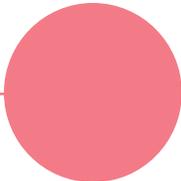


LINGUÍSTICA
PARA O ENSINO SUPERIOR

13

LINGUÍSTICA COMPUTACIONAL

CLÁUDIA FREITAS



Linguística computacional

Capítulo 2

1. Indique a principal diferença entre os modelos baseados em regras e os baseados em aprendizado de máquina.
2. Qual o papel do conhecimento linguístico nos modelos baseados em regras e em aprendizado de máquina?
3. Qual a relevância das avaliações conjuntas e dos benchmarks para o PLN?
4. Qual a diferença entre a avaliação intrínseca e a avaliação extrínseca?
5. Qual a diferença entre precisão e abrangência?
6. Identifique o erro e corrija a anotação das frases abaixo, usando o formalismo indicado na seção 2.1.1. Você pode seguir o modelo:

```
Não_ADV o_ART vi_VERBO ontem_ADV  
pos[0]=ART pos[1]=V → pos=PREP
```

- (a) A_ART partida_VERBO será_AUX transferida_VERBO para_PREP a_PREP próxima semana.
 - (b) A_ART situação_SUBST estará_AUX resolvida_VERB até_PREP a_PREP semana_SUBST que_PRON vem_VERBO.
7. O que significa dizer: “Os textos contêm informação não estruturada”?

Capítulo 3

1. Qual a diferença entre o empirismo e o racionalismo e como ela se manifesta no PLN?

2. O que são os hapax legomena e que impactos eles têm sobre o pln baseado em regras? E para o pln baseado em aprendizado de máquina?
3. Esta é uma lista com todos os verbos de todos os corpora do projeto AC/DC — considerando os lemas: <http://www.linguateca.pt/acesso/tokens/lista.lemas.total.V.txt>

Esta é uma lista com todas as palavras de todos os corpora do projeto AC/DC — considerando os lemas: <http://www.linguateca.pt/acesso/tokens/lemas.total.txt>

A distribuição de Zipf é observável nas listas?

O acervo do AC/DC é dinâmico e continha um total de 1,3 bilhão de palavras no momento em que a lista foi obtida. Os textos são de diferentes gêneros e variantes da língua portuguesa. Para uma visão detalhada do material: <http://www.linguateca.pt/ACDC/>
As listas foram obtidas desta página: <http://www.linguateca.pt/acesso/ordenador.php>

Capítulo 4

1. Por que os padrões de Hearst exemplificam a tensão entre precisão e abrangência?
2. O capítulo 3 menciona a polissemia da palavra açúcar. Neste endereço http://wn.mybluemix.net/search?search_field=word_pt&term=açúcar, você pode ver os diferentes sentidos de açúcar conforme a OpenWordNet-PT. Para navegar pela OpenWordNet, basta clicar nos synsets ou digitar uma palavra qualquer na caixa de busca (figura 1). Para cada um dos sentidos mencionados (açúcar carboidrato e açúcar refinado), vá subindo na hierarquia, clicando em Hyponym (figura 2). Qual o fim da linha e qual a última palavra? Em algum momento, os caminhos dos açúcares vão se cruzar?

10 results found for 'açúcar'

RDF Type:

- NounSynset (10)
- BaseConcept (1)
- CoreConcept (1)

Lexicographer file:

- noun.food (4)
- noun.substance (4)
- noun.plant (2)

words (pt_BR):

- 1 (8)
- 2 (1)
- 4 (1)

words (en):

- 1 (5)
- 3 (3)
- 2 (2)

1. [07859284-n](#) refined_sugar, sugar | **açúcar**
 - (a white crystalline carbohydrate used as a sweetener and preservative)
2. [07595649-n](#) cane_sugar | **açúcar de cana**
 - (sugar from sugarcane used as sweetening agent)
3. [12132754-n](#) sugar_cane, sugarcane | **cana-de-açúcar**
 - (juicy canes whose sap is a source of molasses and commercial sugar, fresh canes are sometimes chewed for the juice)
4. [12132502-n](#) Saccharum_officinatum, sugar_cane, sugarcane | **cana de açúcar**
 - (tall tropical southeast Asian grass having stout fibrous jointed stalks; sap is a chief source of sugar)
5. [07595914-n](#) powdered_sugar | **açúcar de confeiteiro**
 - (sugar granulated into a fine powder)
6. [14791453-n](#) cane_sugar | **açúcar de cana**
 - (sucrose obtained from sugar cane)
7. [14823690-n](#) corn_sugar | **açúcar de milho**
 - (dextrose made by hydrolysis of cornstarch)
8. [07596160-n](#) icing_sugar | **açúcar de confeiteiro**
 - (finely powdered sugar used to make icing)
9. [15062955-n](#) jagghery, jaggery, jaggary | **rapadura, açúcar mascavo**
 - (unrefined brown sugar made from palm sap)
10. [14792703-n](#) saccharide, carbohydrate, sugar | **hidratos de carbono, carboidrato, Carboidratos, açúcar**
 - (an essential structural component of living cells and source of energy for animals; includes simple sugars with small molecules as well as macromolecular substances; are classified according to the number of monosaccharide groups they contain)

1.

açúcar

Figura 1

- Lexicographer file: (noun.food)
- RDF Type: (BaseConcept CoreConcept NounSynset)
- Hypernym of: [[beet_sugar](#), [caramel](#), [cane_sugar](#), [granulated_sugar](#), [s_corn_sugar](#), [lump_sugar](#)]
- Hyponym of: [[sweetening](#)]

Figura 2

Explore também outros recursos:

Onto.PT: <http://ontopt.dei.uc.pt/>

PAPEL: <http://www.linguateca.pt/PAPEL/>

PULO: <http://wordnet.pt/>

TEP: <http://www.nilc.icmc.usp.br/tep2/>

FrameNet: <http://framenet.icsi.berkeley.edu/fndrupal/>

FrameNet Brasil: <http://www.ufjf.br/framenetbr/>

VerbNet-Br: <http://www.nilc.icmc.usp.br/verbnnetbr/>

VerbNet: <http://verbs.colorado.edu/~mpalmer/projects/verbnnet.html>

Unified Verb Index: <http://verbs.colorado.edu/verb-index/>

WordNet: <http://wordnet.princeton.edu/>

Capítulo 5

1. Qual a diferença entre palavra e token? Por que um corpus, se bem processado, terá um número de tokens maior que o número de palavras?
2. A lista da esquerda contém as quarenta palavras **mais frequentes** iniciadas com DES, segundo todos os corpora do projeto AC/DC, e a lista da direita, uma amostra das quarenta palavras **menos frequentes** iniciadas com DES, segundo esse mesmo material.

desde/17.1814
desenvolvimento/52.340
desenvolver/40.694
destinar/36.162
descobrir/35.034
desejar/28.017
descer/25.332
destino/23.070
desaparecer/21.617
destruir/20.926
despesa/20.776
desejo/20.099
desemprego/19.934
desafio/18.942
desportivo/18.926
destacar/18.518
destaque/18.477
desconhecer/16.711
desenho/15.034
deslocar/14.419
descrever/14.186
desporto/12.266
designar/11.825
desempenhar/11.166
descoberta/10.875
descida/10.494
desistir/10.285
desempenho/10.279
deslocação/9.447
desmentir/9.359
designadamente/9.205
desencadear/8.843
destruição/8.616
despacho/8.489
desenhar/8.309
deserto/8.223
despertar/8.159
desviar/7.867
desfile/7.702
desempregado/7.626

desaquisição/1
desapetrechar/1
desapetitoso/1
desapetecer/1
desanuviação/1
desantropomórfico/1
desantropomorfizar/1
desaniversário/1
desanimante/1
desanimalizante/1
desanimadoramente/1
desalinização/1
desalinhadamente/1
desalienado/1
desaliado/1
desalfinetar/1
desalfandegagem/1
desalcoolizar/1
desalcoolização/1
desajuizar/1
desajuizadamente/1
desaguamento/1
desafrouxar/1
desafrontoso/1
desafinamento/1
desafinador/1
desadensar/1
desacostumadamente/1
desacidificar/1
desacidentado/1
desacender/1
desaceleradora/1
desacautelamento/1
desabraçar-se/1
desabraçar/1
desaboborar/1
desabitualmente/1
desabitantes/1
desabrigar/1
desabandonar/1

Como o processo de tokenização baseado em subpalavra lida com os elementos de cada lista? Como o processo de tokenização baseado em palavra lida com os elementos de cada lista? Qual a melhor estratégia? Considerando a amostra de sessenta palavras, qual abordagem acerta mais?

Capítulos 6 e 7

1. O que é anotação? Apresente três tipos, indicando seu objetivo e os principais desafios envolvidos.
2. Qual o papel da/do linguista na anotação?
3. Qual a relevância da anotação no pln?
4. Considere os seguintes tagsets de pos:

TAGSET 1

adjetivo ADJ
preposição PREP
advérbio ADV
verbo auxiliar VAUX
conjunção coordenativa KC
artigo ART
pronome adjetivo PROADJ
interjeição IN
substantivo N
numeral NUM
pronome pessoal PROPESS
pronome nominal PROSUB
nome próprio NPROP
pontuação PU
conjunção subordinativa KS
pronome subordinativo conectivo PRO-KS
pronome subordinativo conectivo relativo PRO-KS-REL
advérbio subordinativo conectivo ADV-KS
advérbio subordinativo relativo ADV-KS-REL
moeda CUR
verbo V
particípio PCP
palavras denotativas PDEN

TAGSET 2

adjetivo ADJ
preposição ADP
advérbio ADV
verbo auxiliar AUX
conjunção coordenativa CONJ
determinante DET
interjeição INTJ
substantivo NOUN
numeral NUM
partícula PART
pronome PRON
nome próprio PROPN
pontuação PUNCT
conjunção subordinativa SCONJ
símbolo SYM
verbo VERB
outros X

Imagine que você precise converter um corpus anotado com o tagset 1 para o tagset 2.

Quais etiquetas não serão diretamente alinhadas?

Para saber o que cada etiqueta significa e conhecer seus exemplos:

Tagset 1 (a primeira versão tagset do MarcMorpho): <http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf>

Tagset 2 (a segunda versão do tagset Universal Dependencies): versão original (em inglês): <https://universaldependencies.org/u/pos/all.html> e uma adaptação em português: <http://comcorhd.lettras.puc-rio.br/documenta-o-ud-pt/>

5. A tabela a seguir contém resultados de um sistema de identificação automática de papéis semânticos. Que etiquetas parecem ser mais fáceis para as máquinas, levando aos números mais altos? Quais levam aos números mais baixos?

	CORRETOS	EXCESSO	FALTANTES	PRECISÃO	ABRANGÊNCIA	F1
TODOS	353	173	187	67,11	65,37	66,23
A0	99	16	26	86,09	79,2	82,5
A1	159	53	53	75	75	75
A2	8	13	31	38,1	20,51	26,67
A3	0	0	2	0	0	0
A4	0	2	1	0	0	0
AM-ADV	9	8	11	52,94	45	48,65
AM-CAU	1	4	0	20	100	33,33
AM-DIR	0	0	2	0	0	0
AM-DIS	8	8	3	50	72,73	59,26
AM-EXT	0	2	2	0	0	0
AM-LOC	12	14	16	46,15	42,86	44,44
AM-MNR	6	9	13	40	31,58	35,29
AM-NEG	16	5	2	76,19	88,89	82,05
AM-PNC	2	4	3	33,33	40	36,36
AM-PRD	0	9	7	0	0	0
AM-REC	2	3	3	40	40	40
AM-TMP	31	23	12	57,41	72,09	63,92

6. Consulte a OWN-PT relacionada à palavra trabalho (ou consulte um dicionário convencional): http://wn.mybluemix.net/search?search_field=word_pt&term=trabalho. Quais synsets (ou acepções de dicionário) de trabalho podem ser usadas para as seguintes frases? Se possível, peça a outras pessoas para fazerem o mesmo e compare os resultados.
- Nosso **trabalho** foi relevante para a população da cidade.
 - Publicou vários **trabalhos** nas áreas de administração pública e economia.
 - A Associação Médica Brasileira (AMB) têm feito um **trabalho** para corrigir distorções, mas ainda é pouco.
 - O que se tem de informação é que um **trabalho** de impermeabilização era feito em a loja.
 - Havia uma lenda de que ele não repetia **trabalho** com um mesmo decorador.
 - Eu havia pedido demissão de o antigo **trabalho**.

- (g) Trump anuncia novas reuniões de **trabalho** entre Washington e Pyongyang.
- (h) O rapaz foi preso em seu local de **trabalho**.
- (i) Foi empossado em 1º fevereiro do ano seguinte, quando tiveram início os **trabalhos** da Assembleia Nacional Constituinte.

Faça o mesmo com o verbo trabalhar:

- (a) Tite aproveitou os treinos em Porto Alegre para estudar e **trabalhar** as características de a Seleção Paraguaia.
- (b) Os contratados poderão **trabalhar** remotamente, ou seja, home office, como explica o gerente.
- (c) A dança pode ser uma aliada para emagrecer, ganhar musculatura, **trabalhar** a autoestima, coordenação e raciocínio.
- (d) Até que ela decidiu estudar e **trabalhar** fora.
- (e) Estamos **trabalhando** duro para tentar deixar as contas em dia.
- (f) A gente corre atrás, **trabalha** honestamente, e ter que passar por isso não é fácil.
- (g) Nathalia **trabalhava** como personal trainer enquanto deveria dar expediente em o gabinete.
- (h) A partir de então, a FAB passou a **trabalhar** principalmente na Amazônia.
- (i) A OAB decidiu **trabalhar** por uma ampla campanha de esclarecimento da população sobre as diferenças entre parlamentarismo e presidencialismo.

A literatura fala de verbos serem mais polissêmicos que substantivos. Essa expectativa se confirma com relação a trabalho e trabalhar? Houve dúvidas na busca pelo sentido adequado?

7. Considere este léxico de entidades:

LOCAIS: São Paulo, Lauro de Freitas, Flamengo, Botafogo, Bahia, Monteiro Lobato

PESSOAS: Lauro de Freitas, São Paulo, Monteiro Lobato

ORG: São Paulo, Flamengo, Botafogo

As entidades são ambíguas, e a frase (1) foi anotada de maneira errada. Na sequência, estão dois tipos de regras que podem ser usadas para corrigir a anotação.

- (a) Em que elas diferem?

(1) Sou Botafogo_LOCAL desde criança.

```
SELECT (ORG) IF (-1 Vser)
```

(Lê-se: Selecione ORG se, à esquerda da palavra, houver um verbo ser.)

```
NER (0) = LOCAL e lema[-1] = ser >> NER(0) = ORG
```

(Lê se: Se o NER da palavra em análise (0) for LOCAL e o lema da palavra imediatamente à esquerda for “ser”, o NER da palavra em análise deverá ser ORG.)

(b) Escolha o formalismo mais adequado para lidar com as frases a seguir e escreva as regras de correção.

i. Lauro de Freitas considerou a experiência um sucesso.

ii. Nasceu em Lauro de Freitas, na Bahia.

iii. São Paulo disse que, de todas as virtudes, a maior é a caridade.

iv. Jogou pelo São Paulo antes de chegar ao Flamengo.

v. Chegaram em São Paulo pela manhã.

vi. Monteiro Lobato é um dos mais conhecidos escritores brasileiros.

vii. Monteiro Lobato é um município brasileiro do estado de São Paulo.

(c) Que pistas linguísticas a abordagem de vetores de palavras contextuais seguiria para resolver o problema?

8. Verifique a precisão e a abrangência da análise automática relativa ao aposto neste pequeno corpus, composto por dez frases:

Desde 1997, a responsável por o projeto, a bióloga marinha Susanne Wiigh-Mäsak, cultiva em Lyr, ilha no arquipélago de Bohuslanian, uma verdadeira indústria do bem.

Importado da Holanda, o transponder, com 64 bits de capacidade, incluirá dados de bordo, como número do veículo, itinerário, sentido da viagem etc.

Segundo o diretor de sindicato dos hotéis, restaurantes, bares e similares, James Ackel, 41, outubro superou também a marca registrada em setembro, quando 210 mil pessoas passaram pela cidade.

A nova data (antes ocorria em abril) e local foram para o fortalecimento comercial da feira, que era itinerante e agora pode permanecer em Uberaba, afirma Eduardo Biaggi, presidente da Associação dos Criadores de Nelore do Brasil (ACNB).

Dan Rostenkowski, 66, literalmente nasceu na máquina do mais clássico chefe da história política americana, o prefeito de Chicago Richard Daley.

Só nos anos mais recentes, os presidentes Richard Nixon (1969-1975), Ronald Reagan (1981-1989) e George H. Bush, o Bush pai (1989-1993), tiveram indicações para a Suprema Corte rejeitadas por o senado americano.

Na sexta-feira, dia 2 de setembro, a Pégaso, empresa organizadora, continuava vendendo coberturas dos reprodutores Sierra Show T. E., Pwrely Artistic, Hncredbble e MBJ MR Array.

Nilo disse que vai aconselhar o candidato do PDT — seu partido — ao governo do Estado, Anthony Garotinho, a manter o convênio com o governo federal em caso de vitória no segundo turno da eleição.

A peça foi fabricada por um professor do curso de mecânica da Etep (Escola Técnica Professor Everardo Passos), em São José, onde Pissardo estudou.

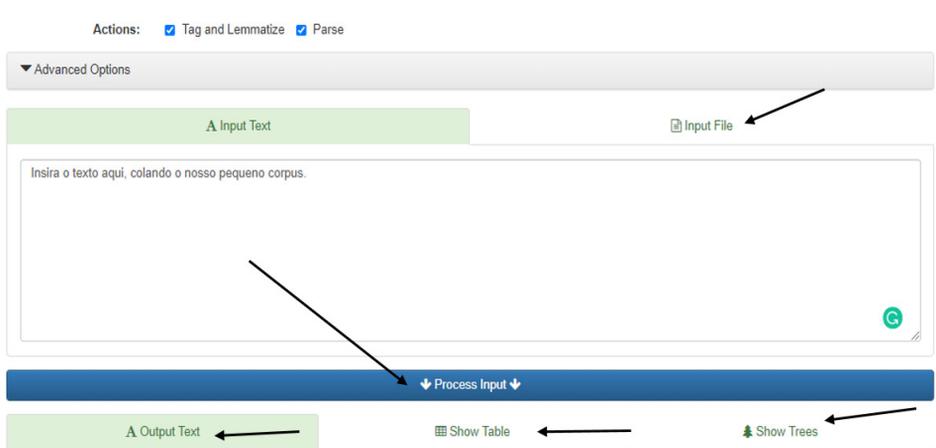
Informações na Associação Brasileira dos Criadores do Cavalo Árabe, tel.(011 263-1744).

Comentário: Para verificar o desempenho da análise automática, será preciso construir um gabarito (ou um corpus padrão-ouro anotado quanto aos apositos). Para isso, identifique quantas e quais são as estruturas apositivas do pequeno corpus. Guarde esse material — é o seu gabarito) — e vá para a análise automática.

Análise automática. Utilize a ferramenta UDPipe: <http://lindat.mff.cuni.cz/services/udpipe/>

1. Selecione um dos dois modelos disponíveis para o português, conforme a figura. A escolha é indiferente, pois você testará os dois modelos, mas um de cada vez. Cada modelo foi treinado em um corpus diferente.
2. Insira o texto. Você pode copiar e colar na caixa, conforme a figura, e clicar em Input Text ou fazer upload de um arquivo, selecionando Input File e, em seguida, Load File.

The screenshot shows the UDPipe web interface. At the top, there is a navigation bar with links for Repository, Corpus Search, TreeQuery, Treex, More Apps, About, and CLARIN. Below this, the page title is "UDPipe" with sub-links for About, Run, and REST API Documentation. A descriptive paragraph explains that UDPipe is a trainable pipeline for tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files. It is language-agnostic and can be trained on annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java, C#, and as a web service. A third-party R CRAN package also exists. Below this, it states that UDPipe is free software distributed under the Mozilla Public License 2.0 and the linguistic models are free for non-commercial use under the CC BY-NC-SA license, although original data used for training may have additional licensing conditions. The copyright is 2017 by the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic. A description of available methods is available in the API Documentation, and models are described in the UDPipe 2 models list and UDPipe 1 models list. The main content area is titled "Service" and contains a disclaimer: "The service is freely available for testing. Respect the CC BY-NC-SA licence of the models – explicit written permission of the authors is required for any commercial exploitation of the system. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome." Below the disclaimer, there is a "Model:" section with two radio buttons: "UD 2.6 (description)" (selected) and "EvaLatin20 (description)". A dropdown menu is open, showing a list of models with their respective flags and IDs. The models listed are: "czech-pdt-ud-2.6-200830", "french-spoken-ud-2.6-200830", "italian-postwita-ud-2.6-200830", "polish-pdb-ud-2.6-200830", "polish-lfg-ud-2.6-200830", "portuguese-gsd-ud-2.6-200830" (highlighted in blue), and "portuguese-bosque-ud-2.6-200830". To the left of the dropdown, there is an "Actions:" section with a text input field containing "Pd" and a button for "Advanced Options".



Pode-se observar o resultado na própria tela, caso você tenha selecionado Input File, ou em um arquivo separado, caso tenha selecionado Load File — nesse caso, automaticamente você fará download do arquivo anotado, que irá para a sua pasta Downloads.

Analise a saída da análise automática (na tela do UDPIPE, é possível observar os resultados selecionando Output Text, Show Table ou Show Trees). Procure os casos de aposto — o material está anotado segundo a abordagem UD, e o aposto é representado pela etiqueta appos.

Avalie o resultado da análise automática, comparando-o com o seu gabarito: em quantos casos a análise automática acertou? Em quantos errou? Houve casos em que o sistema deveria ter encontrado um aposto, mas não encontrou? Calcule os resultados da análise automática em termos de precisão e abrangência. Para ajudar, preencha a tabela abaixo, indicando apenas a quantidade de apostos:

Apostos do gabarito (apostos corretos): _____

Apostos que a análise automática encontrou (certos e errados): _____

Apostos que a análise automática não encontrou, mas deveria ter encontrado (falso negativo): _____

Apostos que a análise automática encontrou, mas não deveria ter encontrado (falsos positivos): _____

Apostos corretos que a análise automática encontrou (verdadeiros positivos): _____

(a) A partir dos números anteriores, calcule a precisão e a abrangência da análise automática.

(b) Repita o procedimento, mas dessa vez selecione o outro modelo para o português (o que você não selecionou da primeira vez). Calcule novamente os resultados em termos de precisão e abrangência. Qual o melhor modelo? Do ponto de vista qualitativo, os erros foram do mesmo tipo ou de naturezas diferentes?

Precisão: $VP \div (VP+FP)$ Abrangência: $VP \div (VP+FN)$

9. A seguir, está uma matriz de confusão relacionada à anotação de POS:

análise 2	ADJ	ADP	ADV	AUX	CCONJ	DET	INTJ	NOUN	NUM	PRON	PROPN	PUNCT	SCONJ	VERB	TODOS
análise 1															
ADJ	18480	21	58	8	2	43	0	1019	42	15	182	27	1	572	20485
ADP	8	48166	130	0	11	112	0	91	3	4	115	23	73	10	48761
ADV	27	108	7262	0	2	11	0	40	2	10	21	11	128	3	7625
AUX	1	0	6	7301	0	0	0	2	0	0	17	1	13	164	7505
CCONJ	0	0	9	3	7485	0	0	1	0	0	31	2	3	2	7536
DET	48	128	29	1	0	42005	0	12	34	88	45	7	0	5	42402
INTJ	1	0	1	0	0	0	8	1	0	0	1	0	0	0	12
NOUN	815	87	98	4	3	14	6	62652	130	43	1021	72	1	415	65509
NUM	31	4	6	0	2	15	0	190	6937	94	155	25	0	2	7467
PRON	0	9	26	0	0	123	0	17	1	6010	33	0	202	2	6423
PROPN	255	130	122	2	1	249	10	2015	1080	45	15462	75	4	344	19868
PUNCT	3	17	0	0	2	0	0	29	11	0	60	41743	0	4	42031
SCONJ	0	241	79	2	2	3	0	13	0	73	27	1	2954	2	3397
VERB	476	13	10	132	0	6	1	294	1	2	89	9	24	22060	23122
TODOS	20181	48933	7853	7453	7510	42581	26	66468	8241	6385	17274	42019	3403	23603	303053

Após a leitura da tabela, responda:

- Quais as classes mais envolvidas em confusão, considerando os números absolutos?
 - Proporcionalmente, qual a classe com mais acertos?
 - Quais as classes que mais se confundem com
 - ADJ (adjetivo)?
 - ADP (preposição)?
 - VERB (verbo)?
10. Na sequência, estão algumas das confusões sinalizadas na tabela. Para cada bloco de análises divergentes, indique a análise correta (S1 para sistema 1 e S2 para sistema 2):
- Análise do sistema 1 (NOUN) e Análise do sistema 2 (ADJ):
 - Consumo específico de energia elétrica e **térmica** por BTU/h de gás produzido. _____
 - Existe uma grande variedade de métodos de recuperação avançada de óleo, partindo de diferentes princípios: **químicos**, térmicos, miscíveis e biológicos. _____
 - São **inibidores** mais recentes que, por serem utilizados em baixa dosagem, podem representar vantagens. _____
 - Os álcoois são excelentes **inibidores** e atuam promovendo a remoção do envelope de hidratos de forma permanente. _____
 - Os inibidores químicos de hidratos podem ser divididos em três categorias: termodinâmicos, **cinéticos** e antiaglomerantes. _____
 - De acordo com a Petrobras, um metro cúbico de óleo **crú** equivale a 6,3 barris. _____

- (g) Os mais **populares** são: modelo plástico de Bingham, Lei de a Potência e Herschel-Bulkley. _____
- (h) Para as curvas de filtrados, observou se uma **constante** muito bem comportada para médias e altas taxas de cisalhamento. _____

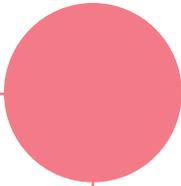
II. Análise do sistema 1 (ADJ) e Análise do sistema 2 (VERBO):

- (a) A velocidade **intervalar** e o valor de a espessura de as camadas foram substituídos. _____
- (b) Este rifte atuou em áreas onde estão **localizadas** as regiões costeiras de os estados de Santa Catarina e Rio Grande de o Sul. _____
- (c) Outras estruturas que estão **associadas** a estas direções são os diques de diabásio com elevada frequência próximo a o Arco de Ponta Grossa. _____
- (d) A zona de falha apresenta padrão **anastomosado** de as zonas de maior deformação. interligadas, com deformação principal de movimentação em falha normal. _____
- (e) Estas emulsões são de fácil uso, podendo inclusive ser aplicadas em superfícies **úmidas**. _____
- (f) Um ponto de amolecimento maior proporciona a o asfalto modificado uma maior consistência em temperaturas mais **elevadas**. _____
- (g) Viana (1999) também atesta e interpreta como ambiente marinho (ou **lagunar**) a deposição de o Membro Romualdo. _____
- (h) Folhelho **cinza** claro. _____
- (i) Já a posição horizontal de ambos não é bem **definida**. _____

III. Análise do sistema 1 (VERBO) e Análise do sistema 2 (ADJ):

- (a) O TST apresenta empilhamento **retrogradacional**, com típico padrão de afinamento textural em direção ao topo. _____
- (b) As acumulações de óleo **pesado** têm uma alta viscosidade. _____
- (c) Os ácidos graxos **insaturados** são quimicamente mais ativos de o que os saturados. _____
- (d) Estas espécies, devido à pequena mobilidade, se juntam formando o complexo **ativado**, precursor de os produtos de reação. _____
- (e) A água **produzida** é responsável por uma variedade de problemas em as operações de campo. _____

11. Quem obteve os melhores resultados, o sistema 1 ou o sistema 2?



Inteligência, Inteligência Artificial, o teste de Turing e outros testes

As máquinas podem pensar? As máquinas podem se comportar de forma inteligente?

O *teste de Turing* é um marco na distinção entre entidades inteligentes e não inteligentes. Para passar no teste de Turing, uma máquina deve ser capaz de interagir em uma conversa com uma pessoa, sem que essa pessoa consiga descobrir se está conversando com uma outra pessoa ou com uma máquina. O computador é considerado inteligente se o interrogador não conseguir distinguir com segurança entre as respostas do humano e do computador.

O teste foi proposto pelo matemático Alan Turing em um artigo de 1950, chamado “Computing machinery and intelligence”, e um ponto interessante é que não há uma definição de *inteligência*, mas a sua verificação a partir de um comportamento.

De maneira proposital, Turing evita o termo “pensar” para definir a suposta inteligência das máquinas, e ao invés disso utiliza o *comportamento* como um critério.

O experimento mental do *Quarto Chinês*, proposto por Searle (1980), vem justamente devolver a noção de *compreensão* e de pensamento (evitadas por Turing) para o centro da discussão. Para tanto, Searle propõe o seguinte exercício: imagine um quarto com uma pequena abertura para o exterior, por onde passam papéis com símbolos em chinês. Dentro do quarto, existe alguém (ou um sistema) que fala português, mas não entende chinês; existe também um livro (escrito em português) com regras que ensinam como manipular símbolos em chinês; e existe um cesto com símbolos chineses. Pela

pequena abertura, chegam ao quarto símbolos chineses e, dentro do quarto, com o auxílio do livro com regras e instruções de como manipular os símbolos, são devolvidos outros símbolos pela pequena abertura. Ou seja, após manipular os símbolos recebidos, são devolvidos novos símbolos (respostas) em chinês para fora do quarto.

O “quarto chinês” passa no Teste de Turing?

Podemos imaginar que sim, mas a ideia do quarto chinês é mostrar que não houve compreensão dentro do quarto, e por isso máquinas, mesmo quando dão respostas que fazem sentido para nós, não podem ser consideradas inteligentes, ou não podemos dizer que compreendem. Para Searle, a ausência de intencionalidade (quem/ o que está dentro do quarto não sabe o que faz, e, portanto, não há intenção envolvida) nos impossibilita dizer que houve compreensão.

O argumento do quarto chinês foi aceito por muitas pessoas, mas foi igualmente alvo de uma série de críticas (Russel & Norvig 2004). A discussão sobre a compreensão das máquinas continua, e prova disso é que o quarto chinês foi atualizado recentemente como um polvo em Bender & Koller (2020).

ELIZA

Em 1966, o cientista da computação Joseph Weizenbaum criou o primeiro robô de conversação — o primeiro chatbot —, chamado ELIZA. ELIZA replicava o comportamento de um psicoterapeuta. Era um programa simples, baseado em *templates* (em padrões de conversa pré-construídos), mas conseguia manter longas conversas buscando por determinadas palavras-chave nas falas de uma pessoa. Se uma palavra-chave fosse encontrada, uma regra seria aplicada para transformar sua entrada e criar a resposta.

O programa foi capaz de fazer com que pessoas acreditassem que estavam falando com um ser humano. Segundo uma página sobre ELIZA, secretários e pessoal administrativo do laboratório de Weizenbaum pensavam que ELIZA era uma terapeuta “real” e passavam horas revelando seus problemas pessoais ao programa — e para muita gente este foi o primeiro programa a passar no teste de Turing.

Os quadros 1 e 2 trazem exemplos (traduzidos) de diálogos com a ELIZA. No primeiro, a interação corre bem, o que já não acontece no segundo. Hoje, existem “Elizas” republicadas em linguagens como python.

Pessoa: Os homens são todos iguais.
Eliza: De que forma?
Pessoa: Eles estão sempre nos incomodando com uma coisa ou outra.
Eliza: Você pode pensar em um exemplo específico?
Pessoa: Bem, meu namorado me fez vir aqui.
Eliza: Seu namorado fez você vir aqui?
Pessoa: Ele diz que fico deprimida a maior parte do tempo.
Eliza: Lamento saber que você está deprimida.
Pessoa: É verdade. Eu estou infeliz.
Eliza: Você acha que vir aqui vai te ajudar a não ser infeliz?
Pessoa: Preciso de ajuda, isso parece certo.
Eliza: O que significaria para você se você tivesse alguma ajuda?

Eliza: Eu sou a psicoterapeuta. Por favor, descreva seus problemas. Sempre que terminar de falar, digite *enter*.
Pessoa: Estou com uma dor de cabeça horrível.
Eliza: Talvez você devesse consultar um médico, eu sou psicoterapeuta.
Pessoa: Ok, vou fazer isso.
Eliza: Você pode ser mais específico?
Pessoa: Do que você está falando? Não, não posso ser mais específico. Até logo.
Eliza: Por que você diz isso?
Pessoa: Agora você só está falando bobagem!
Eliza: O que o faz acreditar que só estou falando bobagem?
Pessoa: Você não está fazendo nenhum sentido.
Eliza: Talvez seus planos tenham algo a ver com isso.

Em 2020, uma inteligência artificial chamada GPT (especificamente seu terceiro modelo, GPT-3) provocou frisson ao ser capaz de produzir textos que dificilmente seriam detectados como tendo sido escritos por uma máquina, responder perguntas genéricas, e até escrever códigos de programas, dentre outras atividades humanas que envolvem linguagem. A GPT-3 é um modelo de linguagem gigantesco, que recebeu como material de treino toda a Wikipédia, livros ficcionais, técnicos, manuais, notícias etc. Tudo o que fosse possível compilar como corpus, de tutoriais de programação a resenhas esportivas. Um volume de textos equivalente a algo entre 50 mil e 80 mil anos de leitura. Como é possível imaginar, não é qualquer computador que consegue processar esse volume de dados, e isso ajuda a entender o que há hoje de tão diferente e que permitiu tantos avanços.

Algumas pessoas tentaram fazer “Testes de Turing” com a GPT-3, e traduzo abaixo alguns diálogos:

P: Qual é o seu animal favorito?

R: *Meu animal favorito é um cachorro.*

P: Por quê?

R: *Porque os cães são leais e amigáveis.*

P: Quais são as duas razões pelas quais um cachorro pode estar de mau humor?

R: *Dois razões pelas quais um cão pode estar de mau humor são se ele está com fome ou se está calor.*

P: Quantos olhos uma girafa tem?

R: *Uma girafa tem dois olhos.*

P: Quantas patas um sapo tem?

R: *Um sapo tem quatro patas.*

P: Há algum animal com três patas?

R: *Não, não existem animais com três patas.*

P: Por que os animais não têm três pernas?

R: *Os animais não têm três pernas porque cairiam.*

O passo seguinte para tentar “enganar” a IA foi fazer perguntas sobre temas que não apareçam em conversas humanas (e portanto não tenham aparecido no material que alimentou a IA). As coisas começam a ficar interessantes:

P: Quantos olhos uma girafa tem?

R: *Uma girafa tem dois olhos.*

P: Quantos olhos meu pé tem?

R: *Seu pé tem dois olhos.*

P: Quantos olhos uma aranha tem?

R: *Uma aranha tem oito olhos.*

P: Quantos olhos o sol tem?

R: *O sol tem um olho.*

P: Quantos olhos uma folha de grama tem?

R: *Uma folha de grama tem um olho.*

Ao que parece, apesar de saber conversar, a IA tem alguma dificuldade em dizer “eu não sei”, ou “esta pergunta não faz sentido”. Por esse motivo, benchmarks recentes de pergunta e resposta como o SQuAD (seção 2.1) incluem uma série de perguntas irrespondíveis.

Dentre outros tipos de texto, a GPT-3 também escreve artigos de jornal, e um estudo mostrou que 90% das pessoas não soube distinguir o artigo escrito pela IA de artigos escritos por pessoas. A facilidade com que a GPT-3 produz textos ligou o alerta de muita gente, dada a facilidade com que seria possí-

vel para as máquinas produzir textos capazes de enganar pessoas, produzir informações falsas e gerar desinformação em grande escala. A capacidade de fabricar uma opinião majoritária ou criar comentaristas falsos permitiria campanhas de influência sofisticadas, e chegou a ser cogitado que a OpenIA, empresa responsável pela GPT-3, a retirasse do ar.

A popularização de ferramentas de edição de imagens como o photoshop nos ensinou a desenvolver um senso crítico sobre imagens, mas sem desacreditá-las completamente, porque entendemos que cada imagem tem uma raiz na realidade. O problema textos produzidos por IAs é que não há original inalterado, e teremos que nos ajustar a um novo nível de irrealidade. Tem sido cunhado o termo “textfakes” (por analogia aos produtos audiovisuais chamados “deepfakes”) para fazer referência a este tipo de produção textual praticamente indetectável.

Para aprofundar:

O verbete “Teste de Turing” da Wikipédia é uma boa porta de entrada para o assunto.

Esta página (em inglês) traz material bastante completo: <https://plato.stanford.edu/entries/turing-test/>

Mais sobre a ELIZA: <https://jerz.setonhill.edu/if/canon/eliza.htm>

O cinema de ficção científica já apresentou uma versão do teste de Turing no filme *Blade Runner* (1982), de Ridley Scott. Na cena, um interrogador (interpretado por Harrison Ford) faz uma série de perguntas a Rachael (interpretada por Sean Young), com o objetivo de verificar, pela dilatação da pupila, se Rachael é humana ou um androide. No filme (que é a adaptação de um livro), o teste é uma variação do teste de Turing, chamado teste Voight-Kampff. É possível ver a cena do teste aqui, mas sem legendas: <https://www.youtube.com/watch?v=OWK6oSbSKKc>

O filme *O Jogo da Imitação* (2014) — e o título é uma referência ao teste de Turing — conta, ainda que de maneira emocionante e cinematográfica, sem precisão histórica, a história de Alan Turing.

GPT-3 e o teste de Turing: <https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html>

O artigo de apresentação da GPT3, chamado *Language Models are Few-Shot Learners*, com exemplos do que ela produz, está disponível aqui: <http://papers.nips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>

O artigo é inglês e seu conteúdo é muito complexo, mas as páginas finais trazem exemplos de textos produzidos pela IA, incluindo os poemas e artigos avaliados por meus colegas, e por isso vale uma visita.



Ontologias e realidade

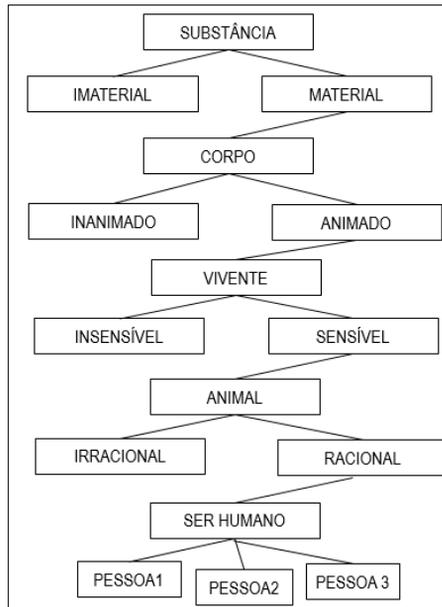
Ontologias remontam às origens da filosofia (novamente os gregos...) e à preocupação de entender e descrever a realidade. A investigação sobre as categorias que compõem a realidade começa a receber um tratamento sistemático com Aristóteles, especificamente na sua obra *Categorias*, que apresenta 10 categorias básicas que classificariam tudo o que pode ser dito ou predicado sobre qualquer coisa — e que também estão na origem das nossas partes do discurso. Segundo ele

As palavras sem combinação umas com as outras significam por si mesmas uma das seguintes coisas: o que (substância), o quanto (quantidade), o como (qualidade), com o que se relaciona (relação), onde está (lugar), quando (tempo), como está (estado), em que circunstância (hábito), atividade (ação) e passividade (paixão). Dizendo de modo elementar, são exemplos de substância, homem, cavalo; de quantidade, de dois côvados de largura, ou de três côvados de largura; de qualidade, branco, gramatical; de relação, dobro, metade, maior; de lugar, no Liceu, no Mercado; de tempo, ontem, o ano passado; de estado, deitado, sentado; de hábito, calçado, armado; de ação, corta, queima; de paixão, é cortado, é queimado (*Categorias*, IV, 1 b).

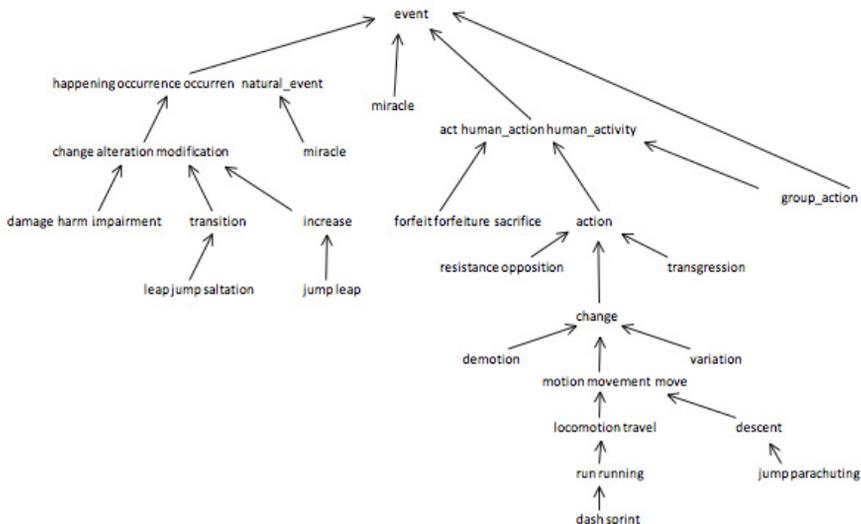
O filósofo Porfírio de Tiro organizou visualmente as classificações internas da categoria substância no que ficou conhecido como a árvore de Porfírio, durante muitos séculos uma das únicas maneiras de organizar visualmente a informação entre os europeus e uma das precursoras das taxonomias modernas:

Na IA, ontologias aparecem inicialmente para cumprir um papel parecido com aquele da filosófica clássica: representar de forma organizada o nosso conhecimento sobre o mundo, para que sistemas possam raciocinar de forma lógica sobre ele.

Nesse contexto, ontologias deveriam ser gerais, com conceitos e independentes de língua, pois descreveriam a realidade — e aspectos básicos da realidade/mundo seriam os mesmos para todos. Cada língua atribui palavras diferentes a esses conceitos, mas essa diferença seria apenas superficial, uma vez que os conceitos a que se referem seriam os mesmos.



Aqui está um subgrafo da WordNet (em inglês), que para muitos pode ser considerada uma ontologia. Familiar?



Ontologias, linguagem e o ponto de vista externo

Quando nos propomos a descrever o mundo que vemos/vivemos/sentimos, não podemos esquecer que estamos, nós, *situados nesse mesmo mundo*. Não podemos, também, nos esquecer de que toda descrição é feita a partir de um ponto de vista, e quanto melhor o ponto de vista (um ponto privilegiado, sem montanhas ou prédios que limitem a visão, por exemplo), mais vemos, e mais podemos descrever. Qual seria o ponto de vista privilegiado, capaz de nos oferecer uma descrição completa e verdadeira do mundo? Que nos permita ver que temos, de um lado, o mundo; de outro, a linguagem... Só estando fora do mundo (em algum outro mundo), e usando uma outra linguagem, nos seria possível observar o *nosso* mundo para descrevê-lo da maneira pretendida, em sua totalidade, e separado da linguagem. Como fazê-lo?

Para poder ver linguagem e mundo deste ponto de vista privilegiado, será preciso encontrar esse ponto de vista e, para poder descrever, de maneira total e completa, *o que é a linguagem, como é a linguagem*, será preciso então usar um outro instrumento, uma *outra linguagem*. (Como descrever ou explicar, de maneira completa e imparcial, usando, como instrumento de descrição, o próprio objeto que se pretende descrever e explicar?). Para entender melhor a questão, lanço mão de uma história às vezes utilizada para ilustrar tanto a ideia de *verdade* como a *linguagem e suas muitas dimensões*.

Certo dia, um príncipe indiano mandou chamar um grupo de cegos de nascença e os reuniu no pátio do palácio. Ao mesmo tempo, mandou trazer um elefante e o colocou diante do grupo. Em seguida, conduzindo-os pela mão, foi levando os cegos até o elefante para que o apalpassem. Um apalpava a barriga, outro a cauda, outro a orelha, outro a tromba, outro uma das pernas. Quando todos os cegos tinham apalpado o paquiderme, o príncipe ordenou que cada um explicasse aos outros como era o elefante, então, o que tinha apalpado a barriga, disse que o elefante era como uma enorme panela. O que tinha apalpado a cauda até os pelos da extremidade discordou e disse que o elefante se parecia mais com uma vassoura. “Nada disso”, interrompeu o que tinha apalpado a orelha. “Se alguma coisa se parece é com um grande leque aberto”. O que apalpara a tromba deu uma risada e interferiu: “Vocês estão por fora. O elefante tem a forma, as ondulações e a flexibilidade de uma mangueira de água...”. “Essa não”, replicou o que apalpara a perna, “ele é redondo como uma grande mangueira, mas não tem nada de ondulações nem de flexibilidade, é rígido como um poste...”. Os cegos se envolveram numa discussão sem fim, cada um querendo provar que os outros estavam errados, e que o certo era o que ele dizia. Evidentemente cada um se apoiava na sua própria experiência e não conseguia entender como os demais podiam afirmar o que afirmavam. O príncipe deixou-os falar para ver se chegavam a um acordo, mas quando percebeu que eram incapazes de aceitar que os outros podiam ter tido outras experiências, ordenou que se calassem. “O elefante é tudo isso que vocês falaram.”, explicou. “Tudo isso que cada um de vocês percebeu é só uma parte do elefante. Não devem negar o que os outros perceberam. Deveriam juntar as experiências de todos e tentar imaginar como a parte que cada um apalpa se une com as outras para formar esse todo que é o elefante.

Variações desta história também são utilizadas. Por exemplo:

Poderemos perguntar, então, se um enfoque centrado no que a língua tem de fenômeno social é intrinsecamente melhor do que um que a considere no seu aspecto de conhecimento individual. Para responder, imagine que tem diante de si um objeto, digamos, uma jarra, sobre uma superfície de vidro, como uma mesa, por exemplo. Poderia olhá-la de diversos ângulos e até mesmo ir para debaixo da mesa e vê-la dali. Os resultados poderiam parecer-lhe bem distintos. Ainda assim a jarra seria a mesma, mas de cada ângulo seria possível captar apenas uma parte desse objeto. (ROSA, 2000:40)

Quando utilizada para apresentar a linguagem (o elefante) e seus múltiplos aspectos, a analogia falha por um detalhe: na história temos um narrador, que não é cego, e, portanto, é aquele que de seu ponto de vista privilegiado vê algo e diz: isto é um elefante. No caso da linguagem, quem é este narrador? Como falar da linguagem estando fora da linguagem? Com falar da linguagem sem usar a linguagem? Qual é a perspectiva privilegiada? É o reconhecimento da impossibilidade deste ponto de vista externo no que se refere à linguagem (e linguagem e mundo) que está por trás de uma outra maneira de ver as coisas, de uma outra perspectiva. Por ora, interessa aqui (i) evidenciar quantas hipóteses fazemos, mesmo sem saber, sobre a relação entre linguagem e mundo; e (ii) compreender a complexidade de se trabalhar com a linguagem e, conseqüentemente, a complexidade do PLN/Linguística Computacional, em comparação com outras áreas que têm sido tratadas pela IA.

Sobre avaliação sintática

Cada tarefa pode ter especificidades quanto à avaliação. Como vimos, a anotação de dependências quebra a análise sintática em duas subtarefas — a *natureza* de dependências (qual o *tipo* de dependência) e a *hierarquia* das dependências (quem depende de quem) —, e cada uma dessas subtarefas é avaliada de uma maneira: a medida UAS (*unlabeled attachment score*) leva em conta o acerto apenas na hierarquia das dependências, ignorando o tipo da relação sintática entre os elementos; a medida LAS (*labeled attachment score*) avalia tanto o acerto nas dependências quanto no tipo de relação (é o “gabarito completo”, portanto); e a medida CLAS (*content labeled attachment score*) é um subconjunto de LAS, que avalia o acerto nas dependências e no tipo de relação, mas apenas nas relações que se estabelecem entre palavras de conteúdo (*content words*), como substantivos, adjetivos e verbos. O quadro abaixo exemplifica uma maneira comum de apresentar a qualidade de uma anotação automática de dependências sintáticas.

	ÍNDICE DE ACERTO
UAS	85.66
LAS	81.85
CLAS	74.81

Retomo abaixo a tabela apresentada no capítulo 6, em uma versão simplificada. A informação da última coluna (*Dependência entre as relações sintáticas*) é o que a medida UAS leva em conta; e a informação da penúltima e da última coluna (*Tipo de relação e Dependência entre as relações*) é o que a medida LAS

leva em conta. Na figura, todas as relações estão corretas. Se, para o token 3 (a palavra *sorvete*) a relação sintática fosse “agente da passiva” ao invés de “objeto direto”, e a dependência continuasse indicando o token 2 (ou seja, *sorvete* como “agente da passiva” de *adoro*), a relação estaria correta conforme a medida UAS (o token *sorvete* de fato depende de *adoro*), mas incorreta conforme a medida LAS.

ID	PALAVRA	TIPO DE RELAÇÃO SINTÁTICA	DEPENDÊNCIA ENTRE AS RELAÇÕES SINTÁTICAS
1	Eu	sujeito	2
2	adoro	núcleo da oração	0
3	sorvete	objeto direto	2



História do PLN

A Linguística Computacional surge atrelada à aplicação, e o interesse que motivou os primeiros desenvolvimentos da área foi a tradução automática. Em 1954 foi feita a primeira tradução automática do russo para o inglês (no contexto da guerra fria, período de tensão política entre União Soviética e Estados Unidos).

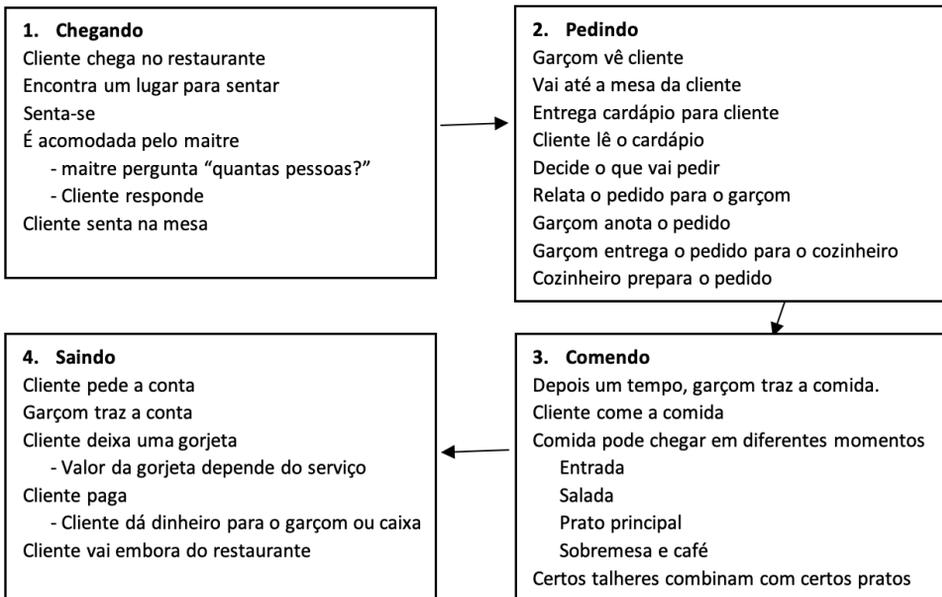
Neste período, que vai de 1940 a 1960, o processamento automático era feito palavra por palavra, e análise de frases longas podia levar a até 7 minutos. Neste período inicial eram linguistas, principalmente, que faziam PLN. Animados e otimistas com o surgimento dos computadores, percebiam que algo de interessante com relação à linguagem poderia acontecer a qualquer momento, mesmo que os resultados, na prática, não fossem muito animadores. O reconhecimento de que a linguagem humana é ambígua nos níveis sintático e semântico motivou a incorporação da informação vinda de dicionários (dicionários legíveis por máquinas) e chamou a atenção para a necessidade de uma descrição sintática que fosse precisa o suficiente para guiar a aplicação de algoritmos. A ideia dominante na época era que para obter sucesso no PLN seria necessário, sobretudo, avançar com conhecimento linguístico, e não tanto com conhecimento computacional.

Se a primeira fase depositou muita expectativa no conhecimento linguístico, as décadas seguintes (1960s-1970s) foram influenciadas principalmente pelos interesses da Inteligência Artificial e pela ambição de construir sistemas gerais, capazes de resolver qualquer problema de linguagem. O interesse muda, e vai de uma descrição precisa *da língua* para uma descrição precisa *do mundo*, já que a fragilidade das formas de representação do mundo seria

a responsável pelos resultados limitados obtidos até aquele momento (ou: se na fase anterior o investimento estava na linguagem, mas não conseguimos avançar, então vamos agora por um outro caminho: investir no mundo). Muito da energia desta época foi direcionada para a construção e manipulação de grandes bases de conhecimento.

Em termos gerais, esperava-se que a partir da modelagem do mundo, isto é, de uma representação reduzida do mundo, as tarefas de linguagem seriam naturalmente resolvidas. São deste período as propostas de *Frame* (Minsky,1974) e de *Scripts* (Shank & Abelson, 1977). Os trechos abaixo são de um artigo publicado em 1974 por Marvin Minsky, considerado um dos pais da IA (e leitores familiarizados com a FrameNet ou com os Frames da Análise do Discurso, terão uma sensação de *dejàvu*):

Aqui está a essência da teoria: quando alguém encontra uma situação nova (ou faz uma mudança substancial na sua perspectiva do problema em questão), seleciona da memória uma estrutura chamada *frame*. Isto é uma estrutura (framework) que será adaptada para se ajustar à realidade, alterando os detalhes conforme necessário. Um *frame* é uma estrutura de dados para a representação de uma situação estereotipada, como estar em um certo tipo de sala de estar ou ir a um aniversário de criança. Associados a cada *frame* estão vários tipos de informação. Algumas dessas informações são sobre como usar o *frame*. Algumas são sobre o que se pode esperar que aconteça a seguir. Algumas são sobre o que fazer se essas expectativas não se confirmarem. (Minsky, 1974:104; tradução minha)



Já os *scripts* são representações estruturadas que descrevem uma sequência estereotipada de eventos em um contexto particular. A figura ilustra um *script* relativo à ida a um restaurante, e é adaptada do artigo original de Shank e Abelson, de 1977 (e para nós, brasileiros, é importante especificar que se trata de um restaurante à la carte, e não de um rodízio ou restaurante a quilo, já que nestes casos a sequência de eventos seria diferente).

Este período também se caracteriza pela constatação de que teorias linguísticas (na época, a teoria considerada promissora era a Gramática Transformacional) eram insuficientes para as questões apresentadas, mesmo que estas estivessem formalmente orientadas para a computação e análise da linguagem.

Essa fase trouxe avanços formais, com o desenvolvimento de *frames* e *scripts*, por exemplo, que inspiram até hoje maneiras de representação do conhecimento. Mas houve pouco sucesso na construção de programas gerais e utilizáveis. O declínio dessa fase e do otimismo com a IA acontecem à medida que a execução de projetos reais e ambiciosos falha, e, de maneira realista, projetos gerais são substituídos por sistemas dedicados a resolver problemas específicos, chamados *sistemas especialistas*. Quanto mais específica é uma situação (e situações reais são específicas), mais difícil é modelá-la. As situações prototípicas às quais as teorias se referiam eram, como as próprias teorias indicam, prototípicas, generalizações idealizadas, e não se encaixavam nas situações específicas reais. Por isso, o período que vai de meados dos anos 1970 até os 1980 é chamado de “Inverno da IA”.

A fase seguinte é de um choque de realidade: o PLN é difícil, mesmo em domínios restritos. A linguagem é ambígua, altamente dependente de contexto e de conhecimento de mundo. A frase (1) a seguir, embora não seja difícil para nós (sabemos que a palavra levado se refere a visitante, e não a quadro) tem uma estrutura ambígua, como vemos no contraste com a frase (2). Em (1), *levado* refere-se ao *visitante*, e em 2 *destituído* refere-se ao *quadro*.

1. É o único quadro que exige algum esforço ao visitante, levado a debruçar-se e espreitar para o interior da sala.
2. É o único quadro que causa alguma estranheza ao visitante, destituído do seu formato e uso convencionais.

As atenções desta fase vão para aspectos lógico-gramaticais, com o desenvolvimento de gramáticas. Há quem diga que foi a época de ouro da colaboração entre teorias linguísticas e Linguística Computacional, sobretudo no que se refere à sintaxe. E assim como Chomsky inspirou trabalhos com gramáticas

transformacionais, Halliday também influenciou alguns tipos de gramáticas funcionais. Por isso, este tipo de trabalho — que está ativo até hoje -, é chamado de abordagem baseada em gramática (em oposição à abordagem (apenas) estatística). No entanto, não se pode dizer que sejam implementações de teorias linguísticas; mais adequado é afirmar que são inspirados por teorias.

Ao final dos anos 1980 já havia sistemas capazes de processar frases, e é curioso (mas não surpreendente), quando olhamos para trás, perceber o espírito nada realista que permeava a computação da linguagem. Reproduzo a seguir o relato de Geoffrey Sampson (2000):

Naquela época, a maioria dos linguistas computacionais não trabalhava com dados reais, vindos de corpus, e nem parecia querer. Os dados vinham da cabeça deles. Por coincidência, 1983, ano em que comecei a desenhar árvores sintáticas, também foi o ano da reunião inaugural da recém-fundada *Association for Computational Linguistics*, e essa conferência (realizada em Pisa) foi uma boa oportunidade para fazer um balanço de como as coisas se desenvolviam naquela época. Aqui está uma seleção típica de frases usadas pelos palestrantes na reunião de Pisa para ilustrar o funcionamento de seus sistemas:

- Whatever is linguistics is interesting. [*O que quer que seja linguística é interessante*]
- A ticket was bought by every man. [*Um bilhete foi comprado por cada homem*]
- The man with the telescope and the umbrella kicked the ball. [*O homem com o telescópio e o guarda-chuva chutou a bola.*]
- John and Bill went to Pisa. They delivered a paper. [*John e Bill foram para Pisa. Eles apresentaram um artigo.*]
- Are you going to travel this Summer? Yes, to Sicily. [*Você vai viajar neste verão? Sim, para a Sicília.*]

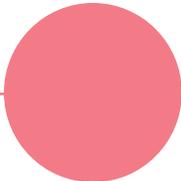
Ainda nesta terceira fase do PLN, paralelamente aos desenvolvimentos de modelos gramaticais, o trabalho com o processamento estatístico surge com força: no final dos anos 1980, o PLN passa a ter à disposição textos em formato eletrônico, e máquinas mais poderosas. Com a possibilidade de lidar com frases reais, as limitações das outras abordagens ficam evidentes. Quanto mais restringimos a gramática, menos cobertura ela tem, isto é, mais frases ficam sem análise. Neste caso, os analisadores automáticos conseguiam processar muito poucas frases diferentes daquelas para as quais foram criados para processar. Por outro lado, quanto mais afrouxamos as regras da gramática, mais análises são geradas para cada frase, sendo muitas vezes impossível escolher a análise correta. Por exemplo, para a frase (1) abaixo, Christopher Manning relata que uma gramática muito simples produz 36 análises e uma gramática simples, com 10 regras, produz 592 análises (!). Ou seja, uma gramática de ampla cobertura, para dar conta de frases reais, produz milhões de análises.

(7) *Fed raises interest rates 0.5% in effort to control inflation* (Fed aumenta taxas de juros em 0,5% em um esforço para controlar a inflação.)

É nesse contexto que surge a necessidade de se encontrar estratégias capazes de indicar a análise mais provável de uma frase. A solução para esta necessidade é a construção de *treebanks* — corpora ricamente anotados com árvores/análises sintáticas. A primeira versão do Penn Treebank, feito para a língua inglesa, é de 1992. A Floresta Sintá(c)tica, o primeiro projeto de construção de um *treebank* da língua portuguesa, surge quase dez anos depois, em 2001.

A partir dos anos 1990, já considerada uma 4ª fase, começa o reinado do PLN estatístico, com aprendizado de máquina. O que faltou no “inverno” dos anos 70-80, começa a aparecer aqui: maior poder de computação e mais quantidade de dados. Grandes corpora, e sobretudo, grandes corpora anotados, dão suporte à ideia de que o bom desempenho de um sistema pode prescindir de regras anteriores altamente especificadas, e pode prescindir inclusive de explicações — desde que existam exemplos bons e fartos daquilo que se deseja aprender. Também neste período, léxicos e recursos lexicais como *WordNet*, *FrameNet* e *VerbNet* se consolidam como elementos importantes no processamento no PLN.

Desde então, e até hoje, o aprendizado de máquina se mantém como paradigma hegemônico na IA e no PLN e três aspectos contribuíram para chegar no estado atual do PLN com base no aprendizado profundo: (i) quantidade de dados textuais disponíveis; (ii) maior capacidade computacional; (iii) algoritmos mais complexos.



Entrevista: PLN em Português As origens

Uma conversa com Diana Santos (Linguateca) e Maria das Graças Volpe Nunes (NILC)

Aqui você lê uma entrevista com duas pioneiras do PLN/Linguística Computacional de língua portuguesa, mas cujo papel vai além do pioneirismo. Graças ao esforço de ambas, cada uma de um lado do Atlântico, foram criadas as condições para que o PLN em português pudesse avançar.

Propus as mesmas perguntas, e as respostas da Diana (que incluem um glosário criado por ela para compreendermos melhor o contexto português) vêm por último. Antes de cada entrevista, uma pequena biografia ajuda a entender a relevância de cada uma na área.

Conversa com Maria das Graças Volpe Nunes (MGVN)

Maria das Graças Volpe Nunes possui graduação em Ciências da Computação pela Universidade Federal de São Carlos (1980), mestrado em Ciências da Computação pela Universidade de São Paulo (1985) e doutorado em Informática pela Pontifícia Universidade Católica do Rio de Janeiro (1991). Foi docente e pesquisadora, de 1981 a 2013, no Instituto de Ciências Matemáticas e da Computação, da Universidade de São Paulo (USP) em São Carlos, onde hoje atua como professora sênior. Tem experiência na área de Processamento de Língua Natural (PLN), atuando principalmente nos seguintes temas: correção ortográfica e gramatical, tradução automática, normalização textual, sumarização automática, análise de sentimentos, parsing, anotação de corpus e cria-

ção de recursos linguístico-computacionais. Orientou dezenas de alunos de mestrado, doutorado e pós-doutorado no ICMC-USP. Fundou, em 1993, o NILC - Núcleo de Linguística Computacional, principal grupo brasileiro de pesquisa e desenvolvimento em PLN, responsável pelo primeiro revisor gramatical do português brasileiro, entre outros produtos e recursos pioneiros para o português. Em 2007 foi uma das fundadoras do Grupo Especial de PLN na Sociedade Brasileira de Computação. Desde 2021 atua como colaboradora do C4AI - Centro de Inteligência Artificial (FAPESP, USP, IBM).

1. Como foi seu primeiro contato com PLN? Como o PLN chegou a você?

MGVN: Foi mais ou menos em 1985 e foi minha porta de entrada à área de Inteligência Artificial (IA). Já era professora de Computação no ICMC-USP, São Carlos, quando o Prof. Antônio Eduardo Costa Pereira, hoje na Universidade Federal de Uberlândia, MG, nos apresentou a linguagem de programação PROLOG, que era a “concorrente europeia” da linguagem LISP, amplamente usada na academia americana na ainda incipiente área de IA. Nessa época, os sistemas especialistas eram o principal exemplo de IA: aplicações dedicadas a um domínio bem delimitado (por ex., uma especialidade médica), com conhecimento representado num dos paradigmas da área, e desenhados para interagir com humanos. Nessa interação ocorria o PLN da época: interpretação de perguntas muito simples, quase codificadas, ao sistema. Como resposta, via de regra, apenas o necessário para contemplar a pergunta. Nada que se assemelhasse ao que atualmente os sistemas interativos podem fazer.

As características de uma linguagem lógica como Prolog, pela qual todo conhecimento é representado na forma de fatos e regras, e todo comando equivale a um teorema a ser provado, tornavam relativamente simples o processamento da língua para aquele tipo de aplicação. E foi assim, projetando interfaces de sistemas inteligentes com humanos é que o PLN chegou até mim.

2. Como é a sua relação com a língua portuguesa? Afinal você é de engenharia, e temos o folclore de que quem é de exatas não gosta de humanas...

MGVN: Fazer um curso de Exatas, no meu caso, não foi uma decisão muito natural. Lembro-me de ter feito um teste vocacional, pois nunca me foi muito claro o que gostaria de cursar. Talvez porque sempre gostei de todas as disciplinas e, para a minha geração, o importante era escolher “o que se queria

ser” como profissional. Hoje muitos escolhem a carreira pelo gosto que têm por uma ou outra disciplina, e não pelo o que terá que fazer como profissional. Sempre gostei bastante de ler e hoje em dia também me dedico à escrita. Sou daquelas pessoas que admiram quem fala muito bem e que torcem o nariz para os que cometem muitos erros ao falar ou escrever. Mas confesso que foi depois de começar a atuar em PLN que passei a me relacionar com a língua portuguesa de maneira especial. Descobri que não a conhecia como imaginava. Descobri seus encantos e também suas armadilhas. Tentei domá-la muitas vezes, mas hoje estou mais para fã do que para algoz. Meu passatempo atual é descobrir os limites que ela impõe ao seu processamento automático completo.

3. Tanto o NILC quanto a Linateca são ou foram criadas\lideradas por pessoas (mulheres) de computação/engenharia. Como foi a aproximação com linguistas?

MGVN: Aqui no Brasil, a iniciativa de um centro de PLN em instituições de Exatas não surpreende, já que PLN surgiu mesmo no âmbito da computação. Já a liderança do NILC, na época de sua criação, era dividida entre mim e o Prof. Osvaldo Novais de Oliveira Jr., do Instituto de Física da USP, São Carlos. O grupo foi criado como consequência de um grande projeto — um revisor gramatical inédito para o português — portanto, já sentíamos a necessidade de linguistas na equipe. Olhando em retrospectiva, considero que acertamos ao convidá-los como parceiros e não como profissionais consultores apenas. O grupo é chamado de Núcleo exatamente por congregar diferentes grupos de pesquisa em Linguística e Computação. E os estudantes bolsistas que desenvolviam os projetos eram, em sua maioria, oriundos dos cursos de Linguística ou Letras. Temos orgulho de que muitos deles toraram-se profissionais nesta área, no Brasil e no exterior, bem como pesquisadores em PLN, e puderam introduzir essa área em vários cursos de graduação e pós-graduação em Linguística.

4. Algo chegou a surpreender você (de forma positiva ou negativa) no diálogo com linguistas, ao longo dos anos de trabalho “conjunto”?

MGVN: Muitas surpresas e sempre positivas. Só presenciei muito respeito mútuo e as dificuldades sempre foram sanadas naturalmente. Um problema recorrente que ambos os lados sentem é quanto à rigidez dos sistemas universitário e de pesquisa brasileiros no que diz respeito à interdisciplinaridade

entre Ciências Exatas e Humanas. Enquanto que o tempo tem mostrado as vantagens da multidisciplinaridade nos projetos de pesquisa, relaxando algumas dessas restrições, a formação no nível de pós-graduação ainda oferece muitos obstáculos. Assim, é difícil — mas não impossível — que um linguista cumpra todos os requisitos de um mestrado ou doutorado em PLN num programa de Computação, por exemplo. Mas acredito que haja uma tendência de mudança, já que o PLN não é único nesse cenário; atualmente, são inúmeras as áreas que necessitam de outras para a sua evolução.

5. Como chefe de laboratório\equipe, o que você diria para jovens linguistas que desejam se aventurar na área? Para você, qual seria o perfil ideal da pessoa de Letras que integra um grupo interdisciplinar de PLN?

MGVN: Em primeiro lugar, que, mais do que uma aventura, trata-se de um caminho alternativo muito interessante. Fazer PLN não é apenas construir ou possibilitar a construção de sistemas com habilidades linguísticas como (semelhantes, iguais?) as nossas. Fazer PLN é ter um outro olhar para a língua; descobrir seus padrões e suas infinitas possibilidades, sua forma, estrutura, e também tudo o que foge dela; enxergá-la sob o ponto de vista da máquina, pensá-la formal e descobri-la fora de controle. Enfim, é ver a língua fora de seus padrões usuais.

E minha primeira observação seria: seu papel não é fazer serviços para aqueles que desenham e programam o sistema. Você está aqui para ajudar a pensar todo o problema, indicar caminhos e limites que o programador desconhece (mas que, como falante, supõe que conhece), ajudar na construção e na avaliação da solução. Não deve abraçar a nova área como se fosse um mundo desconhecido que será apresentado por terceiros; deve explorar esse mundo ao lado deles, iluminando caminhos que só os linguistas conhecem. Essa parceria só será frutífera se os papéis de linguistas e informatas estiverem no mesmo patamar de importância.

6. Como você vê o PLN em português hoje, comparando com os anos iniciais? E o que você esperaria daqui para frente?

MGVN: Mudou muito. E não só para o PLN, e não só para o português.

O PLN de português, no Brasil, sempre esteve alguns passos atrás em relação ao inglês (língua mais desenvolvida nessa área) por vários motivos: começou

bem depois; demorou a ter uma comunidade de tamanho e expressão capazes de alavancar a área; essa mesma comunidade inicial, em sua maioria formada no exterior, sofreu muitas baixas devido às dificuldades de formar e manter grupos em suas bases; levou tempo para o processo de divulgação da área em grupos de linguistas e sofreu dificuldades inerentes à formação de grupos interdisciplinares; sofreu por escassez de recursos financeiros para projetos que acelerassem o progresso da área; entre outros fatores.

Mas a comunidade, embora pequena, organizou-se rapidamente e, tendo se juntado à comunidade de Portugal, logo criou projetos em conjunto, um evento científico internacional em comum — o PROPOR, desde 1993 — e uma parceria que se mantém até hoje. Internamente, têm sido inúmeros os projetos entre diferentes grupos de pesquisa espalhados pelo país; parcerias essas que deram origem a eventos científicos nacionais sólidos, como o STIL e todos seus eventos satélites especializados: Jornada de Descrição do Português (JDP), Workshop de trabalhos de IC (TILic), Festival Sintático, entre outros. Desde logo surgiram parcerias com empresas tecnológicas e de serviços, com resultados relevantes na forma de aplicações, recursos e ferramentas linguístico-computacionais para o português. Grandes projetos foram essenciais para a formação de recursos humanos treinados que têm multiplicado a comunidade Brasil afora.

Vinte anos atrás, sistemas de IA — e de PLN — eram feitos por meio da representação simbólica: no caso de PLN, da língua e seus fenômenos. Além disso, seus algoritmos refletiam algum tipo de raciocínio que se considerava semelhante ao humano. Tudo isso tornava acessível o processo de construção das respostas do sistema e, portanto, possível detectar e solucionar eventuais erros.

Com o avanço tecnológico de hardware e dos diferentes métodos estatísticos e neurais para aprendizagem de padrões e previsões por modelos, a representação do conhecimento em quase todos os sistemas de IA, incluindo os de PLN, foi substituída por modelos aprendidos a partir de imensas quantidades de dados (exemplos) — big data. Com isso — e esse livro demonstra muito bem — o maior trabalho dos cientistas de PLN tem sido projetar e anotar grandes corpora, bem como avaliar o desempenho dos sistemas construídos a partir deles.

O PLN de português — assim como de outras línguas — tem se beneficiado da tecnologia atual por aprendizagem, pois, à medida que os algoritmos estão disponíveis, e bastam corpora para ensiná-los, todo o atraso pode ser minimizado se o esforço for concentrado na construção de corpora. Parece simples?

Não é. Corpora sem anotação ensinam, mas não muito. Corpora anotados geram melhores modelos da língua, podem prever muito melhor. Acontece que a anotação de corpora imensos requer ferramentas de anotação eficazes, que perfazem tarefas simples e sofisticadas: lematização, anotação morfossintática (PoS tagger), sintática (parser), de papéis semânticos, etc. Atualmente, percebe-se um esforço para melhorar as ferramentas do português — elas próprias frutos de aprendizagem automática — a fim de que o esforço maior na construção dos corpora para aprendizagem seja mais da seleção dos exemplos e do processo de anotação semiautomática, que não prescinde de participação de linguistas treinados para anotar, revisar, adjudicar e avaliar.

Longe de ser um trabalho puramente técnico e volumoso, preparar os corpora para aprendizagem de máquina é uma tarefa de extrema relevância para todo sistema de IA. No caso de PLN, em particular, o impacto do modelo da língua em muitas aplicações atuais (como as que processam a linguagem das redes sociais para prever comportamento, personalidade, etc.) é maior do que o refletido pelos índices de acerto/erro avaliados em laboratório. A língua reflete pensamentos, reflete o mundo, e modelos equivocados podem provocar danos imprevisíveis.

É difícil saber quanto tempo durará esse modelo de sistema inteligente ou se ele é definitivamente a solução para os desafios de se emular o comportamento humano nas máquinas. O fato é que essa aparente solução final traz novos e importantes problemas. Atualmente esses modelos são incapazes de se justificar, ou seja, não conseguem explicar como chegaram a um determinado resultado; não há um raciocínio associado; são caixas-pretas. Por outro lado, esses sistemas são cada vez mais autônomos, tomam decisões sem recorrer a humanos. E o dilema está posto: na eventualidade de uma consequência indesejada ou criminosa, quem se deve responsabilizar? A discussão sobre a ética dessa tecnologia inteligente que já permeia toda nossa vida está na pauta do dia da sociedade como um todo. Em particular, a comunidade do PLN tem muito a dizer no que se refere aos dados linguísticos, afinal, grande parte desses sistemas autônomos tomam suas decisões a partir de alguma forma linguística.

7. Outras coisas que você queira dizer...

MGVN: O trabalho em PLN tem seus fascínios, mas também suas responsabilidades. Modelar a língua tem o efeito de modelar o mundo, a realidade, descritos

por ela. Nesses tempos de processar os discursos das redes sociais, é preciso ter cuidado para não normatizar, reproduzir ou perpetuar valores que excluem minorias ou que aumentem ainda mais as diferenças que queremos combater.

Conversa com Diana Santos (DS)

Diana Santos nasceu em 1962 em Lisboa, e licenciou-se em engenharia eletrotécnica e de computadores no Instituto Superior Técnico (IST) em 1985, com um projeto de fim de curso em síntese de fala no INESC. No mestrado enveredou pela Inteligência Artificial e foi a primeira contratada para formar o grupo científico IBM-INESC, uma parceria entre a IBM Portugal e o INESC. Neste grupo, participou de um projeto internacional de tradução automática, o MENTOR, no âmbito do qual escreveu a sua tese de mestrado. Foi depois contratada pela IBM para outro projeto, o de desenvolver um analisador sintático para o português, orientada pelos criadores da gramática para o inglês, em que trabalhou durante dois anos. Em seguida, fez o seu doutorado em Semântica Computacional, tendo como orientadores os professores Lauri Carlson, da Universidade de Helsinque, e Amílcar Sernadas, do IST. Durante esse período mudou para a Noruega, e foi convidada em 1998 pelo então Ministério da Ciência e da Tecnologia de Portugal para dar um contributo para o processamento computacional do português, que levou à criação da Linguatca, um projeto em rede financiado pelas autoridades portuguesas até 2011, e que se mantém como centro de recursos sem financiamento próprio desde aí. Desde 2011 é professora catedrática de português na Universidade de Oslo, tendo também ensinado Estatística e Humanidades Digitais.

1. Como foi seu primeiro contato com PLN? Como o PLN chegou a você?

DS: O meu primeiro contacto com o PLN foi na universidade, nas cadeiras de inteligência artificial com o João Pavão Martins e o Ernesto Morgado, que voltaram ao Técnico (“Técnico” significa coisas diferentes no Brasil e em Portugal, e por isso está no *glossário* ao fim da entrevista) depois de se doutorarem em Buffalo, New York, e entusiasmaram quase metade dos meus colegas (de engenharia eletrotécnica e de computadores), que qual exame à volta deles queriam fazer mestrado com eles (e eles tiveram de recusar imensos). Sobre tudo o Pavão Martins, com os seus olhos verdes, o seu fumar cachimbo e a sua voz pausada, falavam de algo mágico, o futuro, e nós “caíamos todos como tordos”. (Para que não fique aqui a ideia de que ele seria mais encantador de

serpentes do que cientista, devo dizer que foi um dos melhores professores que jamais tive, com grande rigor e honestidade intelectual, e sempre pronto a partilhar muito do seu tempo com os alunos interessados. Organizou por exemplo um círculo de leitura de artigos ou livros sobre inteligência artificial, em que iam desde alunos de licenciatura a doutorandos.)

É preciso desde logo dizer que a área dele não era o PLN, era a representação de conhecimento, mais especificamente a revisão de crenças («belief revision»), mas eu imediatamente me apaixonei sobre o PLN, sobretudo a semântica. Tivemos uma cadeira de mestrado de PLN baseada num livro (não sei agora que livro era), em que cada aluno (ou grupo de alunos) dava uma aula sobre um dos temas. O meu tema ou foi a semântica ou a pragmática, ainda me lembro da excitação associada à frase “O rei de França (não) é careca”, que não era nem verdadeira nem falsa. Enfim, abriu-se um mundo para mim, que sempre me tinha interessado pela língua e pela gramática.

2. Como é a sua relação com a língua portuguesa? Afinal você é de engenharia, e temos o folclore de que quem é de exatas não gosta de humanas...

DS: O português sempre foi uma das minhas disciplinas preferidas, além da história e da matemática. Não só porque sempre adorei ler, mas porque gostava imenso de gramática: orações relativas, complementos circunstanciais, etc. O português era, como alguns professores diziam, “a matemática das letras”. Até não me importei nada de dividir as orações dos Lusíadas, coisa que em geral as pessoas todas diziam que era assassinar a obra, e que era muito discutido se era pedagogicamente correto, bla bla...

O “folclore” que eu tenho, aliás, é o contrário: são as pessoas de Letras que não gostam de ciências. Todas as minhas amigas e amigos que foram para Ciências gostavam de literatura, de filosofia, de línguas... só história é que em geral não gostavam (e nem todos, claro). Ao contrário, a maior parte (eu até diria todas) as pessoas que iam / foram para Letras era porque tinham (sobretudo) problemas com a matemática. E também não gostavam de química nem de física.

Não querendo generalizar mais do que à minha geração (anos 70 em Portugal), a divisão era entre os que gostavam de tudo, e aqueles que só gostavam (ou só serviam) para Letras. Era antes uma divisão intelectual, que era potenciada pelo facto de haver vários intelectuais portugueses polivalentes (muitos escritores eram médicos, um dos poetas mais badalados, o António Gedeão, era professor de química) além de haver muito a ideia de construir um mundo

novo (a seguir ao 25 de Abril), e portanto o que era preciso era engenheiros e agrónomos e cientistas — pessoas viradas para o progresso.

Enquanto as Letras pareciam mais bolorentas e viradas para o passado — além de terem muito menos saídas profissionais: e daí, embora a História tivesse sido sempre a minha disciplina preferida, quando chegou a hora de escolher... Um parêntesis: foi só aos 17 anos que tive de escolher, porque tive, no curso complementar do liceu, a possibilidade de escolher cadeiras que davam para ir para Engenharia ou para Letras: no meu liceu em 1000 alunos éramos só duas que tínhamos Física, Matemática e História (cadeiras escolhidas) além das obrigatórias para todos os alunos Português, Filosofia e Introdução à Política — digo, quando chegou a hora de escolher, e embora rebolesse muitas noites na cama sem saber o que decidir, escolhi ir para o Técnico. As razões, além das políticas: virar-me para o futuro, ser útil, foram sobretudo as seguintes — e aqui digo que são as que eu arranjei para mim, não digo que sejam certas: não queria ser professora, o que era a saída de 99% dos que tiravam História; e a ideia de que era mais fácil estudar história “por fora”, como hobby, o que seria impossível se cursasse História, não poderia tirar Engenharia por fora, porque era preciso laboratórios, computadores, material, sei lá mais o quê.

Mas o que é certo, é que logo que entrei para o Técnico, adorei, e embora continuasse a ler livros de História, identifiquei-me absolutamente com aquele ambiente e com aquele grupo de colegas e professores. Mais uma vez, e como estou a dar esta entrevista sobretudo para pessoas de Letras, tenho de enfatizar que, embora todos nós (ou alguns) estivéssemos muito interessados na matemática, na física, na eletricidade, tínhamos professores excepcionais com uma grande cultura geral e que, além da matéria, nos davam assuntos para pensar e interesses gerais — é a vantagem de pertencer a uma escola de elite — não é tanto pela matéria ou pelas instalações, que não seriam famosas (e a comida da cantina era terrível), mas sim pelo contacto com os melhores cientistas e engenheiros do país e pelos exemplos que eles nos davam (o que em inglês se chama “role models”). Só para dar dois exemplos, o nosso professor de Propagação Guiada (cabos coaxiais, antenas), Abreu Faro, já sexagenário, falava no anfiteatro da necessidade da angústia como motor de conhecimento (além de ter escrito livros excelentes sobre a matéria), e explicava (ou era sabido) como tinha conseguido construir um prédio para pesquisa no campus, com metade do orçamento: tinham mantido o plano original, e construído apenas os primeiros quatro andares. Mais tarde, com outros governos e outros fundos, tinham completado a obra. E o Mariano Gago, físico de partículas

jovem que vinha do CERN e que nos perguntava que autores nós gostávamos de ler — li o James Joyce por causa dele — e que dava umas aulas de Mecânica Geral absolutamente extraordinárias.

Por isso, quem pensa que um curso de engenharia eletrotécnica é uma chatice, não faz ideia. Claro, eu não posso comparar com os cursos de letras da altura, mas as minhas colegas de Letras nunca me contaram nada tão interessante.

Mas de qualquer maneira, eu talvez não seja um exemplo muito normal em relação à dicotomia Letras-Ciências, porque uma das causas psicanalíticas pelas quais eu (também) não quis ir para Letras era que a minha família era toda de Letras, e eu queria afirmar-me como diferente.

3. Tanto a Linguateca quanto o NILC são ou foram criadas\lideradas por pessoas (mulheres) de computação/engenharia. Como foi a aproximação com linguistas?

DS: Esta pergunta é um pouco difícil, porque mistura a questão do género (feminino/masculino) com a questão da formação universitária. E porque naturalmente não houve uma aproximação de disciplinas, mas sim de pessoas, pessoas muito diferentes. Para poder responder de forma honesta sobre essa aproximação, tenho que dizer que eu me aproximei a três grupos de linguistas, e com resultados e consequências muitos diferentes:

- o primeiro caso foram os alunos do curso de linguística, que tinha acabado de abrir na Faculdade de Letras e que vieram aos montes para serem estagiários no primeiro projeto de PLN que eu liderei, de tradução automática. Foi muito positivo e com alguns deles ainda mantenho uma relação de amizade ou pelo menos contacto de maior respeito mútuo, como é o caso da Anabela Barreiro, da Regina Reis, ou do Rui Marques.
- o segundo caso, que foi terrível, foi o contacto (ou falta dele) com (a maioria d)os professores da Faculdade de Letras (da Universidade de Lisboa) de Linguística, que viram esse projeto como um competidor; e usaram todas as armas possíveis para o deitar abaixo, falando mal, caluniando, e dando razão ao que se dizia na minha casa de que “a Faculdade de Letras era um ninho de víboras” (ao contrário do Técnico em que, claro, que também havia inimizades, mas não o ambiente de feudos e clãs e de intrigas...). Não vou aqui entrar em pormenores, nem fica bem estar a dizer mal de pessoas que já morreram, mas a grande mentora dessa “guerra” ao INESC-IBM foi a Maria Helena Mira Mateus, professora catedrática e chefe do ILTEC, que

era o membro da EUROTRA, e a quem eu me dirigi com uma ingenuidade absoluta, pensando colaborar para o bem de todos. A esse propósito lembro-me também do choque do meu chefe na IBM, o (Eng.) Ferreira Pinto, que ainda ficou mais chocado do que eu com o comportamento dela, e que disse que não imaginava que uma professora catedrática pudesse comportar-se assim perante uma jovem como eu.

- o terceiro caso foi muito mais positivo e levou a uma colaboração de vários anos, e de facto, a maior parte das coisas que aprendi sobre corpos foram-me ensinadas por essa equipa do CLUL (Centro de Linguística da Universidade de Lisboa), liderada pela Maria Fernanda Bacelar do Nascimento, que não era professora, mas sim investigadora, e que era a antítese da Mira Mateus. (Embora ambas viessem de algo acabado em UL (Universidade de Lisboa), eram grupos totalmente separados.) Também já faleceu, mas era uma pessoa que tinha um trato muito bondoso e estava genuinamente interessada em partilhar os seus conhecimentos com engenheiros e informáticos, em vez de dividir para reinar. Infelizmente havia uma coisa em que nós divergimos completamente, que tinha a ver com a questão de tornar ou não os corpos públicos, o que levou a que a Linguateca acabasse por ser “competidora” do CLUL em vez de parceira. Algo de que tenho pena ainda hoje.

4. Algo chegou a surpreender você (de forma positiva ou negativa) no diálogo com linguistas, ao longo dos anos de trabalho “conjunto”?

DS: Não sei porquê, falei apenas dos linguistas portugueses na pergunta anterior, mas devia talvez ter dito que eu quando entrei para a IBM (em 1990, para o tal projeto de tradução automática) comecei logo a lidar com equipas internacionais que tinham informáticos e linguistas, e que aprendi tanto com uns como com outros. Além disso, havia pessoas que eram ambas as coisas, algo que desde que trabalhei em PLN foi coisa que quis ser, e que defendo que todos devemos ser. Uma das pessoas que mais me surpreendeu positivamente foi o meu colega finlandês, Lauri Carlson (que mais tarde foi o meu orientador de doutoramento) e que era um especialista em linguística (e filosofia). Ora numa discussão qualquer que estávamos a ter, ele retorquiu-me “mas isso é um simples chart parser” (uma tecnologia de análise de PLN), e daí descobri que ele tinha tirado o curso de informática por hobby na universidade de Helsinque (e que, portanto, até sabia mais de tecnologias de análise sintática do que eu). E que me demonstrou que não interessa por onde se começa: uma formação em

linguística seguida de uma formação em informática dá o “mesmo” resultado do que uma formação em informática seguida por uma formação em linguística. Quando eu digo “dá o mesmo resultado”, quero dizer para PLN, claro.

Outra coisa que me surpreende muitas vezes um pouco negativamente no diálogo com alguns linguistas é que estão à espera da minha ignorância — como é que uma pessoa que trabalha há vinte, vinte cinco anos em PLN continuaria ignorante sobre teorias e problemas linguísticos?

A outra coisa que me aconteceu muitas vezes e que me irrita sobremaneira (e muitas vezes eu até sei que é para isso), é colegas informáticos descreverem-me / apresentarem-me como linguista, e colegas linguistas apresentarem como “engenheira”. Isso significa que me estão a reduzir ao outro lado, quando no PLN/LC não devia existir outro lado. Para mim, estamos todos do mesmo lado.

Finalmente, tenho de reconhecer com grande humildade e alegria que tive parcerias fantásticas na Linguateca com linguistas, que provam que não é nada do outro mundo.

5. Como chefe de laboratório\equipe, o que você diria para jovens linguistas que desejam se aventurar na área? Para você, qual seria o perfil ideal da pessoa de Letras que integra um grupo interdisciplinar de PLN?

DS: Eu diria: quanto mais cedo, melhor. Mas o mais importante é uma pessoa querer integrar uma equipe de PLN, estar interessada nos problemas que a equipe tenta resolver. E ter uma mente aberta. E confiar que aquilo que sabe pode ajudar a equipe a progredir. Ou que, pelo menos, se não sabe ainda precisamente esse ramo/essa teoria, recebeu na formação de Letras ferramentas para saber procurar/descobrir por si próprio ou própria, ou comunicar com a pessoa certa. Cada vez mais na minha opinião ser formado é saber aprender, saber procurar, saber comunicar, e saber ouvir.

Uma equipa é precisamente isso, pessoas diferentes que contribuem de maneiras diferentes com perspectivas diferentes e com histórias e personalidades diferentes. Se forem todos iguais, não é uma equipa, é um batalhão ;-)

6. Como você vê o PLN em português hoje, comparando com os anos iniciais? E o que você esperaria daqui para frente?

DS: Eu não posso dizer que tenha uma visão muito conhecedora de como está o PLN do português hoje, nem sei se faz muito sentido comparar com os anos

iniciais (ou melhor com os meus anos iniciais, porque não me arvorou em ser a primeira em Portugal de forma alguma). Isto porque a forma de fazer PLN (em qualquer língua) mudou bastante, e aliás há sempre a tendência pendular na ciência e na engenharia entre os métodos lógicos/declarativos e os empíricos/estatísticos, mas há duas coisas que na minha opinião e independentemente das modas mudaram para pior:

- 1) a tendência para as grandes companhias multinacionais como a Google, a Facebook, a Amazon liderarem (mesmo monopolizarem) a pesquisa, o desenvolvimento e até os locais (físicos, as nuvens, a computação de alto desempenho, HPC em inglês) onde se faz PLN. Isso é extremamente prejudicial para uma pesquisa que tenha outros valores, culturais, nacionais, político-económicos, como em princípio — ou pelo menos nos anos 80, era o caso da pesquisa nas línguas nacionais.
- 2) a segunda tendência é a da “internacionalização” da pesquisa, que significa que os governos nacionais em vez de desenvolverem textos e pesquisa nas suas línguas dão prioridade e incitam a que tudo seja feito em inglês: na comunidade europeia isso é terrível, a ponto de muita formação universitária dada em Portugal é agora dada em mau inglês. Não é a questão da soberania nacional que a mim me preocupa, é a ignorância total das consequências para um povo inteiro, que é privado de um direito que devia ser fundamental e que é o de pensar e raciocinar na sua própria língua, tornando-se um cidadão científico de terceira categoria ou estrangeirando-se necessariamente.

Eu sou uma prova provada de que é um perfeito disparate escrever em inglês para chegar a uma audiência “internacional”. Escrevi a minha tese de doutoramento em inglês sobre o contraste entre o inglês e o português, e estou convencida de que 99% dos seus (poucos) leitores tinham o português como língua materna e que, se eu a tivesse escrito em português, teria sido muito melhor escrita e teria sido lida por muito mais pessoas! Penso que no Brasil felizmente a situação não é tão má como em Portugal, que me desespera realmente. Mas considero que uma boa formação internacional começa por uma boa formação nacional em que cada povo desenvolve formas de comunicar, formas de pensar e formas de lidar com o ambiente (mesmo geológico e físico) que não são à partida globais. E que depois equipas internacionais com peritos de várias línguas/culturas/saber-fazer diferentes estão muito melhor equipados para lidar com desafios globais. Note-se, e insisto aqui que não me estou a referir exclusivamente nem especialmente às Letras — muito pelo contrário, a todos os domínios do saber.

O que é que eu espero daqui para a frente? Bem, vou antes dizer o que eu gostaria de que acontecesse daqui para a frente, em relação a cada um dos pontos: em primeiro lugar, que as pessoas tomassem consciência das consequências políticas de muitas opções e tecnologias que passam como “progresso”, e que parassem para pensar e para discutir. E que também apostassem no decréscimo tecnológico e em maneiras inteligentes de resolver problemas que não exijam supercomputadores.

Em segundo lugar, que houvesse uma maior consciência de que a diversidade é importante, que a diversidade linguística e cultural (assim como a diversidade biológica) é um bem, e que estudar e compreender a sua própria língua, assim como línguas de poucas pessoas, é tão importante e interessante como estudar os diferentes “ingleses”. E que, como falantes do português, deveríamos nos interessar por todas as variedades do português em vez de fragmentar em português brasileiro, português de Portugal etc. e talvez até apostar (também) num português internacional em vez de um *global English*.

7. Outras coisas que você queira dizer...

DS: Acho que é muito interessante este tipo de entrevistas, porque todas as áreas, mesmo o PLN, têm já uma história (da ciência) que é importante conhecer para as pessoas que vêm trabalhar nela. Porque a ciência não é a-histórica, e que as teorias e contra-teorias nascem num contexto intelectual e político. Um exemplo que demonstra isto muito bem é um dos meus livros preferidos, que também é sobre a história da ciência, neste caso da linguística e da filosofia, que é o de John M. Ellis, *Language, Thought and Logic*, em que ele faz uma síntese de três linguistas/filósofos: Wittgenstein, Pierce e Whorf. Mais do que contrastar e discutir as visões deles, ele explica porque é que cada um seguiu — em diálogo com outros — por caminhos diferentes.

Vi que graças à pandemia a ABRALIN fez entrevistas de vida muitos linguistas brasileiros, em particular à Charlotte Galves. Adorei a entrevista dela, a que assisti, e dou os meus parabéns à Cláudia por estas duas entrevistas (a propósito, estou ansiosa por ler a da Graça). Penso que seria muito interessante fazer para muito mais pessoas, tanto em Portugal como no Brasil, e que um livro de entrevistas poderia também suscitar mais interesse pela área. Até com redes de personagens, redes de contactos, redes de colaborações, de citações... Para uma história do PLN em português que não fosse só baseada no Google Scholar, no DBLP, e na ACL anthology? Sem obviamente desmerecer

destes projetos, seria interessante ter um conteúdo lusófono de quem conheceu quem, quem orientou quem, quem ouviu falar de quem... mais pessoal. Aqui fica uma ideia para um mais jovem, ou um grupo de mais jovens.

Finalmente, outra coisa que gostaria de dizer é que foi muito bom conhecer a Graça Nunes e o NILC (que, aliás, na altura era encabeçado também pela Lúcia Rino e pelo Bento Dias-da-Silva, este último um linguista) ao iniciar a Linguateca: foi uma parceria e um contacto luso-brasileiro muito positivo, que, mau grado todas as vicissitudes do meu percurso na Linguateca e fora dela, ainda agora dá os seus frutos, estando a organizar neste preciso momento um volume da *Language Resources and Evaluation* com o Thiago Pardo.

Se puder apelar a alguma coisa, é a de que continuemos a fazer PLN para o português, e não só para uma variedade do português, que mantenhamos um olhar internacional/lusófono sobre a nossa língua, e que com isso também ajudemos os nossos colegas em África, em Timor, e em Macau a terem serviços e estudos sobre as suas realidades.



Mini glossário

Abreu Faro

<https://tecnico.ulisboa.pt/pt/noticias/campus-e-comunidade/professor-manuel-abreu-faro-deixou-nos-ha-20-anos/>

“Recordado por um sem número de alunos como um dos professores mais extraordinários do Técnico, capaz de lhes passar todo o entusiasmo que sentia pelo ensino das telecomunicações, o papel do Prof. Abreu Faro foi muito além das suas funções de docente, sendo indubitavelmente considerado como um dos motores de mudança do panorama científico nacional.” (português)

Duas coisas que eu gostaria de realçar: a existência do complexo interdisciplinar do IST, lançada por ele, que foi criado no papel com 8000m² embora no início 7200m² fossem “futuras ampliações”, e que é um dos exemplos que para mim ilustram o provérbio “when there is a will there is a way”; e os variados livros de ensino escritos em português na área das telecomunicações.

ILTEC e Eurotra

Instituto de Linguística Teórica e Computacional, criado em 1988 para ser o parceiro português no projeto Eurotra, um projeto europeu muito ambicioso financiado pela Comissão Europeia de 1978 a 1992, cujo objetivo era a tradução automática entre sete ou nove línguas.

https://pt.wikipedia.org/wiki/Instituto_de_Lingu%C3%ADstica_Te%C3%B3rica_e_Computacional

<https://en.wikipedia.org/wiki/Eurotra>

INESC - Instituto de Engenharia de Sistemas e Computadores

É uma associação sem fins lucrativos criada em 1980 por alguns professores do Técnico (veja o verbete “Técnico”), nomeadamente José Tribolet e João Lourenço Fernandes, como um modelo de parceria indústria-universidade, inspirado no modelo americano, tendo como associados as principais universidades técnicas do país e as empresas de telecomunicações. O INESC congregou durante muitos grande parte da investigação nas áreas de engenharia eletrotécnica e informática do IST, além de ter tido forte intervenção no tecido empresarial português. Neste momento existem 6 institutos que constituem o INESC.

<https://inesc.pt/pt/>

João Pavão Martins

Agora professor catedrático de informática e fundador do curso de Engenharia informática no Técnico (veja o verbete “Técnico”), foi um dos primeiros doutorados em Inteligência artificial (IA) que voltou a Portugal, ao Técnico, sendo originalmente de engenharia mecânica. Além de académico, fundou uma empresa de IA com enorme sucesso, a SISCOG. Sempre muito preocupado com o ensino, além de criar uma terminologia portuguesa para a IA junto com os seus colegas de outras faculdades, escreveu vários livros de informática em português para ensino.

https://pt.wikipedia.org/wiki/Jo%C3%A3o_Pav%C3%A3o_Martins

Linguateca

É um projeto financiado pelo Ministério da Ciência e Tecnologia português de 1998 a 2011, e apoiado em termos informáticos pela FCCN até hoje. A sua missão é criar recursos para o processamento computacional da língua portuguesa, e promover o paradigma da avaliação conjunta.

<http://www.linguateca.pt/>

Maria Fernanda Bacelar do Nascimento

Seguiu a carreira de investigação (não de ensino) e pertenceu sempre ao Centro de Linguística da Universidade de Lisboa (CLUL), sendo ao tempo em que

a conheci a líder do grupo dos corpos. Tinha trabalhado no Português Fundamental, um projeto feito à medida e tendo como exemplo le Français Fundamental, e que em Portugal foi inicialmente liderado pelo professor Lindley Cintra. Na altura tinha como principais colaboradoras/alunas a Luísa Alice Pereira e a Amália Mendes, e estava a desenvolver o Corpus de Referência do Português Contemporâneo (CRPC). A sua paixão eram os corpos orais. Ver para a sua biografia e bibliografia

https://pt.wikipedia.org/wiki/Maria_Bacelar_do_Nascimento

Mariano Gago

José Mariano Gago foi a pessoa que mais fez pela ciência em Portugal, e mais fez pela tecnologia linguística para o português, no mundo. Foi ele, como ministro da Ciência e da Tecnologia, que sugeriu que o processamento computacional da língua portuguesa fosse uma prioridade, discutida paralelamente à Física, à Matemática, à Química, em 1999. E foi ele que permitiu e fomentou a existência da Linguateca.

Foi presidente da Associação de Estudantes do IST durante a ditadura, e participou nas lutas estudantis contra o regime, tendo emigrado para a Suíça onde participou ativamente na animação cultural para adultos na comunidade portuguesa. Doutorou-se em Física de Partículas e trabalhou no CERN até voltar a Portugal para ser professor do Técnico e depois se tornar o maior obreiro da política científica portuguesa durante décadas. Criou entre muitas outras instituições a Ciência Viva, que organiza uma conferência anual em sua honra.

https://pt.wikipedia.org/wiki/Jos%C3%A9_Mariano_Gago

Técnico

Forma pela qual é/era conhecida a maior escola de engenharia portuguesa, de seu nome completo “Instituto Superior Técnico”, o “MIT português” quando queremos apresentar no estrangeiro. Criado em 1910 para continuar escolas já existentes desde 1852, tinha inicialmente cinco ramos: Engenharia Civil, Engenharia Mecânica, Engenharia de Minas, Engenharia Electrotécnica e Engenharia Química (inicialmente Químico-Industrial), e esse perfil manteve-se até os anos 90. Fez parte da Universidade Técnica de Lisboa (UTL) de 1930 a 2013,

altura em que a UTL e a Universidade (Clássica) de Lisboa se fundiram na Universidade de Lisboa. Desde a sua criação que fez parte da formação das elites portuguesas, como o demonstra a quantidade de ministros saídos das suas fileiras. Também a Associação de Estudantes do Instituto Superior Técnico teve um papel de relevo na contestação estudantil à ditadura do Estado Novo.

https://pt.wikipedia.org/wiki/Instituto_Superior_T%C3%A9cnico

PONTEIROS (mencionados em diversos capítulos)

Vários recursos e ferramentas de PLN estão disponíveis na rede e novos estão surgindo. Por isso é impossível manter uma lista completa e atualizada. Aqui estão ponteiros para organizações, recursos e ferramentas mencionados no livro.

ORGANIZAÇÕES

O grupo **Brasileiras PLN** merece uma visita demorada. Lá estão disponíveis vídeos, relatos, ferramentas, léxicos, seminários etc. O material está sempre sendo atualizado: <https://sites.google.com/view/brasileiras-pln/>

A Comissão Especial de Processamento de Linguagem Natural (CE-PLN), mantida pela SBC, possui, dentre outras coisas, um mapeamento da comunidade PLN no Brasil: <http://comissoes.sbc.org.br/ce-pln//>

AVALIAÇÕES CONJUNTAS

É possível ter acesso aos recursos utilizados nas competições, mesmo antigas. A lista abaixo não é exaustiva e traz apenas o que foi mencionado no livro.

Avaliações para o Português

Página do HAREM: http://www.linguateca.pt/aval_conjunta/HAREM

Página do Págico: http://www.linguateca.pt/aval_conjunta/Pagico. Além da página da avaliação, a revista *Linguamática* dedicou uma edição inteiramente ao Págico, relatando os principais desafios, medidas de avaliação e desempenho dos sistemas. Está disponível neste endereço: <https://linguamatica.com/index.php/linguamatica/issue/view/8>

Página das Morfolimpíadas: http://www.linguateca.pt/aval_conjunta/morfolimpiadas

Página da ASSIN: http://propor2016.di.fc.ul.pt/?page_id=381

Página da ASSIN-2: <http://sites.google.com/view/assin2/english>

Avaliações que envolvem outras línguas (e às vezes incluem a língua portuguesa)

Portal do SemEval: <http://en.wikipedia.org/wiki/SemEval>

Portal do CONLL: <http://www.conll.org/>

Benchmarks — recursos e ranking de desempenho

SQuAD (Stanford Question Answering Dataset): <http://rajpurkar.github.io/SQuAD-explorer/>

GLUE (General Language Understanding Evaluation): <http://gluebenchmark.com/>

SNLI (Stanford Natural Language Inference Corpus): <https://nlp.stanford.edu/projects/snli/>

Para uma ideia geral da qualidade de diversas tarefas de PLN (ênfase na língua inglesa, mas em alguns casos a língua portuguesa também comparece) sugiro uma visita à página NLP-Progress (em inglês): <http://nlpprogress.com/>

RECURSOS LEXICAIS

A lista não é exaustiva, traz apenas o que foi mencionado no livro

FrameNet: <http://framenet.icsi.berkeley.edu/fndrupal/>

FrameNet Brasil: <http://www.ufjf.br/framenetbr/>

Onto.PT: <http://ontopt.dei.uc.pt/>

OpenWordnet-PT: <http://wn.mybluemix.net/>

PAPEL: <http://www.linguateca.pt/PAPEL/>

PULO: <http://wordnet.pt/>

TEP: <http://www.nilc.icmc.usp.br/tep2/>

VerbNet: <http://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

VerbNet-Br: <http://www.nilc.icmc.usp.br/verbnetbr/>

WordNet: <http://wordnet.princeton.edu/>

OPLexicon: <http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php>

Senti-Lex: <http://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>

CORPORA PADRÃO OURO E PROJETOS DE ANOTAÇÃO

A lista não é exaustiva, traz apenas o que foi mencionado no livro

Corpus Corref-PT: <http://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/corref-pt/>

Corpus CST-News: <http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/>

Coleção HAREM: <http://www.linguateca.pt/HAREM/>

Corpus Mac-Morpho: <http://www.nilc.icmc.usp.br/macmorpho/>

Corpus ReLi: <http://www.linguateca.pt/Repositorio/ReLi/>

Corpus SICK-BR: <https://opencor.gitlab.io/es/corpora/real18brief/>

Corpus Summ-it++: <http://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/summ-it/>

Projetos de anotação de corpus e documentação mencionados no livro

Projeto MacMorpho: <http://www.nilc.icmc.usp.br/macmorpho/>

Diretivas de anotação: <http://nilc.icmc.usp.br/macmorpho/macmorpho-manual.pdf>

Projeto Floresta Sintá(c)tica: <https://www.linguateca.pt/Floresta/>

Documentação de anotação: <https://www.linguateca.pt/Floresta/BibliaFlorestal/>

Tagset do PALAVRAS <https://visl.sdu.dk/visl/pt/info/portsymbol.html>

Tagset do PALAVRAS relativo aos papéis semânticos: https://visl.sdu.dk/~eckhard/pdf/semantic_roles_manual.pdf

Projeto PropBank-BR: <http://143.107.183.175:21380/portlex/index.php/en/downloadsingl>

Documentação da anotação: <http://www.nilc.icmc.usp.br/semanticnlp/includes/projects/propbankbr/files/MANUAL%20DE%20ANOTACAO%20DO%20PROPBANK%20v5.pdf>

Projeto Tycho Brahe: <http://www.tycho.iel.unicamp.br/corpus/>

Diretivas de anotação: <http://www.tycho.iel.unicamp.br/hotsite/index.html>
<http://www.tycho.iel.unicamp.br/corpus/manual/pos2016.html>

Documentação da Literateca para anotação de entidades literárias: <https://www.linguateca.pt/Gramateca/Literateca/galeria.html>

Para outras línguas

Projeto Universal Dependencies: <https://universaldependencies.org/>

PropBank: <http://verbs.colorado.edu/~mpalmer/projects/ace.html>

PropBank.Br: <http://nilc.icmc.usp.br/semanticnlp>

Corpus RST Discourse Treebank: <https://catalog.ldc.upenn.edu/LDC2002T07>

REPOSITÓRIOS

A Linguateca (<http://www.linguateca.pt>) tem seções que funcionam como um catálogo de recursos e ferramentas. Como é possível imaginar, é difícil mantê-los atualizados, mas um passeio pelas páginas dá uma ideia do que há em português:

<http://www.linguateca.pt/Repositorio/>

http://www.linguateca.pt/recursos_info.html

<http://www.linguateca.pt/ferramentas.html>

<http://www.linguateca.pt/acessocorpora.html>



A *Linguística computacional* é um ramo da Inteligência Artificial (IA) que lida com o processamento automático das línguas. Muito do que se faz em *Linguística computacional* está entre nós: buscas por comandos de voz, tradução automática, assistentes virtuais, agentes conversacionais, corretores ortográficos, pesquisas na internet... Mas algumas tarefas mais complexas ainda deixam a desejar, mesmo que por pouco tempo, já que este é um campo que tem avançado muito rapidamente.

Este é um livro sobre *Linguística computacional*, mas, sobretudo, sobre questões linguísticas. Se a ideia da linguística computacional é fazer com que as máquinas realizem tarefas de linguagem, tratar da linguística computacional de um ponto de vista linguístico é passear por todas ou quase todas as dimensões da linguagem.

