Desambiguação de lema e atributos morfológicos na anotação do córpus Porttinari-base

Lucelene Lopes[1], **Magali S. Duran**[1], **Thiago Alexandre Salgueiro Pardo**[1]

[1]Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, São Carlos/SP, Brazil

Abstract. *This paper reports the process of disambiguating lemmas and morphological features in a corpus of Portuguese annotated with Universal Dependencies tagset. We explain the strategies adopted to simplify and reduce the workload of annotators. These strategies contribute to improve the accuracy of linguistic annotation, which is fundamental for various Natural Language Processing tasks.*

Resumo. *Este artigo relata o processo de desambiguação de lemas e atributos morfológicos em um córpus do português anotado com os conjuntos de etiquetas da Universal Dependencies. Explicamos as estratégias adotadas para simplificar e diminuir o trabalho dos anotadores. Essas estratégias contribuem para aumentar a precisão da anotação linguística, a qual é fundamental para diversas tarefas de Processamento de Linguagem Natural.*

1. Introdução

Quanto mais interpretações humanas, lógicas e objetivas, estiverem registradas em um córpus sob forma de anotação, mais ricas as possibilidades de se utilizar esse córpus em tarefas de Processamento de Linguagem Natural (PLN). Uma anotação consistente é particularmente importante para construir um córpus de treino para modelos automáticos [Goldberg 2015, Gamba and Zeman 2023]. Cientes disso vários cientistas uniram seus esforços e vêm aperfeiçoando um modelo multilíngue de anotação chamado *Universal Dependencies* (UD) [Nivre et al. 2016, de Marneffe et al. 2021].

Entre as anotações previstas pela UD, estão a informação do lema, dos atributos morfológicos e da etiqueta morfossintática (*Universal Part-of-Speech* – UPOS) de cada token. Essas informações, muitas vezes, podem ser atribuídas automaticamente, utilizando-se um léxico computacional. No entanto, para tokens que apresentam mais de uma possibilidade de UPOS, lema e/ou atributos morfológicos, é preciso que os anotadores humanos decidam pelas informações mais apropriadas. Em PLN, essas múltiplas possibilidades de anotação são tratadas como casos de ambiguidade, embora, para um humano, o contexto quase sempre seja suficiente para determinar essas informações.

A desambiguação de UPOS em língua portuguesa, usando etiquetas da UD, já foi objeto de outros trabalhos [Lopes et al. 2023, Duran et al. 2022, Duran et al. 2021], porém, até onde é de nosso conhecimento, este é o primeiro trabalho a discutir especificamente a tarefa de desambiguação de lemas e atributos morfológicos em um córpus do português. Compartilhamos os métodos empregados para pré-categorizar automaticamente os tipos de ambiguidades, destacamos casos computacionalmente ambíguos do léxico do português, e apresentamos as quantidades de tokens e sentenças revisados e desambiguados no córpus Porttinari-base [Duran et al. 2023].

Organizamos o artigo como segue: na Seção 2, apresentamos informações sobre a abordagem UD, com vistas a fornecer conceitos básicos utilizados ao longo das discussões; na Seção 3, delineamos a metodologia adotada; na Seção 4, descrevemos em detalhe o processo de desambiguação realizado; na Seção 5, sumarizamos as contribuições deste trabalho e descrevemos as lições aprendidas.

2. Contextualização

A abordagem UD tem sido utilizada para anotar diversos córpus de várias línguas, incluindo os córpus de português, como o Bosque [Rademaker et al. 2017], o Petrogold [Souza et al. 2021], o Porttinari [Duran et al. 2023] e o Cintil [Branco et al. 2022]. O formato de arquivo usado para a anotação UD é o formato CoNLL-U, que contém 10 colunas [Universal Dependencies 2023]. Cada coluna do CoNLL-U tem um nome e as colunas que citaremos neste artigo são as colunas UPOS, LEMMA (lema do token) e FEAT ("*features*", ou atributos morfológicos). Cada token tem uma UPOS e, dada a UPOS, um conjunto específico de atributos morfológicos possíveis, sumarizados na Tabela 1.

Tabela 1. Atributos morfológicos por UPOS

UPOS	Abbr	Case	Definite	Gender	Mood	Number	NumType	Person	Poss	PronType	Tense	VerbForm
ADJ	Yes	–	–	Fem Masc	–	Sing Plur	Ord	–	–	–	–	Part
ADP	Yes	–	–	–	–	–	–	–	–	–	–	–
ADV	Yes	–	–	–	–	–	–	–	–	–	–	–
AUX	Yes	–	–	Fem Masc	Ind Sub Imp Cnd	Sing Plur	–	1 2 3	–	–	Pres Past Fut Imp Pqp	Fin Ger Part Imp
CCONJ	Yes	–	–	–	–	–	–	–	–	–	–	–
DET	Yes	–	Def Ind	Fem Masc	–	Sing Plur	–	1 2 3	Yes	Art Ind Rel Dem Int Prs	–	–
INTJ	Yes	–	–	–	–	–	–	–	–	–	–	–
NOUN	Yes	–	–	Fem Masc	–	Sing Plur	–	–	–	–	–	–
NUM	Yes	–	–	Fem Masc	–	–	Card Frac	–	–	–	–	–
PRON	Yes	Nom Acc Dat	–	Fem Masc	–	Sing Plur	–	1 2 3	Yes	Ind Rel Dem Int Prs	–	–
PROPN	Yes	–	–	–	–	–	–	–	–	–	–	–
PUNCT	–	–	–	–	–	–	–	–	–	–	–	–
SCONJ	Yes	–	–	–	–	–	–	–	–	–	–	–
SYM	–	–	–	–	–	–	–	–	–	–	–	–
VERB	Yes	–	–	Fem Masc	Ind Sub Imp Cnd	Sing Plur	–	1 2 3	–	–	Pres Past Fut Imp Pqp	Fin Ger Part Imp
X	Yes	–	–	–	–	–	–	–	–	–	–	–

Além dos atributos descritos na Tabela 1, é possível também haver atributos não previstos em léxicos do português porque dizem respeito exclusivamente ao eixo sintagmático, como a indicação de voz passiva, **Voice=Pass**, ou porque são palavras estrangeiras, **Foreign=Yes**, ou porque integram nomes próprios, **Proper=Yes**.

3. Metodologia

Ao iniciarmos os trabalhos de anotação, encontramos tokens sem ambiguidade, tokens com ambiguidade de UPOS, de lema ou de atributos morfológicos e tokens que combinam mais de um tipo de ambiguidade. Para automatizar parte da tarefa, fizemos uso do recurso léxico PortiLexicon-UD [Lopes et al. 2022], constituído de formas da língua

portuguesa e suas respectivas possibilidades de anotação com etiquetas da UD. Os tokens que apresentavam mais de uma possível UPOS no PortiLexicon-UD foram desambiguados antes de se passar à verificação da ambiguidade de lema e de atributos morfológicos.

Por não ser nosso foco neste artigo, não detalharemos as dificuldades inerentes à resolução da ambiguidade de UPOS. Essa tarefa foi árdua, pois 44.066 tokens (26% do córpus) apresentavam mais de uma possível UPOS. Grande parte das desambiguações de UPOS já resolveu automaticamente a anotação de lema e de atributos morfológicos, como no exemplo do token "*vestidos*" abaixo, que apresenta uma única alternativa de lema e atributos morfológicos para cada UPOS. Abaixo estão listadas as três opções de UPOS, lema e atributos morfológicos para a palavra "*vestidos*":

- ADJ, "*vestido*", **Gender=Masc**|**Number=Plur**|**VerbForm=Part**;
- NOUN, "*vestido*", **Gender=Masc**|**Number=Plur**;
- VERB, "*vestir*", **Gender=Masc**|**Number=Plur**|**VerbForm=Part**.

Porém, há casos em que a desambiguação de UPOS nem sempre elimina as ambiguidades de lema e de atributos morfológicos, como no exemplo do token "*fora*", a seguir, que apresenta ambiguidades dentro das UPOS AUX e VERB.

- ADP, "*fora*", _;
- ADV, "*fora*", _;
- AUX, "*ser*", **Mood=Ind**|**Number=Sing**|**Person=1**|**Tense=Pqp**|**VerbForm=Fin**;
- AUX, "*ser*", **Mood=Ind**|**Number=Sing**|**Person=3**|**Tense=Pqp**|**VerbForm=Fin**;
- AUX, "*ir*", **Mood=Ind**|**Number=Sing**|**Person=1**|**Tense=Pqp**|**VerbForm=Fin**;
- AUX, "*ir*", **Mood=Ind**|**Number=Sing**|**Person=3**|**Tense=Pqp**|**VerbForm=Fin**;
- NOUN, "*fora*", **Gender=Masc**|**Number=Sing**;
- VERB, "*ser*", **Mood=Ind**|**Number=Sing**|**Person=1**|**Tense=Pqp**|**VerbForm=Fin**;
- VERB, "*ser*", **Mood=Ind**|**Number=Sing**|**Person=3**|**Tense=Pqp**|**VerbForm=Fin**;
- VERB, "*ir*", **Mood=Ind**|**Number=Sing**|**Person=1**|**Tense=Pqp**|**VerbForm=Fin**;
- VERB, "*ir*", **Mood=Ind**|**Number=Sing**|**Person=3**|**Tense=Pqp**|**VerbForm=Fin**.

Sendo assim, para cada token e sua respectiva UPOS anotada no córpus, fizemos a detecção automática de possíveis anotações de lema e atributos morfológicos segundo o PortiLexicon-UD. Anotamos automaticamente os tokens que, para uma dada UPOS, não apresentavam nem ambiguidade de lema nem de atributos morfológicos. Já para aqueles que apresentavam ambiguidade, utilizamos algumas heurísticas (descritas mais abaixo) para anotação automática e submetemos o restante à anotação manual.

É importante ressaltar que, na grande maioria dos casos, uma vez decidida a UPOS de um token, as possibilidades de atributos morfológicos já são conhecidas graças ao PortiLexicon-UD, ou seja, mesmo que haja mais de uma possibilidade de anotação, a tarefa do anotador será escolher entre as possíveis alternativas previstas no léxico.

4. Processo de Desambiguação

Percebemos, durante nossa prática de desambiguação de UPOS, que existe um efeito cascata de um tipo de desambiguação para outro, como já exemplificado acima. Como a desambiguação do lema pode resolver eventualmente as ambiguidades de atributos morfológicos, dividimos o processo em duas etapas:

- Desambiguações de lema dentro de uma mesma UPOS;
- Desambiguações de atributos morfológicos dentro de uma mesma UPOS e de um mesmo lema.

4.1. Desambiguações de Lema - mesma UPOS

Ao procurarmos tokens que admitiam mais de um lema para uma mesma UPOS, encontramos um total de 1.708 tokens, sendo 1.560 verbos plenos, auxiliares ou de cópula (VERB e AUX) que possuem **formas verbais homônimas** para verbos diferentes. As demais 148 ocorrências nessa etapa são **formas nominais homônimas** (substantivos que possuem lemas distintos).

As **formas verbais homônimas** que correspondem a conjugações de verbos distintos ocorreram em 1.560 tokens distribuídos em 1.431 sentenças. Um exemplo é a forma verbal "*viram*" encontrada nas sentenças:

- "*A recepção do Neymar vocês **viram** como foi.*", que trata do verbo "*ver*";
- "*Os moradores se **viram** com grades reforçadas, câmeras e até portão novo para fechar a via.*", que trata do verbo "*virar*".

A desambiguação dessas 1.560 ocorrências foi feita através de planilhas com as 1.431 sentenças que foram manualmente analisadas. Embora várias formas com lema verbal ambíguo tenham sido encontradas, mais da metade (841 das 1.560) eram formas comuns aos verbos "*ser*" e "*ir*". A Tabela 2 mostra alguns exemplos de formas verbais cujos lemas foram desambiguados.

Tabela 2. Exemplos de formas verbais com lema ambíguo

Forma	Opções de Lema		Forma	Opções de Lema	
"*for*"	"*ser*"	"*ir*"	"*fosse*"	"*ser*"	"*ir*"
"*dita*"	"*ditar*"	"*dizer*"	"*pode*"	"*podar*"	"*poder*"
"*postas*"	"*postar*"	"*pôr*"	"*traga*"	"*tragar*"	"*trazer*"
"*sentem*"	"*sentar*"	"*sentir*"	"*vira*"	"*virar*"	"*ver*"

Os **substantivos com formas homônimas**, ou seja, substantivos com mais de uma possibilidade de lema, somaram 148 tokens distribuídos em 146 sentenças. Um exemplo é o token "*críticas*", utilizado como NOUN, que pode ser o feminino plural do substantivo masculino "*crítico*" (como em "*elas são críticas de arte*") ou pode ser plural do substantivo feminino "*crítica*" (como em "*recebeu boas críticas*"). Outro exemplo é o substantivo "*suspeita*", que pode ser feminino singular do substantivo "*suspeito*" ou singular do substantivo feminino "*suspeita*", como exemplificado nas sentenças a seguir:

- "*Uma porta-voz da promotoria disse que a **suspeita** não fez ameaças ou declarações extremistas.*", em que o lema é "*suspeito*" ;
- "*UNE processa Lollapalooza por **suspeita** de burlar lei da meia entrada.*", em que o lema é "*suspeita*".

A desambiguação foi feita através de uma planilha com as 146 sentenças que foram analisadas manualmente. A Tabela 3 mostra alguns exemplos de substantivos cujos lemas foram desambiguados.

4.2. Desambiguações de Atributos Morfólogicos - mesma UPOS e mesmo Lema

Na segunda etapa do processo, analisamos tokens que apresentam mais de uma possibilidade de atributos morfológicos, mesmo apresentando uma mesma UPOS e um mesmo lema. Foram encontrados 11.397 tokens, sendo:

- 7.543 **verbos** com UPOS VERB ou AUX;
- 3.822 **pronomes** com UPOS PRON;
- 32 **substantivos** com UPOS NOUN.

Tabela 3. Exemplos de formas nominais desambiguadas

Forma	Opções de Lema		Forma	Opções de Lema	
"*crítica*"	"*crítico*"	"*crítica*"	"*técnica*"	"*técnico*"	"*técnica*"
"*química*"	"*químico*"	"*química*"	"*porteira*"	"*porteiro*"	"*porteira*"
"*mineradora*"	"*minerador*"	"*mineradora*"	"*música*"	"*músico*"	"*música*"

4.2.1. Ambiguidades de Atributos de Verbos

Nos 7.543 tokens anotados como verbos (VERB ou AUX), a maioria (4.531 tokens) possui ambiguidade tanto nos atributos morfológicos de Pessoa (**Person**) quanto de Tempo, Modo e Forma Verbal (**Tense**, **Mood** e **VerbForm**). As ambiguidades de atributos verbais ocorrem quando diferentes conjugações de um mesmo verbo têm a mesma forma escrita (são homônimas). Por exemplo, a conjugação da segunda pessoa do singular do Imperativo é idêntica à terceira pessoa do singular do presente do Indicativo, como no verbo "*apresentar*", cuja forma "*apresenta*" é utilizada na segunda pessoa do singular do Imperativo ("*apresenta tu*") e na terceira pessoa do singular no presente do Indicativo ("*ele apresenta*").

Dentro de nossa abordagem, tratamos distintamente as ocorrências onde a ambiguidade de formas verbais se dá devido a múltiplas opções de:

- Pessoa (**Person**): primeira, segunda, ou terceira;
- Número (**Number**): singular ou plural;
- Tempo, Modo e Forma Verbal, que são definidos respectivamente por três atributos em UD: **Tense**, **Mood** e **VerbForm**, os quais apresentam 14 possíveis combinações em português. A forma verbal pode ser Infinitivo (pessoal e impessoal), Gerúndio, Particípio ou Finito. Dentro das formas marcadas como Finito, temos os modos e seus respectivos tempos: Indicativo (presente, pretérito perfeito, pretérito imperfeito, pretérito mais-que-perfeito, futuro do presente e futuro do pretérito), Subjuntivo (presente, pretérito imperfeito e futuro) e Imperativo.

Desambiguação do Atributo Pessoa - Dos 7.543 **verbos** (VERB e AUX) com ambiguidade de atributos morfológicos, 6.379 têm ambiguidade de Pessoa (**Person**), e estes estão distribuídos em 4.580 sentenças. Um exemplo é a forma verbal "*para*" nas sentenças:

- "*A escola não **para**, as crianças estão lá todos os dias, afirmou.*", em que "*para*" é a terceira pessoa do singular do Presente do Indicativo;
- "***Para** de chorar porque o seu marido vai cansar, relata.*", em que "*para*" é segunda pessoa do singular do Imperativo.

Dado que o córpus é de gênero jornalístico, observamos que é mais provável que as formas ambíguas pertençam à terceira pessoa do Presente do Indicativo e não à segunda pessoa do Imperativo. Por essa razão, quando os tokens ambíguos apresentavam só essas duas opções (ou seja, em 3.782 ou 59% das ocorrências com ambiguidade de pessoa), utilizamos uma heurística para diminuir o número de casos submetidos à análise dos anotadores. A heurística consistiu em selecionar somente as sentenças onde o token ambíguo era o primeiro token da sentença, ou era precedido de aspas (“), dois pontos (:) ou reticências (...), configurações em que o Imperativo teria maior probabilidade de ocorrer. Essa heurística resultou em 226 sentenças, contendo 318 tokens ambíguos que, analisados manualmente, revelaram 9 casos de segunda pessoa do singular do Imperativo.

Para as demais 2.789 sentenças (contendo 3.464 tokens ambíguos), atribuímos automaticamente a terceira pessoa do singular do Presente do Indicativo.

Para o restante dos verbos onde a desambiguação de Pessoa ainda era necessária, os verbos foram analisados individualmente nas 1.565 sentenças (do total de 4.580 sentenças com ambiguidade de pessoa). Um exemplo é o token "*fazia*" nas sentenças:

- "*Ela não **fazia** a menor ideia de como ou por onde começar a procurar trabalho", recorda.*", onde "*fazia*" é a terceira pessoa do singular do Pretérito do Imperfeito do Indicativo;
- "*É uma percepção que tenho desde que **fazia** residência.*", em que "*fazia*" é a primeira pessoa do singular do Pretérito Imperfeito do Indicativo.

Essa análise resultou na desambiguação de 2.597 tokens, dos quais 2.159 foram anotados como terceira pessoa do singular e 438 como primeira pessoa do singular. A Tabela 4 mostra alguns exemplos de tokens com ambiguidade de pessoa.

Tabela 4. Exemplos de formas verbais com ambiguidade de pessoa

Forma	Opções de Pessoa	
"*demande*"	primeira pessoa do sing. no Pres. do Sub.	terceira pessoa do sing. no Pres. do Sub.
"*conta*"	segunda pessoa do sing. do Imperativo	terceira pessoa do sing. do Pres. do Ind.
"*cresce*"	segunda pessoa do sing. do Imperativo	terceira pessoa do sing. do Pres. do Ind.
"*diz*"	segunda pessoa do sing. do Imperativo	terceira pessoa do sing. do Pres. do Ind.
"*absolva*"	primeira pessoa do sing. no Pres. do Sub.	terceira pessoa do sing. no Pres. do Sub.

Desambiguação do Atributos Tempo e Modo - Considerando que, dos 7.543 tokens verbais com ambiguidade nos atributos morfológicos, 6.379 já foram desambiguados quando se definiu a pessoa do verbo (**Person**), restaram somente 1.013 para se desambiguar o tempo e modo do verbo (**Tense**). Esses 1.013 tokens estavam distribuídos em 714 sentenças. Um exemplo é a forma "*diga*" nas sentenças:

- "*Dá uma boa olhada em os números de a tabela e me **diga** se você não ficou com água em a boca?*", em que "*diga*" é a terceira pessoa do singular do Imperativo;
- "*É simplesmente alguém que coloque as coisas em ordem, e **diga**: atenção, minha gente vamos nos acertar aqui e deixar as coisas de forma que o país consiga andar e não como estamos.*", em que "*diga*" é a terceira pessoa do singular do Subjuntivo.

A desambiguação foi feita através de uma planilha com as 714 sentenças para revisão manual, resultando na anotação de 301 tokens como Presente do Subjuntivo, 207 tokens como Presente do Indicativo, 188 tokens como Pretérito Perfeito do Indicativo e 19 tokens como Futuro do Subjuntivo. A Tabela 5 mostra exemplos dessas desambiguações.

Tabela 5. Exemplos de formas verbais com ambiguidade de tempo

Forma	Opções de Tempo Verbal	
"*aproveite*"	terc. pess. do sing. no Imperativo	terc. pess. do sing. no Pres. do Sub.
"*possam*"	terc. pess. do plural do Imperativo	terc. pess. do plural do Pres. do Sub.
"*tenha*"	terc. pess. do sing. no Imperativo	terc. pess. do sing. no Pres. do Sub.
"*precisamos*"	prim. pess. do plural do Pres. do Sub.	prim. pess. do plural do Pret. do Sub.
"*vão*"	terc. pess. do plural do Imperativo	terc. pess. do plural do Pres. do Sub.
"*mandaram*"	terc. pess. do plural do Pret. Perfeito	terc. pess. do plural do Pret. Mais-que-perf.
"*conseguirem*"	terc. pess. do plural do Fut. do Sub.	terc. pess. do plural do Infinitivo pessoal

4.2.2. Ambiguidades de Atributos de Pronomes

Um total de 3.822 tokens anotados como **pronomes** (PRON) possuem basicamente dois tipos de ambiguidade. A primeira diz respeito ao tipo do pronome (**PronType**), que apresenta ambiguidades entre os valores Relativo, Demonstrativo, Indefinido e Interrogativo, e a segunda diz respeito ao caso gramatical (**Case**): Nominativo, Dativo ou Acusativo.

Desambiguação do Atributo Tipo do Pronome - Essa ambiguidade ocorre em 2.903 tokens em 2.159 sentenças, sendo que em 2.102 deles só existe ambiguidade quanto ao tipo de pronome, enquanto os outros 801 tokens têm também ambiguidade de caso. Um exemplo de ambiguidade de tipo de pronome é o token "*quem*" nas sentenças:

- "*Até o momento,* ***quem*** *ganhou mais com a nova tecnologia foram os clientes.*", em que "*quem*" é Relativo;
- "*A noite está brilhante e acetinada,* ***quem*** *você acha que é?*", em que "*quem*" é Interrogativo.

Para desambiguar as 2.159 sentenças contendo os 2.303 tokens de pronome com ambiguidade de tipo, foram geradas planilhas para revisão manual. A ambiguidade de tipo de pronome ocorreu apenas sobre 11 palavras distintas que são apresentadas na Tabela 6.

Tabela 6. Exemplos de pronomes com ambiguidade de tipo

formas	Opções de Tipo de Pronomes		
"*que*"	Relativo	Interrogativo	–
"*qual*"	Relativo	Interrogativo	–
"*quais*"	Relativo	Interrogativo	–
"*quem*"	Relativo	Interrogativo	Indefinido
"*quantos*"	Relativo	Interrogativo	Indefinido
"*tal*"	Demonstrativo	Indicativo	–
"*tais*"	Demonstrativo	Indicativo	–
"*a*"	Demonstrativo	Pessoal (Caso Acusativo)	–
"*as*"	Demonstrativo	Pessoal (Caso Acusativo)	–
"*o*"	Demonstrativo	Pessoal (Caso Acusativo)	–
"*os*"	Demonstrativo	Pessoal (Caso Acusativo)	–

Desambiguação do Atributo Caso - Considerando que, dos 3.822 pronomes com ambiguidade iniciais, 2.903 já foram desambiguados ao decidir o tipo de pronome, restaram 919 tokens com ambiguidade de caso. Um exemplo de ambiguidade de caso é o pronome "*se*" nas sentenças:

- "*Gilmar não participou e nem o ministro Marco Aurélio, que* ***se*** *declarou suspeito.*", onde "*se*" é Acusativo;
- "*As garotinhas podem* ***se*** *perguntar agora: eu quero ser cabeleireira ou chanceler?*", onde "*se*" é Dativo.

A desambiguação dos 919 tokens resultou em 703 pronomes do caso Nominativo, 219 do caso Acusativo e 98 do caso Dativo. É interessante citar que essa ambiguidade ocorreu apenas para 4 pronomes distintos, todos pronomes pessoais, descritos na Tabela 7.

4.2.3. Ambiguidades de Atributos de Substantivos

O último caso de desambiguação de atributos morfológicos refere-se aos **substantivos** (NOUN), que podem ser utilizados nos gêneros Masculino, Feminino ou não apresentar

Tabela 7. Exemplos de casos de pronomes ambíguos

Palavras	Opções de Caso de Pronomes			Palavras	Opções de Caso de Pronomes	
"*me*"	Acusativo	Dativo	–	"*te*"	Acusativo	Dativo
"*se*"	Acusativo	Dativo	–	"*nos*"	Acusativo	Dativo

nenhum valor, quando são comuns de dois gêneros. Este tipo de ocorrência apareceu em apenas 32 tokens distribuídos em 32 sentenças. Por exemplo, o token "*corte*" pode ser um substantivo masculino (como em "*corte de tecido*") ou um substantivo feminino (como em "*a decisão da corte*"), mas ambos os casos mantêm a UPOS NOUN e o lema "*corte*". Outro exemplo de ambiguidade é o substantivo "*reservas*" nas sentenças:

- "*As chamas, que começaram nas **reservas** naturais, avançaram nos últimos dias com ventos de 80 km/h e chegaram às cidades.*", em que "*reservas*" é feminino;
- "*Tite precisa definir, sem mudar o esquema tático, os **reservas** imediatos de Renato Augusto e Paulinho.*", em que "*reservas*" é masculino.

As 32 sentenças foram analisadas por anotadores, resultando em 22 tokens anotados como masculino e 10 tokens anotados como feminino. Apenas 7 substantivos distintos com ambiguidade de gênero foram encontrados no córpus: "*meia*", "*reserva*", "*reservas*", "*bandeirinha*", "*corte*", "*cortes*" e "*paquera*".

5. Conclusão

Este estudo demonstrou a complexidade da desambiguação de lemas e atributos morfológicos em português. A metodologia empregada tem três grandes vantagens: 1) anotar automaticamente o lema e os atributos dos tokens que não apresentam ambiguidade; 2) anotar automaticamente os atributos que deixam de ser ambíguos quando uma das etapas de desambiguação é concluída; 3) sistematizar a anotação manual dos casos ambíguos, por meio de planilhas, restringindo as alternativas de anotação e diminuindo a probabilidade de erros de anotação.

A grande lição que aprendemos nesse trabalho que faz parte do Projeto POeTiSA (`https://sites.google.com/icmc.usp.br/poetisa/`) é que há um efeito de desambiguação em cascata, que justifica uma ordem de desambiguação das colunas do CoNLL-U: primeiro UPOS, depois lema e, por fim, atributos morfológicos.

O estudo evidencia o volume de tokens ambíguos que requerem uma anotação manual cuidadosa, por anotadores com bons conhecimentos linguísticos, a fim de garantir a precisão dos dados. Esperamos que este relato contribua para uma melhor compreensão dos desafios envolvidos na construção de recursos linguísticos para PLN.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -`http://c4ai.inova.usp.br/`), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

Referências

Branco, A., Silva, J. R., Gomes, L., and António Rodrigues, J. (2022). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5617–5626, Marseille, France. European Language Resources Association.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (2023). The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

Duran, M., Lopes, L., and Pardo, T. (2021). Descrição de numerais segundo modelo universal dependencies e sua anotação no português. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 344–352, Porto Alegre, RS, Brasil. SBC.

Duran, M. S., Oliveira, H., and Scandarolli, C. (2022). Que simples que nada: a anotação da palavra que em corpus de UD. In Pardo, T. A. S., Di-Felippo, A., and Roman, N. T., editors, *Proceedings of the Universal Dependencies Brazilian Festival*, pages 1–11, Fortaleza, Brazil. Association for Computational Linguistics.

Gamba, F. and Zeman, D. (2023). Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In Grobol, L. and Tyers, F., editors, *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Goldberg, Y. (2015). A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

Lopes, L., Duran, M., Fernandes, P., and Pardo, T. (2022). Portilexicon-ud: a portuguese lexical resource according to universal dependencies model. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6635–6643, Marseille, France. European Language Resources Association.

Lopes, L., Fernandes, P., Inacio, M. L., Duran, M. S., and Pardo, T. A. S. (2023). Disambiguation of universal dependencies part-of-speech tags of closed class words in portuguese. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 241–255, Cham. Springer Nature Switzerland.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. ELRA.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In Montemagni, S. and Nivre, J., editors, *Pro-*

ceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Souza, E., Silveira, A., Cavalcanti, T., Castro, M., and Freitas, C. (2021). Petrogold – corpus padrão ouro para o domínio do petróleo. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 29–38, Porto Alegre, RS, Brasil. SBC.

Universal Dependencies (2023). CoNLL-U format - UD version 2. `https://universaldependencies.org/format.html`. Accessed: 2021-06-14.