

# SCC0633/5908 PROCESSAMENTO DE LINGUAGEM NATURAL



# A BATALHA FINAL



# UMA CONVERSA ENTRE MÁQUINAS





# CARACTERÍSTICAS DE ULTRON

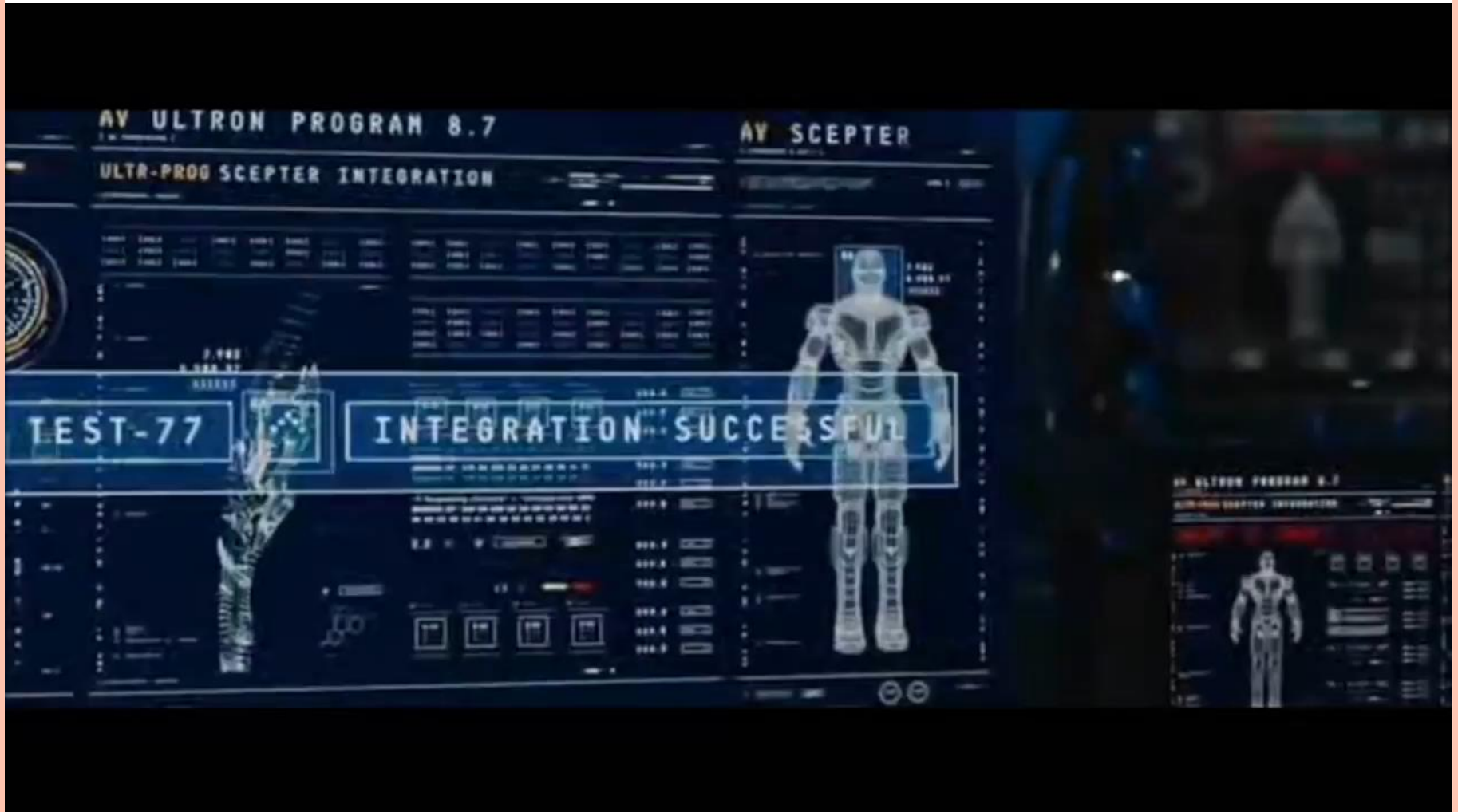
- Para variar, quer exterminar a humanidade, mas é uma máquina impressionante
  - Interpretação e geração de fala
  - Protocolos de diálogo
  - Medo da morte
  - Autopercepção
  - Um toque de religiosidade
  - Humor
  - Ironia
  - Expressão corporal
  - Criatividade
  - Conhecimento de mundo
  - Etc.

*Por onde começar construir algo com essas capacidades?*

Muito além do HAL!



# PARTE DA RESPOSTA



# PARTE DA RESPOSTA

- Capacidade inata de linguagem? Capacidade de aprendizado inata?
- Inteligência “real” requer um corpo?
- Web (córpus) como base de conhecimento?
- Aprendizado?
  - Dados comuns?
  - Dados representativos?
- Etc.

# LUGER (2013) EM SEU LIVRO CLÁSSICO SOBRE INTELIGÊNCIA ARTIFICIAL

- *O problema de definir o campo inteiro da inteligência artificial é semelhante ao de definir a própria inteligência: ela é uma única faculdade ou é apenas um nome para a coleção de capacidades distintas e não relacionadas? Até que ponto a inteligência é aprendida e não existe desde o nascimento? O que acontece exatamente quando ocorre o aprendizado? O que é criatividade? O que é intuição? A inteligência pode ser deduzida do comportamento observável ou ela requer evidências de um mecanismo interno em particular? Como o conhecimento é representado no tecido nervoso de um ser humano e que lições isso traz para o projeto de máquinas inteligentes? O que é autopercepção? Que papel ela desempenha na inteligência? Além disso, o conhecimento sobre a inteligência humana é necessário para construir um programa inteligente, ou uma técnica estritamente de “engenharia” é suficiente para tratar o problema? É possível conseguir inteligência em um computador, ou uma entidade inteligente requer a riqueza de sensações e experiências que só poderiam ser encontradas em uma existência biológica?*



# “Produtos” de PLN

*SCC5908 Introdução ao Processamento de Língua Natural*  
*SCC0633 Processamento de Linguagem Natural*

Thiago A. S. Pardo

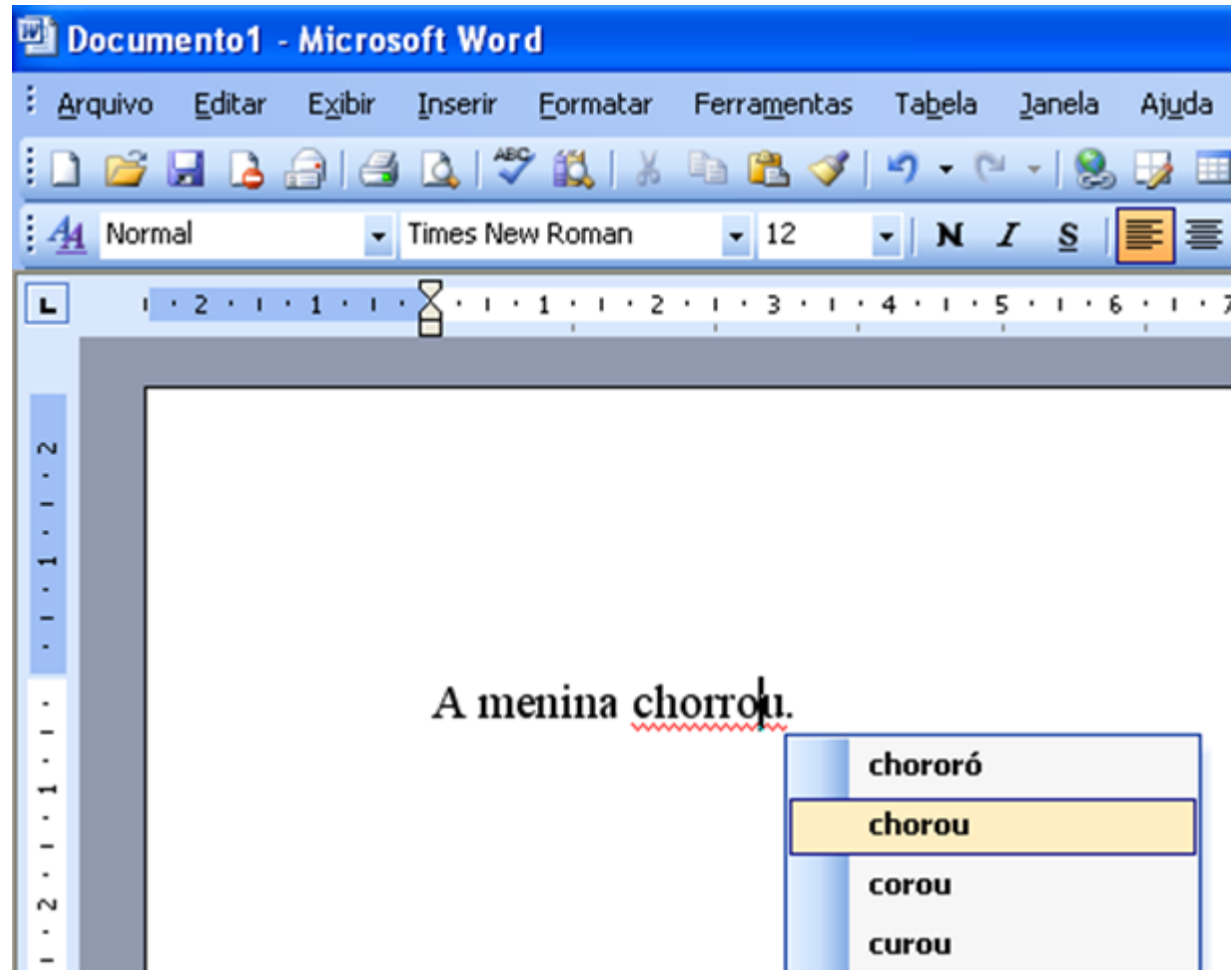


# Como desenvolver um produto de PLN?

- Diversas questões a considerar
  - Etapas do trabalho em PLN
  - Recursos e ferramentas
    - Material de base
  - Perfil de usuário
  - “Desenvolvedores”

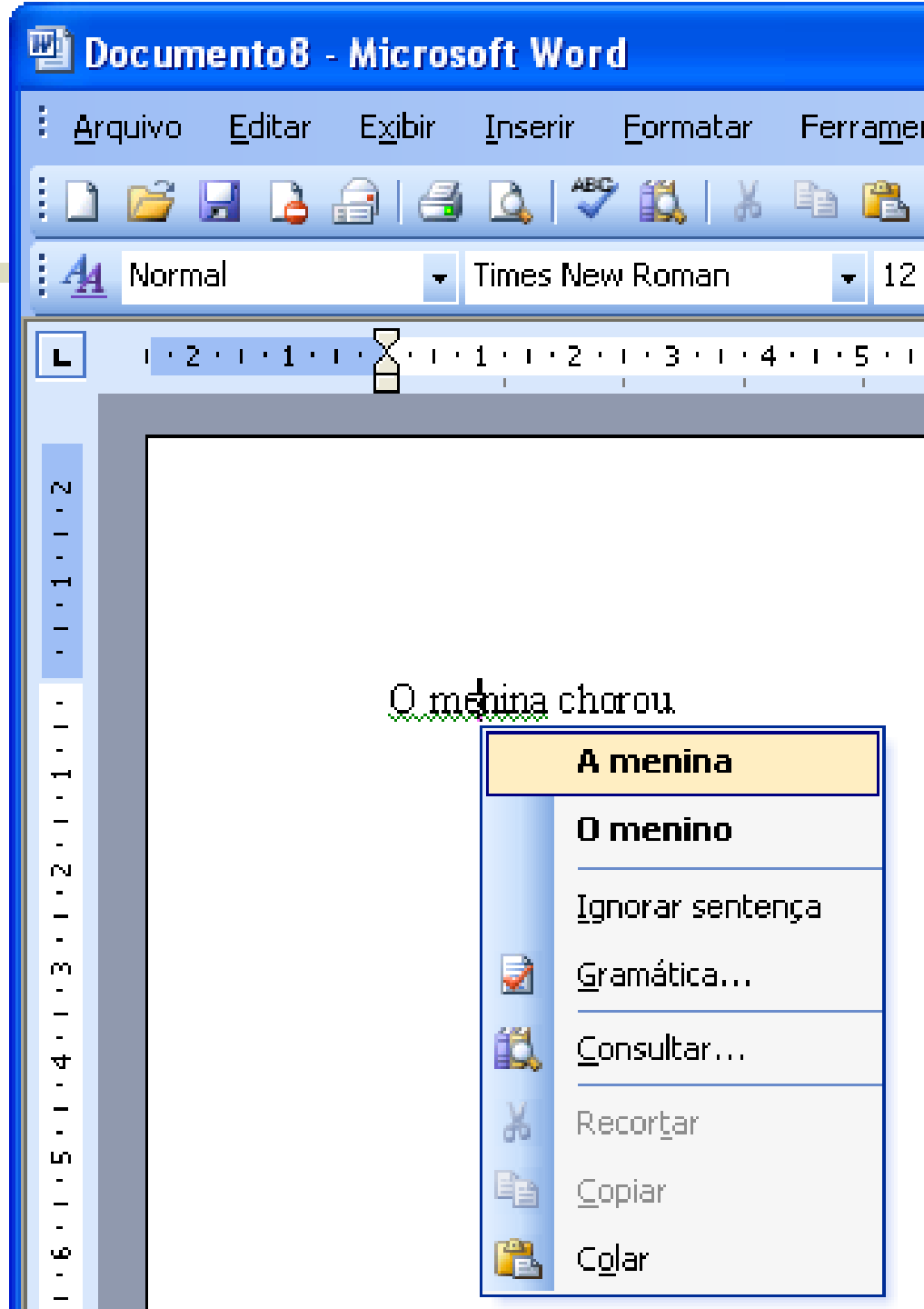
# Exemplos

- Revisão ortográfica
  - Tokenizador
  - Léxico
  - Proposta de sugestões (com base em ortografia, fonética, perfil de usuário, etc.)
  - Regras para ordenar sugestões



# Exemplos

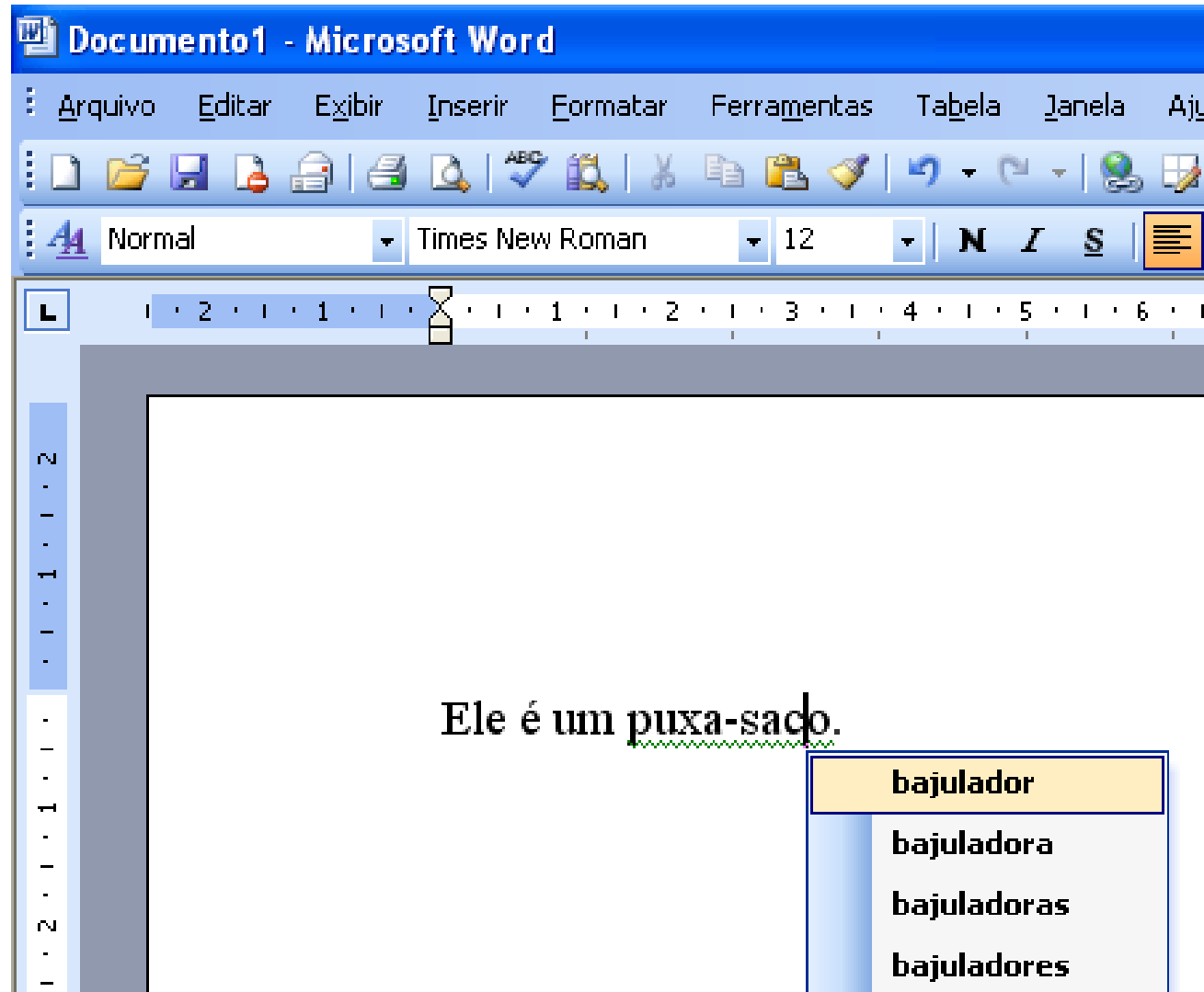
- Revisão gramatical
  - Tokenizador
  - Segmentador sentencial
  - Léxico
  - Etiquetador morfossintático
  - Analisador sintático
  - Desambiguação de análises
  - Regras gramaticais



# Exemplos

## ■ Revisão estilística

- Tokenizador
- Regras estilísticas e associação com perfil de usuário
- ...



# Exemplos

## ■ Sumarização automática

- Segmentação textual (nível?)
- 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 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.

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.

Um avião sérvio foi destruído no ataque. Foi escolhido um alvo isolado, para atingir os sérvios. Por causa do mau tempo as armas não foram devolvidas.

Os sérvios retiraram um tanque, dois caminhões e um canhão antiaéreo de um depósito da zona de exclusão.

O Exército sérvio tentou perseguir os sérvios, que estavam se retirando.

O Exército sérvio criou uma área de exclusão de 20 km em torno de Sarajevo.



# [ Exemplos ]

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

# SciPo

**Resumos** | **Introduções**

*Redação:* Recrear estrutura | Crítica automática

*Suporte:* Exemplos de estratégias | Exemplos de resumo | Marcadores discursivos

Ajuda  
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

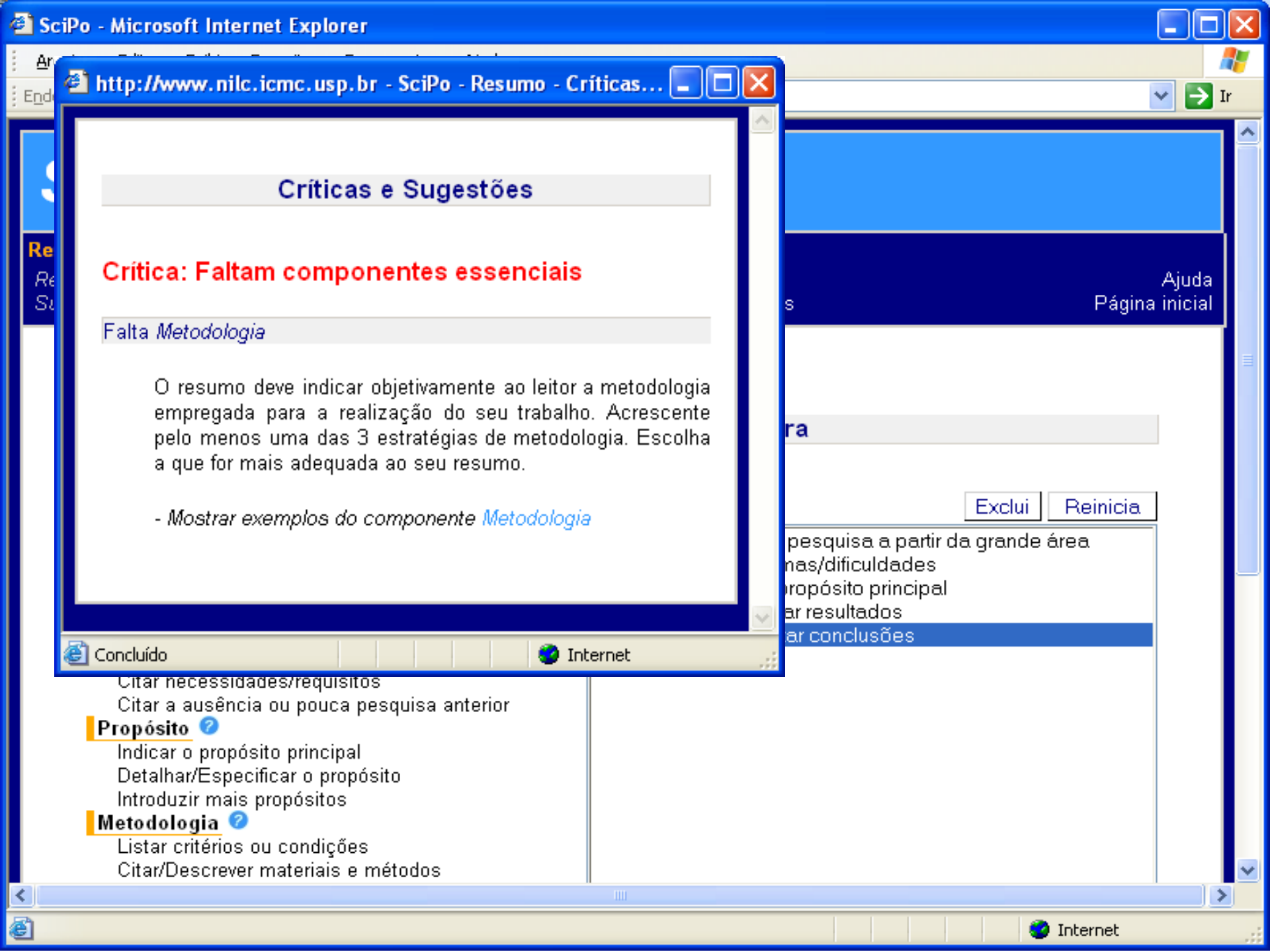
### Propósito ?

- 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**



### 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](#)*

Ajuda  
Página inicial

Exclui

Reinicia

- pesquisa a partir da grande área
- nas/dificuldades
- propósito principal
- ar resultados
- ar conclusões

Concluído

Internet

- 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

# 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](#)

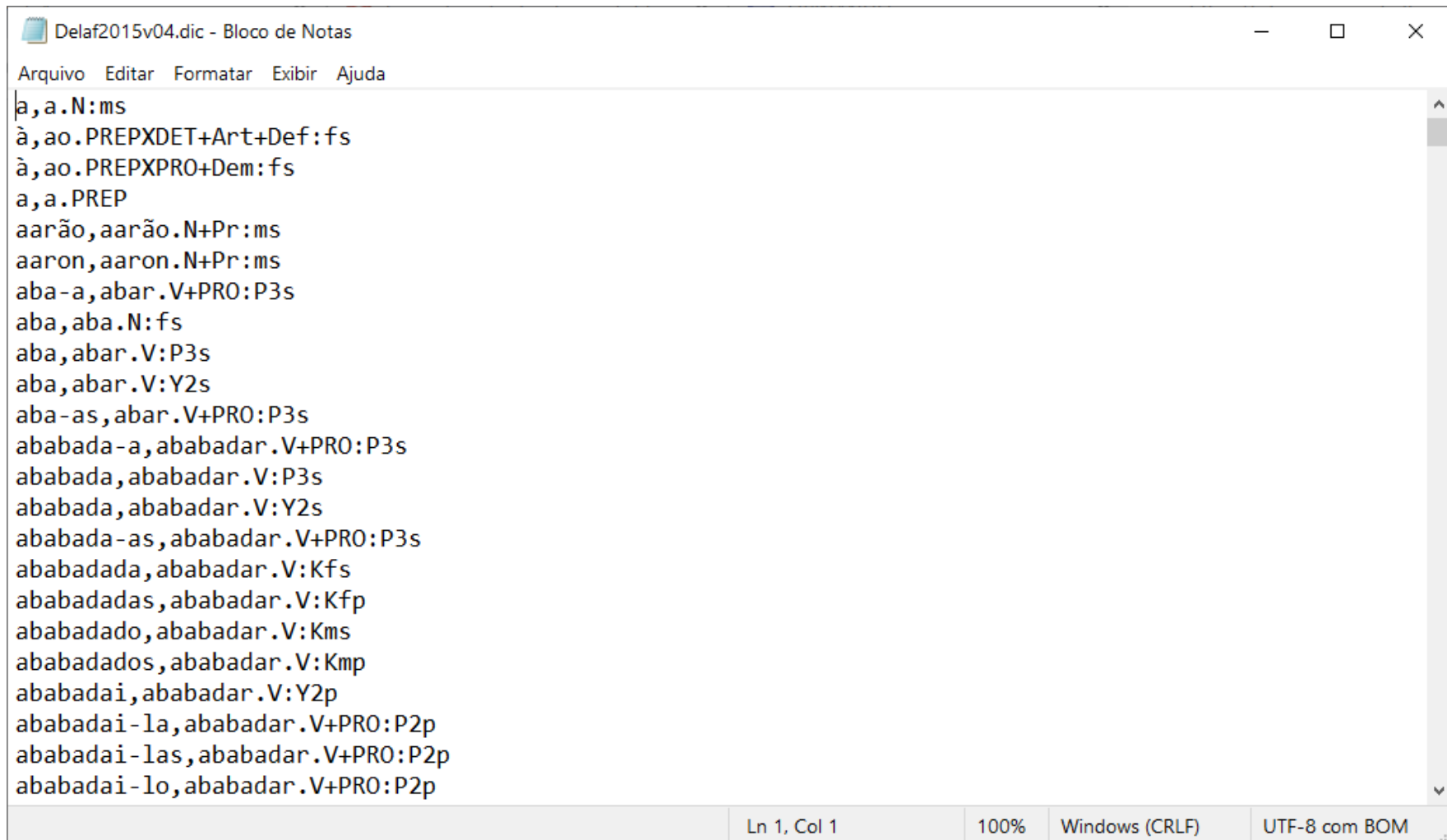
# [ PLN ]

- **Conhecimento linguístico** é a base para muitos sistemas que manipulam língua natural, direta ou indiretamente modelado
  - Extração de conhecimento de **córpus** e **bases de dados**
    - Regras gramaticais, sintáticas e discursivas
    - Estrutura textual
    - Regras de tradução
    - Critérios para resumir
    - Estatísticas sobre uso de termos
    - Conhecimento de domínio
    - Etc.

Ainda estamos distantes da capacidade de Ultron ;-)



# Léxico de uso geral



Delaf2015v04.dic - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
a, a.N:ms
à, ao.PREPXDET+Art+Def:fs
à, ao.PREXP+Dem:fs
a, a.PREP
aarão, aarão.N+Pr:ms
aaron, aaron.N+Pr:ms
aba-a, abar.V+PRO:P3s
aba, aba.N:fs
aba, abar.V:P3s
aba, abar.V:Y2s
aba-as, abar.V+PRO:P3s
ababada-a, ababadar.V+PRO:P3s
ababada, ababadar.V:P3s
ababada, ababadar.V:Y2s
ababada-as, ababadar.V+PRO:P3s
ababadada, ababadar.V:Kfs
ababadadas, ababadar.V:Kfp
ababadado, ababadar.V:Kms
ababadados, ababadar.V:Kmp
ababadai, ababadar.V:Y2p
ababadai-la, ababadar.V+PRO:P2p
ababadai-las, ababadar.V+PRO:P2p
ababadai-lo, ababadar.V+PRO:P2p
```

Ln 1, Col 1    100%    Windows (CRLF)    UTF-8 com BOM

# Léxico de sentimentos

%		abafar	125	127	129
1	funct	abafara	125	127	129
2	pronoun	abafaram	125	127	129
3	ppron	abafaras	125	127	129
4	i	abafardes	125	127	129
5	we	abafarei	125	127	129
6	you	abafareis	125	127	129
7	shehe	abafarem	125	127	129
8	they	abafaremos	125	127	129
9	ipron	abafares	125	127	129
10	article	abafaria	125	127	129
11	verb	abafariam	125	127	129
12	auxverb	abafarias	125	127	129
13	past	abafarmos	125	127	129
14	present	abafará	125	127	129
15	future	abafará	125	127	129
16	adverb	abafarás	125	127	129
17	preps	abafarão	125	127	129
18	conj	abafaríamos	125	127	129
19	negate	abafaríeis	125	127	129
20	quant	abafas	125	127	129
21	number	abafasse	125	127	129
22	swear	abafassem	125	127	129
121	social	abafasses	125	127	129
122	family	abafaste	125	127	129
123	friend	abafastes	125	127	129
124	humans	abafava	125	127	129
125	affect	abafavam	125	127	129
126	posemo	abafavas	125	127	129
127	negemo				
128	anx				
129	anger				
130	sad				



# Google n-grams

- *The n-gram counts were generated from approximately 1 trillion word tokens of text from publicly accessible Web pages.*
- *File sizes: approx. 24 GB compressed (gzip'ed) text files Number of tokens: 1,024,908,267,229 Number of sentences: 95,119,665,584 Number of unigrams: 13,588,391 Number of bigrams: 314,843,401 Number of trigrams: 977,069,902 Number of fourgrams: 1,313,818,354 Number of fivegrams: 1,176,470,663*
- *ceramics collectables collectibles 55 ceramics collectables fine 130 ceramics collected by 52 ceramics collectible pottery 50 ceramics collectibles cooking 45 ceramics collection , 144 ceramics collection . 247 ceramics collection 120 ceramics collection and 43 ceramics collection at 52 ceramics collection is 68 ceramics collection of 76 ceramics collection | 59 ceramics collections , 66 ceramics collections . 60 ceramics combined with 46 ceramics come from 69 ceramics comes from 660 ceramics community , 109 ceramics community . 212 ceramics community for 61 ceramics companies . 53 ceramics companies consultants 173 ceramics company ! 4432 ceramics company , 133 ceramics company . 92 ceramics company 41 ceramics company facing 145 ceramics company in 181 ceramics company started 137 ceramics company that 87 ceramics component ( 76 ceramics composed of 85 ceramics composites ferrites 56 ceramics composition as 41 ceramics computer graphics 51 ceramics computer imaging 52 ceramics consist of 92*

# Universal Dependencies

## Current UD Languages

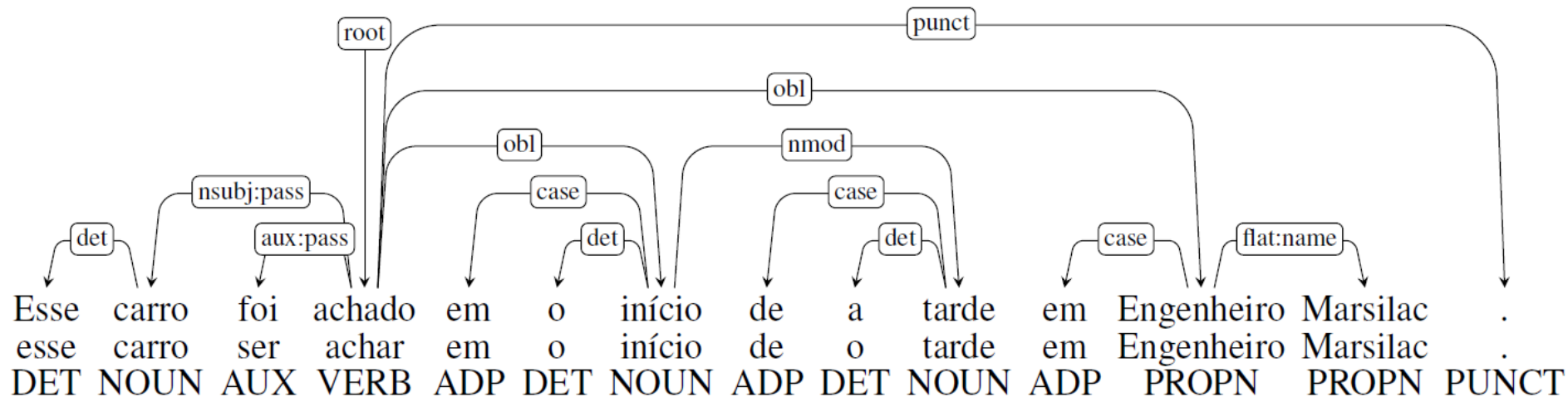
Information about language families (and genera for families with multiple branches) is mostly taken from [WALS Online](http://WALS-Online) (IE = Indo-European).

▶		Abaza	1	3K	☞	Northwest Caucasian
▶		Afrikaans	1	49K	↔	IE, Germanic
▶		Akkadian	1	1K	📖	Afro-Asiatic, Semitic
▶		Albanian	1	<1K	W	IE, Albanian
▶		Amharic	1	10K	📖✍️📖📖	Afro-Asiatic, Semitic
▶		Ancient Greek	2	416K	📖📖	IE, Greek
▶		Arabic	3	1,042K	📖W	Afro-Asiatic, Semitic
▶		Armenian	1	52K	📖✍️↔📖	IE, Armenian
▶		Assyrian	1	<1K	📖	Afro-Asiatic, Semitic
▶		Bambara	1	13K	📖	Mande
▶		Basque	1	121K	📖	Basque
▶		Belarusian	1	13K	↔📖	IE, Slavic
▶		Bhojpuri	2	6K	📖	IE, Indic
▶		Breton	1	10K	✍️📖📖🎵W	IE, Celtic
▶		Bulgarian	1	156K	↔📖	IE, Slavic
▶		Buryat	1	10K	✍️📖	Mongolic
▶		Cantonese	1	13K	☞	Sino-Tibetan
▶		Catalan	1	531K	📖	IE, Romance
▶		Chinese	5	285K	📖📖☞W	Sino-Tibetan
▶		Classical Chinese	1	130K	📖	Sino-Tibetan
▶		Coptic	1	42K	📖📖	Afro-Asiatic, Egyptian
▶		Croatian	1	199K	📖W	IE, Slavic
▶		Czech	5	2,226K	↔✍️📖📖W	IE, Slavic
▶		Danish	2	100K	📖📖☞	IE, Germanic
▶		Dutch	2	306K	📖W	IE, Germanic
▶		English	9	648K	📖📖✍️✍️📖📖📖📖📖W	IE, Germanic
▶		Erzya	1	16K	📖	Uralic, Mordvin
▶		Estonian	2	481K	📖📖📖📖📖	Uralic, Finnic
▶		Faroese	2	10K	📖📖W	IE, Germanic
▶		Finnish	3	377K	📖✍️📖W	Uralic, Finnic
▶		French	8	1,157K	📖✍️📖📖☞W	IE, Romance
▶		Galician	2	164K	↔📖	IE, Romance
▶		German	4	3,753K	📖📖📖W	IE, Germanic
▶		Gothic	1	55K	📖	IE, Germanic

# Universal Dependencies

## ■ Análise sintática

- Léxico
- Etiquetador morfosintático
- Conhecimento sintático (regras, estatística, pesos neurais, etc.)
- Desambiguação da análise





# WordNet de Princeton

## WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

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

Display options for sense: (gloss) "an example sentence"

### Noun

- [S:](#) (n) **house** (a dwelling that serves as living quarters for one or more families) *"he has a house on Cape Cod"; "she felt she had to get out of the house"*
- [S:](#) (n) **firm, house, business firm** (the members of a business organization that owns or operates one or more establishments) *"he worked for a brokerage house"*
- [S:](#) (n) **house** (the members of a religious community living together)
- [S:](#) (n) **house** (the audience gathered together in a theatre or cinema) *"the house applauded"; "he counted the house"*
- [S:](#) (n) **house** (an official assembly having legislative powers) *"a bicameral legislature has two houses"*
- [S:](#) (n) **house** (aristocratic family line) *"the House of York"*
- [S:](#) (n) **house** (play in which children take the roles of father or mother or children and pretend to interact like adults) *"the children were playing house"*
- [S:](#) (n) **sign of the zodiac, star sign, sign, mansion, house, planetary house** ((astrology) one of 12 equal areas into which the zodiac is divided)
- [S:](#) (n) **house** (the management of a gambling house or casino) *"the house gets a percentage of every bet"*

Buscar

Todas



Mostrar Exemplo



Mostrar Antônimos

Quis dizer [cântaro](#)?**cantar** (Verbo)1. **cantar**, ditar2. **cantar**, descantar, ensoar, entoar, soar3. **cantar**, alevantar, consagrar, divinizar, elevar, enaltar, enaltecer, encumear, engrandecer, enobrecer, exalçar, exaltar, extremar, heroificar, levantar, sobalçar, soberanizar, sobredoirar, sobredourar, sublimar4. **cantar**, cantarejar, cantorolar, musicar, trautear5. **cantar**, gavionar, paqueirar, paquerar**cantar** (Substantivo)1. **cantar**, canção, cantadela, cantiga, trova

# FrameNet Brasil

FrameNet Brasil Webtool 3.0 [fnbr-docker]

Reports Grapher

Frames

Search Frame Search LU

- Frames
  - Abandono
  - Abertura
  - Absorção\_de\_calor
  - Abster-se
  - Abundância
  - Abundância\_distribuída
  - Abundar\_com
  - Abusar
  - Acabar\_de\_descobrir
  - Ação\_transitiva
  - Aceitar\_ou\_recusar\_a\_agir
  - Acertar\_ou\_errar
  - Acertar\_o\_alvo
  - Acessórios\_de\_vestuário
  - Acidente
  - Ações\_do\_árbitro
  - Acomodação
  - Acompanhamento
  - Acordar
  - Acumular
  - Acusação
  - Adiar**
  - Adição
  - Adjacência

## Adiar

### Definição

Um **Agente** decide temporariamente não executar um **Ação\_desejável** ou não interagir com uma **Entidade\_saliente**.

### Exemplo(s)

### Elementos de Frame Nucleares

#### FE Core:

**Ação\_desejável** [Desirable\_action]

**excludes:** Entidade\_saliente Identifica o evento ou atividade em que o **Agente** não se engaja por algum tempo.

**Agente** [Agent]

**semantic\_type:** @sentient

O **Agente** decide não participar de uma **Ação\_desejável** por um tempo ou não interagir com uma **Entidade\_saliente**.

**Entidade\_saliente** [Salient\_entity]

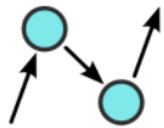
Uma entidade que evoca uma **Ação\_desejável** contextualmente inferível de que o **Agente** está temporariamente ausente.

### Elementos de Frame Não-Nucleares

### Relações

### Unidades Lexicais

# Senso comum



## ConceptNet

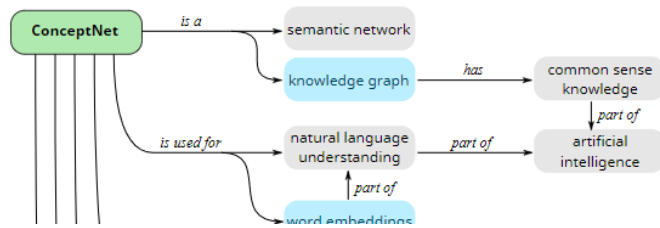
An open, multilingual knowledge graph

- [Documentation](#)
- [FAQ](#)
- [Chat](#)
- [Blog](#)

## What is ConceptNet?

**ConceptNet** is a freely-available semantic network, designed to help computers understand the meanings of words that people use.

ConceptNet originated from the crowdsourcing project Open Mind Common Sense, which was launched in 1999 at the MIT Media Lab. It has since grown to include knowledge from other crowdsourced resources, expert-created resources, and games with a purpose.



## Examples

To explore what's in ConceptNet, try browsing what it knows about any of these terms:

- [en](#) word
- [fr](#) mot
- [nl](#) woord
- [es](#) palabra
- [pt](#) palavra
- [ja](#) 単語

- [en](#) graph
- [en](#) knowledge
- [en](#) learn
- [en](#) natural language
- [en](#) semantic network
- [mul](#)

## Word vectors and recent publications

# Conhecimento de mundo



WIKIPÉDIA  
A enciclopédia livre

[Página principal](#)  
[Conteúdo destacado](#)  
[Eventos atuais](#)  
[Esplanada](#)  
[Página aleatória](#)  
[Portais](#)  
[Informar um erro](#)  
[Loja da Wikipédia](#)

[Colaboração](#)

[Boas-vindas](#)

[Ajuda](#)

[Página de testes](#)

[Portal comunitário](#)

[Mudanças recentes](#)

[Manutenção](#)

[Criar página](#)

[Páginas novas](#)

[Contato](#)

[Donativos](#)

[Ferramentas](#)

[Páginas afluentes](#)

[Alterações relacionadas](#)

[Carregar ficheiro](#)

[Páginas especiais](#)

[Hiperligação permanente](#)

[Informações da página](#)

[Não autenticado](#) [Discussão](#) [Contribuições](#) [Criar uma conta](#) [Entrar](#)

[Página principal](#)

[Discussão](#)

Ler

[Ver código-fonte](#)

[Ver histórico](#)



[\[ocultar\]](#)

## BEM-VINDOS À WIKIPÉDIA

A enciclopédia livre que todos podem editar

1 043 418 artigos em português

6 016 usuários ativos

[Ajuda](#) · [Índice](#) · [Perguntas](#) · [Políticas](#) · [Portais](#)

[Arte](#)

[Biografias](#)

[Ciência](#)

[Filosofia](#)

[Geografia](#)

[História](#)

[Matemática](#)

[Sociedade](#)

[Tecnologia](#)

## ARTIGO EM DESTAQUE

**Metrô de São Paulo** ou **Metropolitano de São Paulo**, conhecido popularmente como **Metrô**, é um sistema de transporte metroviário que serve a cidade de São Paulo, no Brasil. O Metrô de São Paulo é operado pela [Companhia do Metropolitano de São Paulo](#), sociedade de economia mista do estado de São Paulo.

Fundada em 24 de abril de 1968, a empresa é responsável pelo planejamento, projeto, construção e operação do sistema de transporte metroviário na [Região Metropolitana de São Paulo](#). Tendo a maior parte de seu controle acionário associada ao governo do estado, é subordinada à [Secretaria dos Transportes Metropolitanos do Estado de São Paulo](#). Integra também a [Rede Metropolitana de Transporte de São Paulo](#). O [Grupo CCR](#) (por meio das concessionárias [ViaQuatro](#) e



## EVENTOS ATUAIS

### Pandemia de COVID-19

[Doença](#) · [Vírus](#) · [Diagnóstico](#) · [Cronologia](#) · [Por país](#) · [Impactos](#) · [Mortos notáveis](#) · [Portal](#)

- Na música, a revista *Billboard* (logo)  anuncia sua primeira [parada musical global](#), a *Billboard Global 200*.
- Ex-presidente da Índia **Pranab Mukherjee** morre aos 84 anos.
- Primeiro-ministro do Japão Shinzō Abe renuncia** ao cargo após alegar problemas de saúde.
- Furacão Laura** mata mais de 69 pessoas no Caribe e nos Estados Unidos.
- Atentados terroristas** em Jolo, Filipinas, matam 15 pessoas e ferem outras 75.
- No automobilismo, [Teluma Sete anos](#) se 500 Milhas de



# Word Embeddings

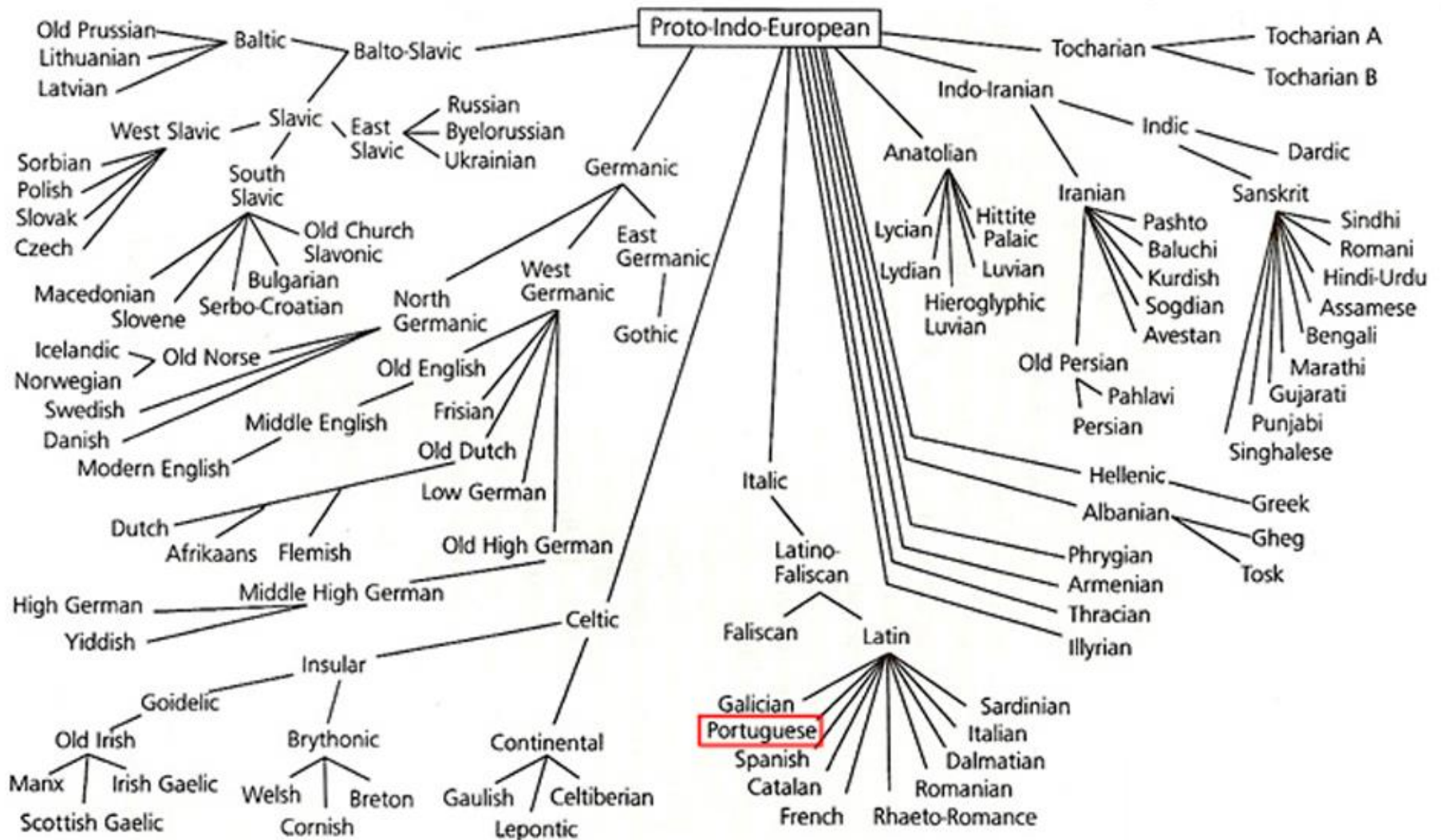
- Vetores para palavras
  - “rei” – “homem” + “mulher” = “rainha”

```
cbow_s50.txt - Bloco de Notas
Arquivo Editar Formatar Exibir Ajuda
puxada -0.034887 0.171691 0.286350 0.001959 -0.265040 -0.363308 0.008134 0.028733 0.244310 0.070709
0.650404 0.547310 0.020620 0.406854 -0.080173 -0.263757 -0.294278 0.074062 -0.354263 0.029041
0.072985 0.032874 -0.856120 0.248842 0.180632 -0.570520 -0.046392 0.211533 0.291666 -0.039001 -
0.135342 0.459373 0.240077 0.167408 -0.502166 -0.216179 0.464056 0.222689 -0.246912 0.082746
0.584343 -0.200991 0.137767 0.300268 0.026158 0.162092 -0.222494 -0.074502 -0.022367 -0.437183
mentalidades -0.044354 -0.169533 -0.203965 0.641967 0.391989 0.492085 0.553901 0.105235 -0.003434 -
0.018459 -0.307926 0.014193 0.457690 0.386151 0.576904 0.169937 -0.355718 0.008306 -0.364774
0.674785 -0.131864 0.361762 0.167776 0.091368 -0.297379 0.086894 -0.107647 0.025288 -0.153701
0.543972 0.408720 -0.513095 0.242194 0.346348 -0.144786 -0.511438 0.369222 -0.019037 0.072361 -
0.304172 -0.207620 0.334368 0.555059 0.400517 0.345321 0.210812 0.206703 -0.286901 0.073816 -
0.552975
rega 0.376240 -0.052562 0.003062 0.199682 -0.055232 0.073866 0.238939 0.251127 0.350433 0.492652
0.145822 -0.200336 -0.179975 0.072417 0.067692 0.540316 -0.015247 -0.080517 -0.230508 0.387325
0.248653 0.253949 -0.315998 0.417741 -0.000485 -0.376826 -0.097732 -0.050296 0.221432 0.093637
0.494296 -0.115676 0.046192 -0.108311 0.266834 -0.123263 -0.347469 0.137446 -0.004984 0.222727
0.136311 0.428985 0.111122 0.284440 -0.015603 0.245002 0.033343 -0.016158 -0.477271 -0.264597
filtragem 0.202029 0.064431 0.069496 0.257424 0.177513 0.091983 0.108795 0.071681 0.166440 0.442256
0.247053 -0.027207 -0.101376 0.511969 0.241767 0.586548 -0.111406 0.123894 -0.167538 0.554824 -
0.023976 -0.101418 -0.202871 0.446157 0.093930 -0.529788 0.002787 0.011946 -0.095409 0.212605
0.274954 -0.018061 0.125794 -0.330955 0.398550 0.114111 -0.297261 0.349666 0.140497 0.084230
0.182703 0.309072 0.265252 0.093108 -0.065538 0.117508 0.235194 -0.165781 -0.505004 -0.077262
odisseia 0.145710 0.271790 0.043788 0.510743 0.306707 0.204887 0.062529 0.486732 0.100621 -0.095087
-0.288090 -0.395822 0.092362 0.431765 -0.513534 -0.045715 0.095737 0.241324 -0.471110 0.358929 -
```



# OS DESAFIOS DO PORTUGUÊS

# FAMÍLIAS DE LÍNGUAS



# O PORTUGUÊS NO MUNDO

- Falado em 10 países



# DADOS SOBRE A LÍNGUA

## ○ Português

- 6ª língua mais falada no mundo
- Grande número de vocábulos
  - Dicionário Houaiss
    - 228.500 entradas
    - 376.500 acepções
    - 415.500 sinônimos
    - 26.400 antônimos
    - 57.000 palavras arcaicas
  - Academia Brasileira de Letras: mais de 356.000 palavras

# O PORTUGUÊS NO MUNDO

- Comunidade dos Países de Língua Oficial Portuguesa (CPLP)
  - Instituto Internacional da Língua Portuguesa (IILP)
    - Divulgação e promoção da língua portuguesa
- Português na **web** (segundo mapeamento de 2012)
  - 5ª língua mais utilizada na web, atrás do inglês, chinês, espanhol e japonês
  - 82,5 milhões de utilizadores
    - Entre 2000 e 2010, expansão de 990%
  - 3ª língua mais usada no Twitter, atrás do inglês e do japonês
    - Razão: aumento de acesso à web no Brasil

# DESAFIOS PARA PROCESSAMENTO

- **Variações** nos países em que é falado
  - Pronúncias variadas
  - Ortografias variadas, apesar do acordo ortográfico
  - Diferenças nas construções sintáticas mais usuais
  - Sentidos diferentes de palavras
  - “Perfis de usuários” diferentes: hábitos, expectativas e cultura variados



# DESAFIOS PARA PROCESSAMENTO

- Dialetos (Branco et al., 2012)

Em Portugal, a divisão geográfica dos dialetos [13] distingue os dialetos do Centro-Sul, os dialetos do Norte e os dialetos das ilhas atlânticas. Os dialetos do Norte podem ser identificados pela ausência da distinção fonológica entre /b/ e /v/, com prevalência do /b/, pela preservação de antigos ditongos, e pela existência de fricativas ápicoalveolares. As diferenças entre estes dialetos encontram-se sobretudo ao nível da fonética e fonologia e ao nível lexical, sendo todos eles mutuamente compreensíveis de forma imediata (possivelmente com a exceção de alguns dialetos das ilhas).



# DESAFIOS PARA PROCESSAMENTO

- Dialetos (Branco et al., 2012)

Quanto ao Brasil, dada a dimensão geográfica deste país, não é viável apresentar aqui as suas variedades linguísticas. Por razões geográficas, políticas e sociais, não é possível falar de uma variedade padrão do português do Brasil. Os especialistas tendem a mencionar “normas urbanas cultas”.

A situação das variedades africanas do português é variada: enquanto em Angola e Moçambique o número de falantes de português tem vindo a aumentar desde a independência destes países, noutros casos, como São Tomé e Príncipe ou Cabo Verde, em muitas circunstâncias utiliza-se amplamente o crioulo e o português é adquirido como língua segunda.

# DESAFIOS PARA PROCESSAMENTO

- Zuchini (2011)
  - Elevado número de vocábulos existentes
  - Elevado número de sinônimos entre vocábulos
  - Elevado número de flexões verbais
  - Diversas possibilidades de construção sintática
  - Elevado número de flexões em gênero, número e grau
  - Grande número de exceções para praticamente todas as regras

# Hora da prática



# [ Passo a passo ]

- Em grupos de ~10 alunos
- Cada grupo vai ganhar um “projeto” e deverá especificar como o desenvolverá, pensando nos detalhes (de forma mais realista possível) de cada etapa do trabalho em PLN
- Cada grupo ingressará em uma outra sala virtual e, uma vez na outra sala, escolherá um líder para coordenar o trabalho
  - Se preferir, também pode escolher um porta-voz para representar o grupo e comunicar as decisões, assim como ajudantes para tomarem notas 😊

# [ Passo a passo ]

- Tópicos para discutir
  - Como exatamente é a tarefa que vai implementar: dados de entrada & dados de saída?
  - Qual o perfil do usuário de seu sistema?
  - Que níveis de conhecimento linguístico são necessários para atender a demanda?
  - Como coletar dados/córpus e as informações necessárias?
  - Que recursos e ferramentas de apoio são necessários?
  - Qual o tempo necessário para o desenvolvimento?
  - Quem são os profissionais envolvidos no desenvolvimento? Quantos são?
  - Como “medir a qualidade” do resultado final (afinal, ninguém vai querer comprar algo sem saber se é bom)
  
- Discutam alternativas “sortidas”, suas vantagens e desvantagens, etc.

# [ Passo a passo ]

- Em seguida
  - Vamos (eu e estagiário PAE) visitar esporadicamente cada sala para discutir a proposta de cada grupo
  - Ao final, voltaremos para a sala principal para continuar/finalizar a aula

# [ Projetos possíveis ]

- Corretor ortográfico (realmente bom 😊) para smartphones
- Revisor gramatical para um editor de texto qualquer
- Classificador de polaridade de comentários na web
- Detector de discurso de ódio na web
- Detector de notícias falsas
- Avaliador de redações de vestibular
- Simplificador textual
- Sumarizador automático
- Etc.

# Tarefas da semana

- Nunes, M.G.V. (2008). O Processamento de Línguas Naturais: para quê e para quem? Notas Didáticas do ICMC, N. 73. 12p.
  - No e-Disciplinas
- Provinha 3 disponível à tarde