

POeTiSA

POrtuguese processing - Towards Syntactic Analysis and parsing

Coordenador:

Thiago Alexandre Salgueiro Pardo

Magali Sanches Duran

15/04/2021



**Center for
Artificial
Intelligence**

Parceria USP, FAPESP, IBM

As atividades de pesquisas no C4AI estão organizadas em cinco Grandes Desafios:

NLP2 Recursos para Levar o Processamento de Linguagem Natural em Português para o Estado-da-Arte

KEML Aprendizado de Máquina Enriquecido por Conhecimento para Raciocínio sobre Dados Oceânicos

AgriBio Tomada de Decisão Causal Multicritério em Redes de Produção Alimentar

GOML Aprendizado de Máquina Orientado a Grafos para Diagnóstico e Reabilitação de AVCs

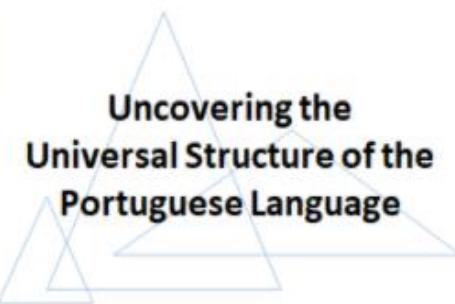
AI Humanity IA em Países Emergentes: Políticas Públicas e o Futuro do Trabalho

NLP2 <http://c4ai.inova.usp.br/pt/nlp2-pt/>

Três frentes:

1. POeTiSA <https://sites.google.com/icmc.usp.br/poetisa>
2. Modelos de distribuição e NLI (“Natural Language Inference”)
3. Treinamento de dois modelos em corpus de fala: um para identificação de locutor e outro para reconhecimento de fala

POrtuguese processing - Towards Syntactic Analysis and parsing



POeTiSA is a long term project that aims at growing syntax-based resources and developing related tools and applications for Brazilian Portuguese language, looking to achieve world state-of-the-art results in this area. On the resource side, we focus on the production of a large and comprehensive multi-genre corpus of [Universal Dependencies](#)-based part of speech and syntactically annotated texts, including mainly news texts and user-generated content (tweets and online comments). Regarding the tools, we aim to investigate recent neural and distributional-based methods for training robust parsing models for Portuguese. The project also envisions the production of applications on opinion mining and sentiment analysis tasks that may benefit from syntactic knowledge, as opinion summarization, helpfulness prediction, aspect identification, deception detection and emotion classification.

Objetivos do POeTiSA

- Treebank multigênero de tamanho significativo para o português
 - Filiado ao modelo *Universal Dependencies* (Nivre, 2015; Nivre et al., 2020)
- Sistemas de parsing completo e parcial para o português, com capacidade de análise multigênero
 - Avaliação e aprimoramento de parsers atuais (Straka, 2018; Zilio et al., 2018)
 - Investigação e desenvolvimento de novos métodos para o português, principalmente com base em abordagens neurais e distribucionais, com resultados do estado da arte
- Aplicação principal relacionada: análise de sentimentos
 - Tarefas que podem se beneficiar da sintaxe: extração de aspectos, detecção de conteúdo enganoso, classificação de utilidade de comentários, classificação de emoções, etc.
 - Avanços metodológicos: sumarização de opiniões, detecção de transtornos psicológicos, etc.

Montando uma fábrica de anotação...

PLANEJAMENTO

Decisões estratégicas

PRODUÇÃO

processo de anotação

RECURSOS HUMANOS

recrutamento, seleção, treinamento e
avaliação de anotadores

PESQUISA E DESENVOLVIMENTO

sistemas

PLANEJAMENTO

- Estabelecimento de metas
- Controle de resultados
- Escolha dos córpus

	Número de sentenças	Tamanho médio das sentenças	Número de tokens	Número de types
Folha-Kaggle	3.556.700	21,30	75.818.329	422.228
MAC-MORPHO	43.519	17,40	757.574	49.661
B2W-reviews1	238.567	12,60	2.995.379	55.919
Tweets_stocks	7.281	10,80	78.791	9.286
Comentários de livros	422	19,00	8.058	2.485
Total	3.846.489	16,22	79.658.131	539.579

RECURSOS HUMANOS

- Recrutamento de anotadores
- Treinamento de anotadores
- Avaliação de anotadores
- Reunião semanal de avaliação do processo de anotação e reforço do treinamento
- Elaboração de material de treinamento (constante)

RH - Material de suporte à anotação

- Manual de anotação de POS tags, rico em exemplos
- Manual de anotação de relações de dependência (em desenvolvimento)
- Lista de expressões fixas e semifixas e suas respectivas POS tags
- Lista de estrangeirismos e suas respectivas POS tags (em desenvolvimento)
- Estudos individuais de palavras ambíguas, com suas possíveis POS tags e respectivos exemplos

RH - Temas de dúvidas dos anotadores

nível de POS tagging

- Atribuição de SCONJ a preposições e advérbios (distinção entre SN e oração)
- Atribuição de POS tags segundo o contexto:
 - Preposição anotada como SCONJ
 - Adjetivo anotado como ADV
 - Pronome anotado como DET ou como PRON
 - Numeral anotado como PRON
 - Particípio anotado como ADJ, VERB ou NOUN
 - Etc
- Palavras ambíguas
- Expressões fixas

RH - PALAVRAS AMBÍGUAS: uma fonte constante de dúvidas

“porque”: SCONJ ou CCONJ?

Quando introduz oração subordinada adverbial de causa, é **SCONJ**

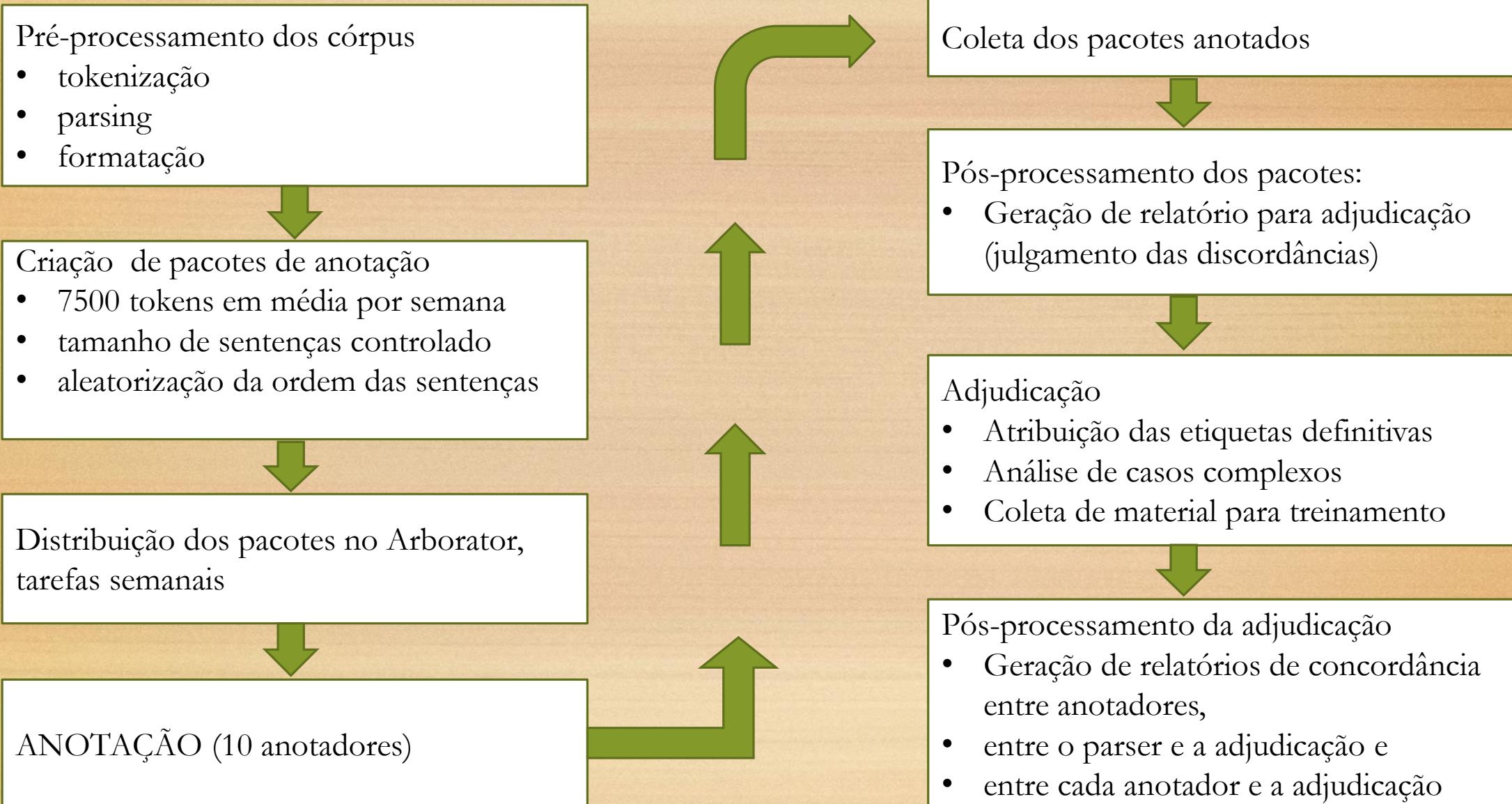
Eu não posso demonstrar raiva, **porque** tenho o lado profissional. FOLHA_DOC000_SENT000690

Quando introduz uma oração independente, é **CCONJ** (conj. coordenativa explicativa)

Porque todos viram um ensaio de discurso para a eventual saída de Doria. FOLHA_DOC000_SENT000977

Outra ainda é festejar, como Galvão Bueno, **porque** ganhar de eles “é sempre mais gostoso”. FOLHA_DOC000_SENT000733

PRODUÇÃO



PRODUÇÃO

DECISÃO: anotação “issue based” (início da anotação: 24/03)

- começamos anotando apenas POS tags, para simplificar a tarefa de anotação e o treinamento

RESULTADOS PRELIMINARES

- em sentenças de 7 a 20 tokens, **no corpus Folha-Kagle**, o parser erra, em média, 2,14% das POS tags
- muitos casos carecem de definição em nosso manual de anotação
- os anotadores têm diferentes níveis de conhecimento sobre categorias morfossintáticas
- os “erros” do parser são, em sua maioria, recorrentes, e poderão ser corrigidos de forma semi-automática

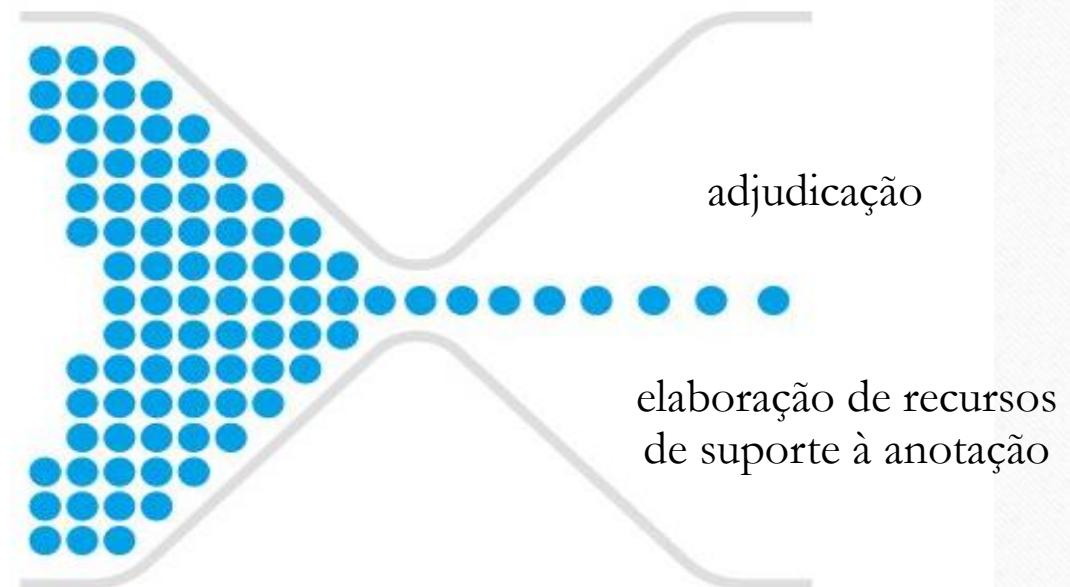
AÇÕES:

- enriquecer o material de suporte à anotação
- reforçar o treinamento nos pontos mais críticos
- listar erros recorrentes para abordagem de revisão em lote
- anotar sentenças maiores para observar se o desempenho do parser se mantém

Relatório para Adjudicação

Um gargalo temporário em nossa produção

10 anotadores,
500 sentenças
por semana



PESQUISA E DESENVOLVIMENTO

- Treinar modelos
- Automatizar processos
- Desenvolver novos tipos de relatórios
- Desenvolver/customizar ferramentas de anotação

Aprendendo com o Processo...

Questões críticas

Necessidade: ferramenta de anotação confiável, com salvamento automático, com log de atividades do anotador, com recurso de busca por palavra ou expressão ou etiqueta e possibilidade de substituição automática e semi-automática.

Tomada de decisões de anotação:

- Decidir casos não explícitos nas Guidelines da UD
- Decidir tokenização de formas híbridas
- Decidir questões específicas de UD em português (sozinhos?)

Decisões para anotação UD em português

- Verbos auxiliares
- Verbos de cópula
- Expressões fixas
- Elementos expletivos
- Diferença entre “core” e “obliques”

Se não obtivermos um consenso nas questões mais básicas, nossos códigos não vão conversar nem entre si, quanto mais com outras línguas!

Mãos à obra!

Questão para discussão: quando atribuir AUX?

- Quais verbos anotamos como auxiliares?
- Quais verbos de cópula devemos considerar? Só SER e ESTAR?

Tabela de Auxiliares

Temos uma lista, utilizada no projeto Propbank, com 67 verbos auxiliares: de tempo, modo, aspecto e diátese (voz passiva).

- Escolher os mais frequentes?
- Escolher só os de tempo, modo e diátese, seguindo o inglês?

VERBO	PALAVRA DE INTRODUÇÃO	FORMA DO AUXILIADO	TIPO DE AUXÍLIO	SEMÂNTICA
ir	-	infinitivo	temporal	futuro
ter	-	particípio	temporal	passado
haver	-	particípio	temporal	passado
ter	de	infinitivo	modal	deôntico
ter	que	infinitivo	modal	deôntico
conseguir	-	infinitivo	modal	epistêmico
dever	-	infinitivo	modal	deôntico/epistêmico
ficar	de	infinitivo	modal	deôntico
necessitar	-	infinitivo	modal	deôntico
poder	-	infinitivo	modal	deôntico/epistêmico
precisar	-	infinitivo	modal	deôntico
saber	-	infinitivo	modal	epistêmico
ser	-	particípio	diátese	
parar	de	infinitivo	aspectual	terminativo
acabar	-	gerúndio	aspectual	terminativo
acabar	de	infinitivo	aspectual	terminativo
acabar	por	infinitivo	aspectual	terminativo

Por que a UD desestimula vários AUX?

- Essa é uma questão que gera muita divergência entre as línguas?
- No nível semântico, quanto mais AUX, melhor, pois AUX não tem estrutura argumental e podemos focar na estrutura argumental dos verbos plenos (economia).
- No nível sintático, se um verbo é AUX, ele é dependente, se é pleno, é head. Portanto, muda totalmente a anotação.

OBRIGADA!