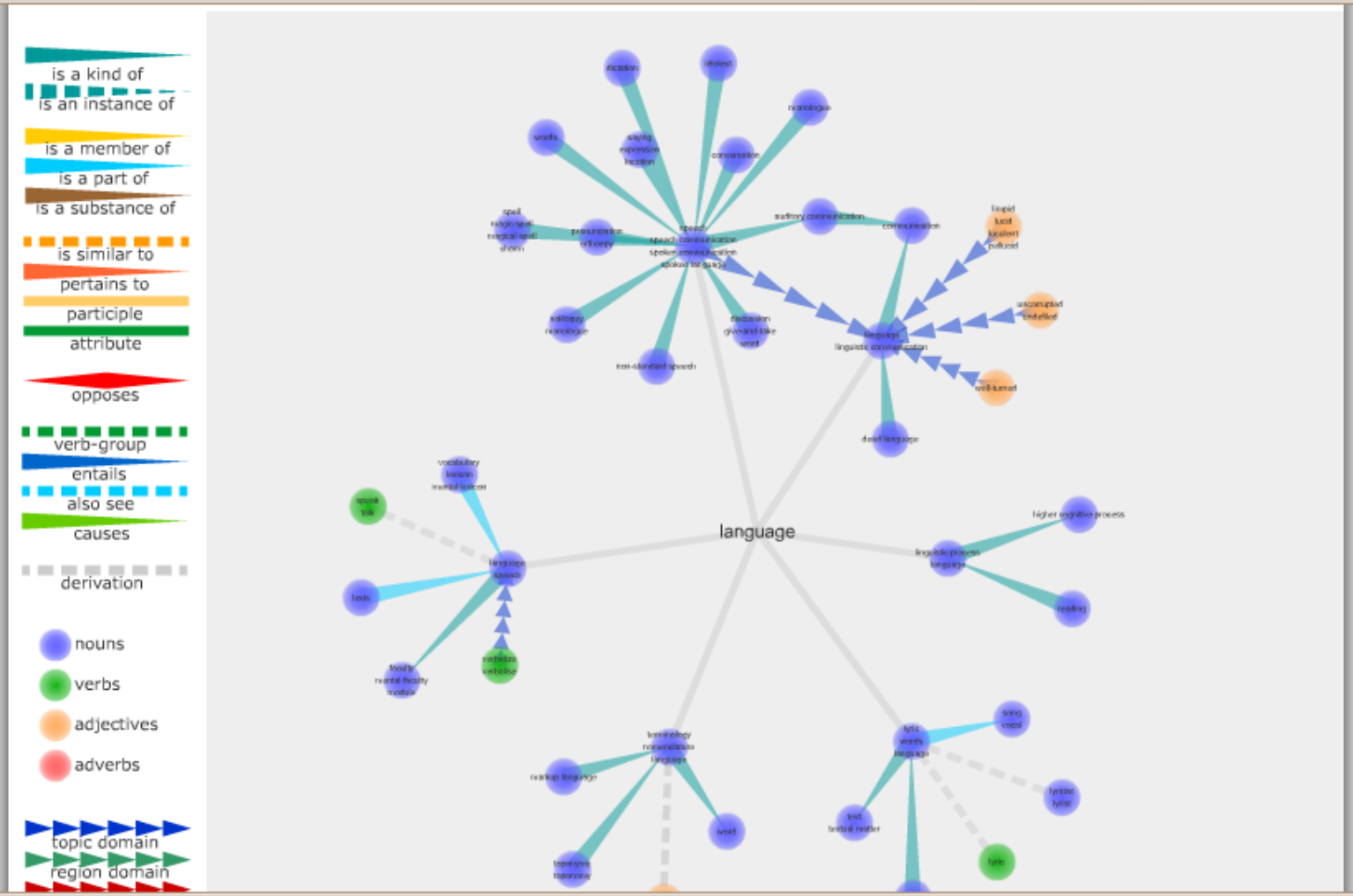
Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda

http://www.visuwords.com/?word=language

age is spoken in brazil?

Search Web

MAS 2008 | the Asso... | propor2008... | WordNet... | Powerset B... | Microsoft K... | wordviz ... | Visu... | languag... | o que a ...




Localizar: prt | Próxima | Diferenciar maiúsc./minúsc. | Atingido o fim da página, continuando do início

Concluído | 1.859s | Images

Iniciar | Visuwords™ o... | IA-07 | http://mail.goog... | Erbase2008.ppt | pln-erbase2008l... | 11:42

# Tarefas PLN

– Alguns exemplos...
  - Tradutores
  - Corretores ortográficos
  - Sumarizadores
  - Geração automática de recursos linguísticos e semânticos (dicionários, ontologias)

FACIN PUCRS

## Traduzir texto

Texto original:

In the philosophy of language, a natural language (or ordinary language) is a language that is spoken, written, or signed by humans for general-purpose communication, as distinguished from formal languages (such as computer-programming languages or the "languages" used in the study of formal logic, especially mathematical logic) and from constructed languages.

inglês para português | Traduzir

Texto traduzido automaticamente:

Na filosofia da linguagem, a linguagem natural (ou linguagem comum) é uma língua que é falada, escrita, assinada por seres humanos ou para fins gerais de comunicação, como distinguir de linguagens formais (como as linguagens de programação de computador ou os "línguas" Utilizados no estudo da lógica formal, especialmente lógica matemática) e construída a partir de línguas.

Sugira uma tradução melhor

## Traduzir uma página da web

http:// | inglês para português | Traduzir

# (2008)

- Na filosofia da linguagem, a linguagem natural (ou linguagem comum) é uma língua que é falada, escrita, **assinada** por seres humanos **ou** para fins gerais de comunicação, **como distinguir de linguagens formais** (como as linguagens de programação de computador ou os **"línguas" Utilizados** no estudo da lógica formal, especialmente lógica matemática) e **construída a partir de línguas**.

- Na filosofia da linguagem, a linguagem natural (ou linguagem comum) é uma língua que é falada, escrita, **ou gestualizada** por seres humanos  para fins gerais de comunicação, **e são distintas das linguagens formais** (como as linguagens de programação de computador ou **as linguagens utilizadas** no estudo da lógica formal, especialmente lógica matemática) **e das linguagens construídas**. (Esperanto)

FACIN PUCRS

# Está evoluindo (2018)

- Na filosofia da linguagem, **uma** linguagem natural (ou linguagem comum) é uma **linguagem** falada, escrita ou **assinada** por seres humanos para comunicação de propósito geral, **como distinta** das linguagens formais (como linguagens de programação de computador ou "línguas" usadas no estudo da lógica formal, especialmente a lógica matemática) e das linguagens construídas.

- Na filosofia da linguagem, **a** linguagem natural (ou linguagem comum) é uma **língua que** é falada, escrita, ou **gestualizada** por seres humanos  para fins gerais de comunicação, **e são distintas** das linguagens formais (como as linguagens de programação de computador ou as linguagens utilizadas no estudo da lógica formal, especialmente lógica matemática) e das linguagens construídas. (esperanto)

FACIN PUCRS

# Extração de informação de textos

- Reconhecimento de entidades nomeadas
  - Reconhecer a ocorrência de nomes próprios e suas categorias
    - pessoas, organizações, locais, etc...
- Reconhecimento de relações entre entidades
  - Fusões/aquisições de empresas
  - Relações pessoas organizações
  - Etc...
- Análise de sentimentos

FACIN PUCRS

# Resolução de correferência textual

- Pesquisadores do Museu Nacional do Rio de Janeiro anunciaram a descoberta de **uma nova espécie de dinossauro** no Brasil. **O animal** que na cadeia evolutiva dos dinossauros ocuparia uma posição no grupo Tyrannoraptora, o mesmo do Tyrannossaurus rex, habitou o nordeste brasileiro há 110 milhões de anos, no período Cretáceo. **O fóssil, batizado de Santanaraptor placidus**, é **o único a ser encontrado no país com tecidos preservados**. Isso pode permitir que os cientistas saibam mais sobre o modo de vida e a evolução dos répteis.

VISL - Visual Interactive Syntax Learning

File Symbols Display Extras Language Settings Tools Help

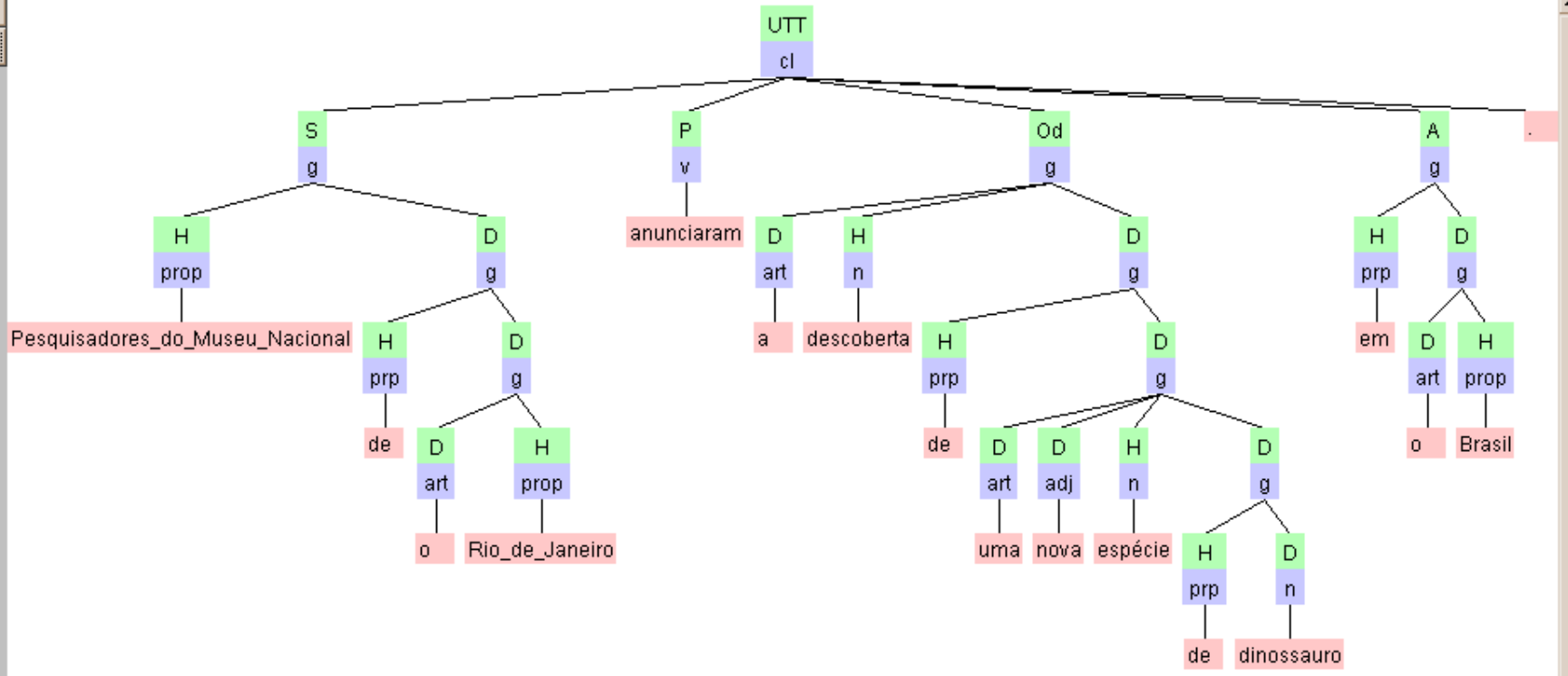
Sentence: Pesquisadores_do_Museu_Nacional de o Rio_de_Janeiro anunciaram a descoberta de uma nova espécie de dinossauro em o Brasil .

Function: S P Od Oi Op Cs Co fC A SUB CO CJT D H UTT STA QUE COM EXC X

Form: n prop v adj adv art pron prp conj num intj cl par g x

Collapse Tree

Expand Tree




Analysis 1 of 1

Java Applet Window

Iniciar VISL - Tre... http://mail... erbase Downloads pln-erbase... have-i-me... VISL - Vis... 18:36

# Relatorio Correferencia Member: CIENCIA_2000_17088.txt

| | Classificação | Sintagma |
|---|---|---|
| CADEIA : set_24 | | |
| word_1..word_4 | --- | **Pesquisadores de o Museu_Nacional_do_Rio_de_Janeiro** |
| word_79..word_80 | indirect---old | **os paleontólogos** |
| word_135..word_136 | indirect---old | **os cientistas** |
| word_301..word_302 | direct---old | **os pesquisadores** |
| word_342..word_343 | direct---old | **os paleontólogos** |
| CADEIA : set_28 | | |
| word_10..word_14 | --- | **uma nova espécie de dinossauro** |
| word_19..word_20 | indirect---old | **O animal** |
| word_49..word_53 | indirect---old | **Batizado_de Santanaraptor_placidus , o fóssil** |
| word_96..word_97 | indirect---old | **o dinossauro** |
| word_117..word_118 | direct---old | **o fóssil** |
| word_164..word_165 | indirect---old | **o Santanaraptor** |
| word_199..word_200 | direct---old | **o animal** |
| word_231..word_232 | direct---old | **O Santanaraptor** |
| word_255..word_263 | indirect---old | **O exemplar de Santanaraptor encontrado por a equipe carioca** |
| word_273..word_274 | direct---old | **o fóssil** |
| word_286 | --- | **ele** |
| word_306..word_307 | indirect---old | **o bicho** |
| word_354 | --- | **ele** |
| CADEIA : set_13 | | |

# Resolução de correferência

- Aplicações das cadeias de correferência?
  - Sumarização

- Sumário:
  - Segundo ele, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas.

| | | |
|---|---|---|
| CADEIA : set_10 | | |
| word_8..word_9 | ---new | **a internet** |
| word_126..word_127 | direct---old | **a internet** |
| CADEIA : set_3 | | |
| word_13..word_17 | ---new | **os contatos entre as pessoas** |
| word_20 | --- | **os** |
| word_62..word_66 | indirect---old | **a socialização de as pessoas** |
| CADEIA : set_4 | | |
| word_16..word_17 | ---new | **as pessoas** |
| word_65..word_66 | direct---old | **as pessoas** |
| CADEIA : set_14 | | |
| word_32..word_45 | --- | **Barry_Ellman , de o Centro para Estudos_Urbanos e Comunitários de a Universidade_de_Toronto , Canadá** |
| word_48..word_49 | indirect---old | **o pesquisador** |
| word_73 | --- | **ele** |
| word_134 | --- | **Ellman** |
| word_137 | --- | **ele** |
| word_159..word_160 | indirect---old | **o pesquisador** |
| CADEIA : set_11 | | |
| word_56 | --- | **computadores** |
| word_142 | --- | **computadores** |
| CADEIA : set_9 | | |
| word_89..word_95 | ---associative | **as pessoas plugadas em uma rede local** |
| word_108..word_109 | --- | **vizinhos conectados** |
| CADEIA : set_15 | | |

# Correferência e sumarização

– Resultado:

- Segungo Barry Ellman, do Centro para Estudos Urbanos e Comunitários da Universidade de Toronto, Canadá, pessoas ligadas por computadores tiveram mais contatos pessoais com seus amigos e parentes do que pessoas não-conectadas.

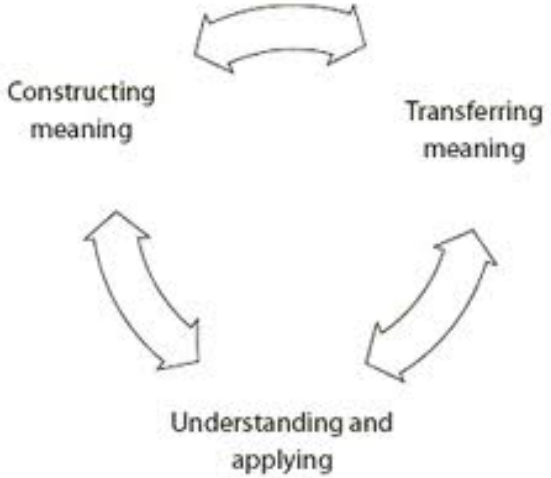
FACIN PUCRS

# Diálogos

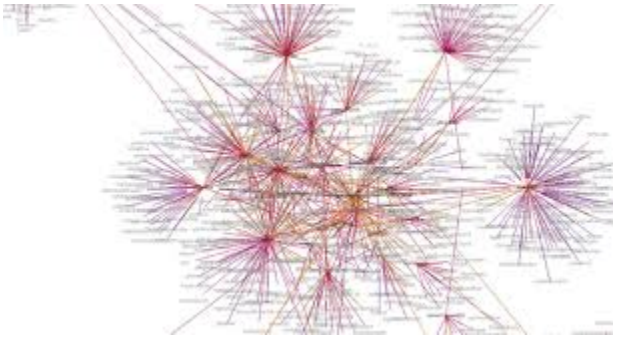
- Sistemas de perguntas e respostas
- Chat bots
- Comunicação Humano-Robô

FACIN PUCRS

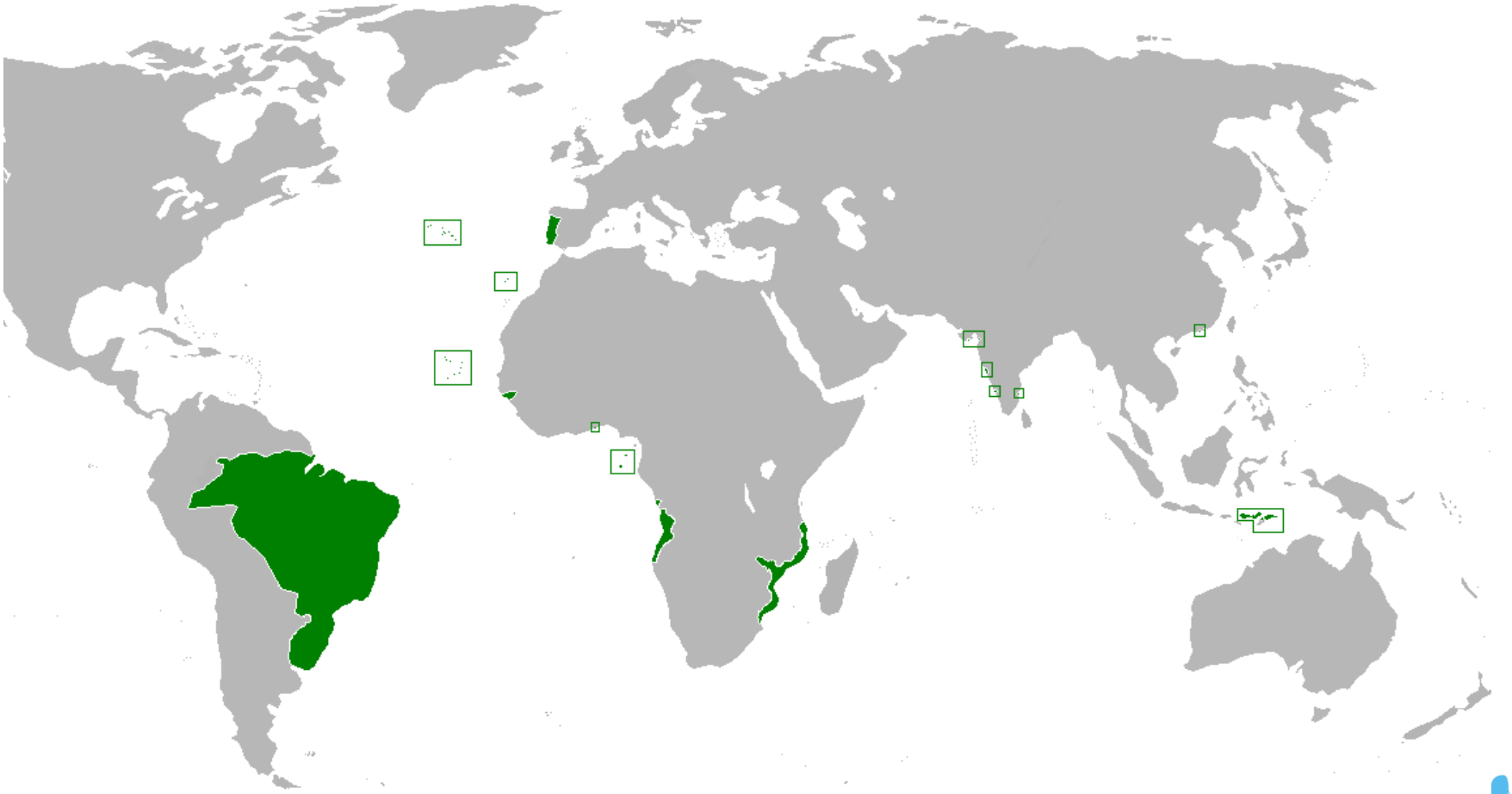
# Muitos desafios





# PLN: e a língua portuguesa?



# PLN no Brasil

- Necessidade de estudos fundamentais
- Construção de recursos de base
  - Léxicos e bases de dados lexicais
  - Ontologias e vocabulários
  - Gramáticas e *parsers*
  - Grandes coleções anotadas
- PROPOR, STIL
  - Propor 2020 em Évora PT

FACIN PUCRS

# 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 Profas. 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 Lingüística 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 córpus eletrônicos, desenvolvimento de taxonomias e ontologias, investigações em lingüística de córpus, desenvolvimento de esquemas de marcação e anotação de conhecimento lingüístico-computacional, resolução anafórica, análise morfossintática 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 Lingüística 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.

# Part-Of-Speech and Parsers

**COLLOVINI, Sandra et al. Cross-Framework Evaluation for Portuguese POS Taggers and Parsers. In: 19th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing2018, 2018, Hanoi, Vietnam. 19th**

## POS Taggers

- TreeTagger (Schmid, 1994)
  - http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
- NLTK (Bird, 2006)
  - https://www.nltk.org/
- NLPNet (Fonseca and Rosa, 2013)
  - http://nilc.icmc.usp.br/nlpnet/
- UDPipe (Straka et al., 2016)
  - http://ufal.mff.cuni.cz/udpipe

FACIN PUCRS

# Parsers

- Palavras  (Bick, 2000)
  - http://visl.sdu.dk/visl/pt/
- MaltParser (Nivre et al., 2007)
  - http://www.maltparser.org/
- LX-Parser (Silva et al., 2010)
  - http://lxparser.di.fc.ul
- Freeling (Padró and Stanilovsky, 2012)
  - http://nlp.lsi.upc.edu/freeling/demo/demo.php
- CoGrOO (Silva, 2013)
  - http://cogroo.sourceforge.net/

# Bases Semânticas

Thiago Machado Lima. Analysing Semantic Resources for Coreference Resolution. In Master's Dissertation. PUCRS, 2019.

- **OntoPT**
  - http://ontopt.dei.uc.pt/
- **ContoPT**
  - http://ontopt.dei.uc.pt/index.php?sec=contopt
- **ConceptNet**
  - http://conceptnet.io/
- OpenWordNet-PT
  - https://github.com/own-pt/openWordnet-PT
- DBPedia
  - http://pt.dbpedia.org/
- BabelNet
  - https://babelnet.org/

# Outras Abordagens

- Vamos transformar tudo em números
- Abordagens estatísticas
- Observação da frequência dos vizinhos
  - diga me com quem andas que te direi quem és...
  - word embeddings