UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

ATRIBUIÇÃO DE LEMAS E ATRIBUTOS MORFOLÓGICOS SEGUINDO AS DECISÕES ADOTADAS NA ANOTAÇÃO DO CÓRPUS PORTINARI-BASE DENTRO DAS DIRETRIZES DA *UNIVERSAL DEPENDENCIES* (UD)

LUCELENE LOPES
MAGALI SANCHES DURAN
THIAGO ALEXANDRE SALGUEIRO PARDO

Nº 445

RELATÓRIOS TÉCNICOS

São Carlos – SP
Ago./2023

Natural Language Processing initiative (NLP2) of the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP

POeTiSA

POrtuguese processing – Towards Syntactic Analysis and parsing

Atribuição de Lemas e Atributos Morfológicos seguindo as decisões adotadas na anotação do córpus Portinari-base dentro das diretrizes da *Universal Dependencies* (UD)

Lucelene Lopes, Magali Sanches Duran,

Thiago Alexandre Salgueiro Pardo

Agosto/2023

Relatório Técnico do
Núcleo Interinstitucional de Linguística Computacional (NILC)

Sumário

1. Introdução

Este relatório apresenta um conjunto de definições para atribuição de lemas e atributos morfológicos aos tokens de sentenças anotadas segundo o padrão *Universal Dependencies* - UD (de Marneffe et al., 2021; Nivre et al., 2020) no formato CoNLL-U de acordo com as diretrizes para UD para a língua portuguesa (Duran et al., 2022).

A definição da anotação dos lemas e atributos morfológicos dos tokens foi feita dentro do projeto POeTiSA[1] (Pardo et al., 2021) no contexto da construção de um léxico para português usando o formato UD, o PortiLexicon-UD (Lopes et al., 2022), e durante a anotação de um córpus em português anotado segundo o padrão UD, o Porttinari-base (Duran et al., 2023). Nesse sentido, as decisões tomadas refletem um esforço de padronização típico da construção de léxicos, bem como uma visão prática da aplicação em textos reais através da anotação de um córpus de textos jornalísticos.

O escopo deste documento é relatar como anotar lemas e atributos morfológicos em tokens de um córpus seguindo as mesmas diretrizes adotadas atualmente no projeto POeTiSA. Portanto, está fora do escopo deste documento discutir a ambiguidade natural de palavras. Essa ambiguidade é previamente resolvida ao escolher a etiqueta morfossintática (PoS - *Part of Speech*) e a função sintática das palavras nas sentenças. Igualmente, está fora do contexto deste documento decisões de tokenização que podem ser ambíguas, pois assume-se que os tokens e suas funções sintáticas já foram estabelecidos. Dessa forma, a tarefa de anotação apresentada aqui é a atribuição de lema e atributos morfológicos, ou seja, descobrir como representar essas duas informações segundo o padrão UD individualmente para cada token em português.

De uma forma pragmática, este relatório está organizado em uma seção de contextualização do formato e diretrizes da UD com ênfase nas definições da anotação para o português. Em seguida, descrevem-se nossas diretrizes divididas por etiqueta PoS, com exemplos de anotação.

[1] https://sites.google.com/icmc.usp.br/poetisa/

2. Diretrizes de Anotação da UD

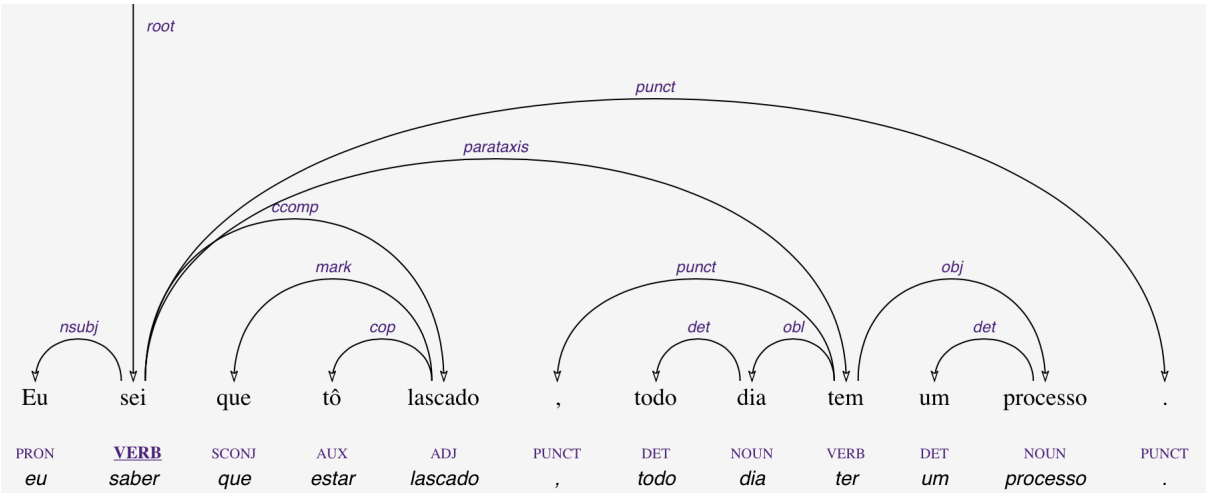
Esta seção descreve algumas diretrizes básicas da UD no que diz respeito às anotações de tokens para seus lemas, etiquetas PoS e seus atributos morfológicos. Nesse sentido, apresentam-se, nas subseções a seguir, detalhes sobre o formato CoNLL-U (Subseção 2.1), quais são as etiquetas PoS (Subseção 2.2) e quais são os atributos morfológicos (Subseção 2.3).

2.1. Formato CoNLL-U

O formato CoNLL-U é usado para anotar informações relativas aos tokens que compõem as sentenças. De acordo com este formato, cada token possui 10 campos com as seguintes informações:

- ID: o identificador do token na sentença (um número sequencialmente atribuído a partir do número 1);
- FORM: o token na sua forma utilizada na sentença;
- LEMMA: o lema ao qual o token corresponde;
- POS: a etiqueta PoS associada ao token;
- XPOS: a etiqueta PoS estendida associada ao token (campo não utilizado no decorrer da construção do córpus Porttinari-base até o momento);
- FEAT: os atributos morfológicos do token;
- HEAD: o ID do token head da relação de dependência do token;
- DEPREL: a relação de dependência do token com seu token head;
- DEPS: a relação de *enhanced dependency* do token (campo não utilizado no decorrer da construção do córpus Porttinari-base até o momento);
- MISC: informações adicionais sobre o token.

Para ilustrar o uso do formato CoNLL-U, a Figura 1 apresenta um exemplo de anotação da sentença "*Eu sei que tô lascado, todo dia tem um processo.*". Mais informações sobre o formato CoNLL-U podem ser encontradas na página oficial da UD em https://universaldependencies.org/format.html.




```
# sent_id = FOLHA_DOC000001_SENT003
# text = Eu sei que tô lascado, todo dia tem um processo.
1   Eu  eu  PRON    _   Case=Nom|Number=Sing|Person=1|PronType=Prs  2   nsubj   _   _
2   sei saber   VERB    _   Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin   0   root    _   _
3   que que SCONJ   _   _   5   mark    _   _
4   tô  estar   AUX _   Abbr=Yes|Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin  5   cop _   _
5   lascado lascado ADJ _   Gender=Masc|Number=Sing|VerbForm=Part   2   ccomp   _   SpacesAfter=No
6   ,   ,   PUNCT   _   _   9   punct   _   _
7   todo    todo    DET _   Gender=Masc|Number=Sing|PronType=Ind    8   det _   _
8   dia dia NOUN    _   Gender=Masc|Number=Sing 9   obl _   _
9   tem ter VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   2   parataxis   _   _
10  um  um  DET _   Definite=Ind|Gender=Masc|Number=Sing|PronType=Art   11  det _   _
11  processo    processo    NOUN    _   Gender=Masc|Number=Sing 9   obj _   SpacesAfter=No
12  .   .   PUNCT   _   _   2   punct   _   _
```

Figura 1. Exemplo de formato CoNLL-U para uma sentença e árvore de dependência correspondente.

2.2. Etiquetas PoS

O esquema de anotação *Universal Dependencies* (UD) possui 17 etiquetas PoS, sendo que apenas 16 delas foram utilizadas na anotação do córpus Porttinari-base. A definição completa das diretivas para anotação do português de etiquetas PoS pode ser encontrada em (Duran, 2021), e um breve resumo das etiquetas é descrito a seguir:

- **ADP**, adposições, uma classe fechada que, em português, corresponde às preposições, como "de", "para" e "com";
- **ADJ**, adjetivos, uma classe aberta que inclui palavras como "bonitas", "último" e "vermelha"; os numerais ordinais escritos por extenso, como "primeiro", "centésima" e "duodécimo"; os numerais ordinais expressos em dígitos como "20º" ou "13ª"; e formas de verbos no particípio como "cansado" e "adequado ";
- **ADV**, advérbios, uma classe aberta com um subconjunto fechado, os advérbios primitivos (aqueles não formados com o sufixo "-mente"). Exemplos do subconjunto fechado são as palavras "cedo", "agora" e "acima". Exemplos do subconjunto aberto são as palavras "normalmente" e "insanamente". Também são incluídas formas abreviadas desses advérbios que aparecem em expressões como "social e economicamente" onde o advérbio "social" é usado como uma forma abreviada do advérbio "socialmente";
- **AUX**, verbos auxiliares e de cópula, uma classe fechada de verbos gramaticalizados, ou seja, que possuem função gramatical. No decorrer da construção do córpus Porttinari-base, essa classe engloba todas as conjugações dos verbos "ser", "estar", "haver", "ir", "ter" e "vir";
- **CCONJ**, conjunções coordenativas, uma classe fechada que contém, por exemplo, "e", "mas" e "portanto";
- **DET**, determinantes, uma classe fechada que inclui artigos como "um" e "o", bem como pronomes que estejam modificando substantivos, por exemplo, "cujo" em "cujo pai", "aquele" em "aquele problema" e "meus" em "meus amigos";
- **INTJ**, as interjeições, uma classe aberta que inclui, por exemplo, "tchau", "oi" e "poxa";
- **NOUN**, substantivos, uma classe aberta que inclui palavras como "presidente**, "**quartos**, "**bandeirinha**, "**salões" e **"**bola**";**
- **NUM**, numerais cardinais, uma classe aberta que inclui os os numerais cardinais escritos por extenso, como "duas", "trinta" e "quinhentos", bem como todos os numerais escritos com dígitos como "51", "-3.1415", "¾", "1,5" e até datas como "25/12/1974";
- **PART**, partículas, uma classe que não é utilizada no decorrer da construção do córpus Porttinari-base;
- **PRON**, pronomes, uma classe fechada atribuída aos pronomes substantivos como "eu", "mim", "aquilo" ou pronomes em função substantiva (função de

sujeito ou objeto de verbos) como "aqueles" em "aqueles que vieram" ou "estes" em "prefiro estes";

- **PROPN**, nomes próprios, uma classe aberta que contempla denominações. No projeto POeTiSA, definiu-se que a classe compreende todos os nomes de pessoas, empresas, órgãos governamentais, projetos, locais, títulos de obras, etc. Esses nomes podem ser simples ou compostos e normalmente são escritos em maiúsculas, como "João", "Paris", "SENAI", "Senado Federal", "Sem Parar" (exceção: iPhone);
- **PUNCT**, todas as pontuações, como ponto final, exclamação, interrogação, vírgula, dois pontos, ponto-e-vírgula, aspas e parênteses;
- **SCONJ**, conjunções subordinativas, uma classe fechada que contém, por exemplo, "conquanto", "se" e "segundo";
- **SYM**, todos os símbolos simples e compostos, por exemplo, "R$", "US$" e "%";
- **VERB**, verbos, uma classe aberta que inclui todas as conjugações dos verbos plenos, por exemplo, "canta", "jogar", "chorastes" e "terias";
- **X**, tudo que não pertence ao vocabulário da língua, como as palavras estrangeiras "petit-four" e "bullying", bem como onomatopéias como "oinc" e "crec".

Mais informações sobre as etiquetas PoS podem ser encontradas na página oficial da UD em https://universaldependencies.org/u/pos/index.html.

2.3. Etiquetas para atributos morfológicos

No momento, a UD define 24 atributos morfológicos principais, sendo 7 atributos lexicais e 17 atributos flexionais, sendo que, destes, 7 são nominais e 10 são verbais. No entanto, apenas 15 são utilizados na anotação para o português dentro do contexto atual do nosso projeto.

2.3.1. Atributos morfológicos lexicais

Os atributos morfológicos apontados como lexicais da UD são:

- **PronType** - Utilizado para indicar o tipo pronominal para palavras das classes PRON e DET, e seus valores possíveis na nossa anotação são:
    - **PronType=Art** - para um artigo (exclusivo para DET e deve ser usado em conjunto com o atributo **Definite**);
    - **PronType=Dem** - para indicar um pronome demonstrativo;
    - **PronType=Ind** - para indicar um pronome indicativo;
    - **PronType=Int** - para indicar um pronome interrogativo;
    - **PronType=Rel** - para indicar um pronome relativo;
    - **PronType=Prs** - para indicar um pronome pessoal (exclusivo para PRON) ou para pronome possessivo (neste caso, deve ser usado em conjunto com o atributo **Poss=Yes**);
- **NumType** - Utilizado para indicar palavras das classes NUM e ADJ, e seus valores possíveis na nossa anotação são:
    - **NumType=Card** - para indicar palavras da classe NUM que são numerais cardinais ou tokens numéricos;
    - **NumType=Ord** - para indicar palavras da classe ADJ que são numerais ordinais;
- **Poss** - Utilizado para indicar palavras das classes PRON e DET, e seu único valor possível na nossa anotação é:
    - **Poss=Yes** - para indicar um pronome possessivo (neste caso, deve ser usado em conjunto com o atributo **PronType=Prs**);
- **Foreign** - Utilizado nas palavras da classe X, e seu único valor possível na nossa anotação é:
    - **Foreign=Yes** - para indicar uma palavra da classe X que não pertence a língua (palavra estrangeira);
- **Abbr** - Utilizado em praticamente todas as classes, exceto PROPN, PUNCT e SYM, e seu único valor possível na nossa anotação é:
    - **Abbr=Yes** - para indicar uma palavra que foi abreviada;
- **Typo** - Utilizado em todas as classes, e seu único valor possível na nossa anotação é:
    - **Typo=Yes** - para indicar uma palavra cuja grafia está equivocada;
- **Reflex** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto.

2.3.2. Atributos morfológicos flexionais nominais

Os atributos morfológicos apontados como flexionais nominais da UD são:

- **Gender** - Utilizado para indicar o gênero de palavras das classes NOUN, ADJ, NUM, PRON, DET, AUX e VERB, e seus valores possíveis na nossa anotação são:
    - **Gender=Fem** - para indicar gênero feminino;
    - **Gender=Masc** - para indicar gênero masculino;
    - Palavras comuns de dois gêneros não possuem o atributo;
- **Number** - Utilizado para indicar o número de palavras das classes NOUN, ADJ, NUM, PRON, DET, AUX e VERB, e seus valores possíveis na nossa anotação são:
    - **Number=Sing** - para indicar número singular;
    - **Number=Masc** - para indicar número plural;
    - Casos de número invariável não possuem o atributo;
- **Definite** - Utilizado em palavras da classe DET para indicar a natureza de artigos, e seus valores possíveis na nossa anotação são:
    - **Definite=Def** - para indicar artigos definidos;
    - **Definite=Ind** - para indicar artigos indefinidos;
- **Case** - Utilizado para indicar palavra da classe PRON, e seus valores possíveis na nossa anotação são:
    - **Case=Acc** - para indicar pronomes pessoais acusativos;
    - **Case=Dat** - para indicar pronomes pessoais dativos;
    - **Case=Nom** - para indicar pronomes pessoais nominativos;
- **Degree** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto;
- **NounClass** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto;
- **Animacy** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto.

2.3.3. Atributos morfológicos flexionais verbais

Os atributos morfológicos apontados como flexionais verbais da UD são:

- **VerbForm** - Utilizado para indicar palavras das classes ADJ, AUX e VERB, e seus valores possíveis na nossa anotação são:
    - **VerbForm=Inf** - para indicar verbos no infinitivo;
    - **VerbForm=Ger** - para indicar verbos no gerúndio;
    - **VerbForm=Part** - para indicar verbos no particípio;

    - **VerbForm=Fin** - para indicar verbos no no indicativo, subjuntivo ou imperativo (utilizado em conjunção com os atributos **Mood** e **Tense**);
- **Mood** - Utilizado para indicar palavras das classes AUX e VERB, e seus valores possíveis na nossa anotação são:
    - **Mood=Ind** - para indicar verbos no indicativo, exceto Futuro do Pretérito (utilizado em conjunção com o atributo **Tense**);
    - **Mood=Cnd** - para indicar verbos no Futuro do Pretérito (não necessita do atributo **Tense**);
    - **Mood=Sub** - para indicar verbos no subjuntivo (utilizado em conjunção com o atributo **Tense**);
    - **Mood=Imp** - para indicar verbos no imperativo (utilizado em conjunção com o atributo **Tense**);
- **Tense** - Utilizado para indicar palavras das classes AUX e VERB, e seus valores possíveis na nossa anotação são:
    - **Tense=Pres** - para indicar o tempo presente do indicativo e do subjuntivo;
    - **Tense=Past** - para indicar o tempo pretérito perfeito do indicativo ou pretérito imperfeito do subjuntivo;
    - **Tense=Imp** - para indicar o tempo pretérito imperfeito do indicativo;
    - **Tense=Fut** - para indicar o tempo futuro do presente do indicativo e futuro do subjuntivo;
    - **Tense=Pqp** - para indicar o tempo pretérito mais-que-perfeito do indicativo;
- **Person** - Utilizado para indicar palavras das classes PRON, DET, AUX e VERB, e seus valores possíveis na nossa anotação são:
    - **Person=1** - para indicar primeira pessoa;
    - **Person=2** - para indicar segunda pessoa;
    - **Person=3** - para indicar terceira pessoa;
- **Voice** - Utilizado para indicar palavras das classes AUX e VERB, e seu único valor possível na nossa anotação é:
    - **Voice=Pass** - para indicar quando um verbo no particípio está sendo usado na voz passiva;
- **Polarity** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto;
- **Aspect** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto;
- **Evident** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto;
- **Polite** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto;
- **Clusivity** - Não utilizado na anotação do português dentro do contexto atual do nosso projeto.

Mais informações sobre as etiquetas de atributos morfológicos e seus valores podem ser encontrados na página oficial da UD em https://universaldependencies.org/u/feat/index.html.

3. Atribuição de Lema e Atributos Morfológicos para o Português

Nesta seção, apresentam-se as diretrizes de atribuição de lema e atributos morfológicos para o português dentro do contexto atual do nosso projeto. Devido à natureza distinta das palavras de acordo com a classe, as diretrizes são apresentadas para cada uma das 17 etiquetas PoS da UD.

Cabe lembrar que, para todas as definições aqui apresentadas, assume-se que os tokens tiveram sua PoS e sua função sintática estabelecidas previamente. Portanto, em todas as definições de lema, atributos e valores, bem como exemplos apresentados nessa seção, assume-se que a análise do contexto (sentença onde o token aparece) já foi feita, logo o contexto está estabelecido sem ambiguidade. Dessa forma, mesmo que a escolha de anotação seja dependente de contexto, a atribuição de lema e atributos morfológicos se torna independente após o contexto ser estabelecido.

3.1. Adjetivos (ADJ)

Classe aberta, com um subconjunto fechado (numerais ordinais).

O lema de um adjetivo que possui flexão de gênero e/ou número é sua forma no masculino singular, sempre em minúsculas. Quando um adjetivo for comum de dois gêneros, seu lema é a própria forma no singular, em minúsculas. Se o adjetivo apresentar número invariável, seu lema é a própria forma em minúsculas.

3.1.1. Atributos Possíveis para ADJ:

- **Gender=Fem/Masc** - atributo opcional a ser omitido quando o adjetivo for comum de dois gêneros
- **Number=Sing/Plur** - atributo opcional a ser omitida quando o adjetivo for utilizado de forma invariável em número
- **Abbr=Yes** - atributo opcional que só deve ser utilizado se o adjetivo estiver abreviado e, neste caso, o lema deve ser a forma não abreviada
- **VerbForm=Part** - atributo opcional que só deve ser utilizado se o adjetivo for originalmente um verbo no particípio passado
- **NumType=Ord** - atributo opcional que só deve ser utilizado se o adjetivo for um numeral ordinal

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de ADJ:

Termo	Lema	PoS	Atributos
baixas	baixo	ADJ	Gender=Fem|Number=Plur
quinta	quinto	ADJ	Gender=Fem|Number=Sing|NumType=Ord
úteis	útil	ADJ	Number=Plur
simples	simples	ADJ	_
aposentada	aposentado	ADJ	Gender=Fem|Number=Sing|VerbForm=Part
Centésimo	centésimo	ADJ	Gender=Masc|Number=Sing|NumType=Ord

3.2. Preposições (ADP)

Classe fechada.

O lema de uma preposição é a repetição da palavra em minúsculas, exceto se abreviada.

3.2.1. Atributos Possíveis para ADP:

- **Abbr=Yes** - atributo opcional que só deve ser utilizado se a preposição estiver abreviada e, neste caso, o lema deve ser a forma não abreviada

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de ADP:

Termo	Lema	PoS	Atributos
de	de	ADP	_
Até	até	ADP	_
para	para	ADP	_
pra	para	ADP	Abbr=Yes
Malgrado	malgrado	ADP	_

3.3. Advérbios (ADV)

Classe aberta, com um subconjunto fechado (advérbios primitivos - advérbios sem a terminação -mente).

O lema de um advérbio é a repetição da palavra em minúsculas, exceto se abreviada. Alguns advérbios não primitivos podem ser anotados de forma abreviada, como na expressão "completa e totalmente", em que o advérbio "completa" é uma abreviação de "completamente".

3.3.1. Atributos Possíveis para ADV:

- **Abbr=Yes** - atributo opcional que só deve ser utilizado se o advérbio estiver abreviado e, neste caso, o lema deve ser a forma não abreviada

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de ADV:

Termo	Lema	PoS	Atributos
abaixo	abaixo	ADV	_
Cedo	cedo	ADV	_
utilmente	utilmente	ADV	_
completa	completamente	ADV	Abbr=Yes
Não	não	ADV	_

3.4. Verbos Auxiliares (AUX)

Classe fechada, formas conjugadas dos verbos "ser", "estar", "ir", "haver", "ter" e "vir".

O lema de um verbo auxiliar é a palavra no infinitivo impessoal, sempre em minúsculas.

3.4.1. Atributos Possíveis para AUX:

- **VerbForm=Inf/Ger/Part/Fin** (obrigatório)
- **Mood=Ind/Sub/Cnd/Imp** - Não utilizado para os tempos verbais Infinitivo, Gerúndio e Particípio, sendo obrigatório para os demais
- **Tense=Pres/Past/Fut/Imp/Pqp** - Não utilizado para os tempos verbais Infinitivo, Gerúndio, Particípio e Futuro do Pretérito, sendo obrigatório para os demais
- **Gender=Fem/Masc** - Obrigatório para verbos no particípio, sendo ausente nos demais tempos verbais
- **Number=Sing/Plur** - Não utilizado para os tempos verbais Infinitivo impessoal e Gerúndio, sendo obrigatório para os demais
- **Person=1/2/3** - Não utilizado para os tempos verbais Infinitivo impessoal, Gerúndio e Particípio, sendo obrigatório para os demais
- **Voice=Pass** - Opcional só para verbos no particípio, sendo ausente nos demais tempos verbais
- **Abbr=Yes** - atributo opcional que só deve ser utilizado se o auxiliar estiver abreviado e, neste caso, o lema deve ser a forma não abreviada

Exemplos de AUX:

Termo	Lema	PoS	Atributos
Ter	ter	AUX	VerbForm=Inf
			Infinitivo Impessoal
haveres	haver	AUX	Number=Sing|Person=2|VerbForm=Inf
			Infinitivo Pessoal
indo	ir	AUX	VerbForm=Ger
			Gerúndio
sido	ser	AUX	Gender=Masc|Number=Sing|VerbForm=Part
			Particípio

venho	vir	AUX	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
			Presente do Indicativo
era	ser	AUX	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin
			Pretérito Imperfeito do Indicativo
tava	estar	AUX	Abbr=Yes|Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin
			Pretérito Imperfeito do Indicativo - abreviado
vieras	vir	AUX	Mood=Ind|Number=Sing|Person=2|Tense=Pqp|VerbForm=Fin
			Pretérito Mais-que-Perfeito do Indicativo
estareis	estar	AUX	Mood=Ind|Number=Plur|Person=2|Tense=Fut|VerbForm=Fin
			Futuro do Presente do Indicativo
serias	ser	AUX	Mood=Cnd|Number=Sing|Person=2|VerbForm=Fin
			Futuro do Pretérito do Indicativo
Haja	haver	AUX	Mood=Sub|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
			Presente do Subjuntivo
tivéssemos	ter	AUX	Mood=Sub|Number=Plur|Person=1|Tense=Past|VerbForm=Fin
			Pretérito Imperfeito do Subjuntivo
forem	ser	AUX	Mood=Sub|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin
			Futuro do Subjuntivo
Vamos	ir	AUX	Mood=Imp|Number=Plur|Person=1|VerbForm=Fin
			Imperativo Afirmativo

3.5. Conjunções Coordenativas (CCONJ)

Classe fechada.

O lema de uma conjunção coordenativa é a repetição da palavra em minúsculas, exceto se abreviada.

3.5.1. Atributos Possíveis para CCONJ:

- **Abbr=Yes** - atributo opcional que só deve ser utilizado se a conjunção estiver abreviada e, neste caso, o lema deve ser a forma não abreviada

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de ADV:

Termo	Lema	PoS	Atributos
Porém	porém	CCONJ	_
como	como	CCONJ	_
logo	logo	CCONJ	_
todavia	todavia	CCONJ	_
pq	porque	CCONJ	Abbr=Yes

3.6. Determinantes (DET)

Classe fechada.

O lema de um determinante é a palavra no masculino singular, sempre em minúsculas.

3.6.1. Atributos Possíveis para DET:

- **PronType=Art/Dem/Ind/Rel/Int/Prs** - atributo obrigatório
- **Definite=Def/Ind** - Obrigatório para artigos, sendo ausente nos demais determinantes
- **Gender=Fem/Masc** - Obrigatório para artigos, sendo opcional nos demais determinantes
- **Number=Sing/Plur** - Obrigatório para artigos, sendo opcional nos demais determinantes
- **Person=1/2/3** - Obrigatório para pronomes possessivos, sendo ausente nos demais
- **Poss=Pass** - Obrigatório para pronomes possessivos, sendo ausente nos demais

Exemplos de DET:

Termo	Lema	PoS	Atributos
uma	um	DET	Definite=Ind|Gender=Fem|Number=Sing|PronType=Art
			Artigo
própria	próprio	DET	Gender=Fem|Number=Sing|PronType=Dem
			Pronome Demonstrativo
que	que	DET	PronType=Ind
			Pronome Indefinido
quantos	quanto	DET	Gender=Masc|Number=Plur|PronType=Rel
			Pronome Relativo
qual	qual	DET	Number=Sing|PronType=Int
			Pronome Interrogativo
sua	seu	DET	Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs
			Pronome Possessivo

3.7. Interjeições (INTJ)

Classe aberta.

O lema de uma interjeição é a repetição da palavra em minúsculas.

3.7.1. Atributos Possíveis para INTJ:

- "_" Nenhum atributo admitido

Exemplos de INTJ:

Termo	Lema	PoS	Atributos
Amém	amém	INTJ	_
Nossa	nossa	INTJ	_
olá	olá	INTJ	_
poxa	poxa	INTJ	_
Ufa	ufa	INTJ	_

3.8. Substantivos (NOUN)

Classe aberta.

O lema de um substantivo que apresenta flexão de gênero é a palavra no masculino singular, sempre em minúsculas. Se o substantivo tiver gênero invariável, o lema é a forma no singular. Se o substantivo for comum de dois gêneros, o lema é a própria forma no singular. Igualmente, quando um substantivo for invariável em número, entende-se que a forma invariável serve ao número singular também.

3.8.1. Atributos Possíveis para NOUN:

- **Gender=Fem/Masc** - atributo opcional
- **Number=Sing/Plur** - atributo opcional
- **Abbr=Yes** - atributo opcional

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de NOUN:

Termo	Lema	PoS	Atributos
casas	casa	NOUN	Gender=Fem|Number=Plur
estudante	estudante	NOUN	Number=Sing
baixa	baixa	NOUN	Gender=Fem|Number=Sing
tríceps	tríceps	NOUN	_
tadinha	coitadinho	NOUN	Abbr=Yes|Gender=Fem|Number=Sing

3.9. Numerais (NUM)

Classe aberta, com um subconjunto fechado (numerais escritos por extenso).

O lema de um numeral é a própria forma, se escrito em dígitos, e a forma no masculino e em minúsculas, se escrito por extenso.

3.9.1. Atributos Possíveis para NUM:

- **NumType=Card** - atributo obrigatório
- **Gender=Fem/Masc** - atributo opcional

Exemplos de NUM:

Termo	Lema	PoS	Atributos
uma	um	NUM	Gender=Fem|NumType=Card
dois	dois	NUM	Gender=Masc|NumType=Card
34	34	NUM	NumType=Card
7/7/2023	7/7/2023	NUM	NumType=Card
Quinhentos	quinhentos	NUM	Gender=Masc|NumType=Card

3.10. Partículas (PART)

Classe não utilizada no decorrer da construção do córpus Porttinari-base.

3.11. Pronomes (PRON)

Classe fechada.

O lema de um pronome que possui flexão de gênero e/ou número é a palavra no masculino singular, sempre em minúsculas; caso contrário, o lema é a própria forma em minúsculas.

3.11.1. Atributos Possíveis para PRON:

- **PronType=Dem/Ind/Rel/Int/Prs** - atributo obrigatório
- **Gender=Fem/Masc** - atributo opcional para todos os pronomes
- **Number=Sing/Plur** atributo opcional para todos os pronomes
- **Person=1/2/3** - atributo obrigatório para pronomes possessivos, sendo opcional para pronomes demonstrativos, ausente nos demais
- **Poss=Pass** - atributo obrigatório para pronomes possessivos, sendo ausente nos demais
- **Case=Acc/Dat/Nom** atributo obrigatório para pronomes pessoais, sendo ausente nos demais pronomes

Exemplos de PRON:

Termo	Lema	PoS	Atributos
a	o	PRON	Gender=Fem|Number=Sing|Person=3|PronType=Dem
			Pronome Demonstrativo
algumas	algum	PRON	Gender=Fem|Number=Plur|PronType=Ind
			Pronome Indefinido
quais	quais	PRON	Number=Plur|PronType=Rel
			Pronome Relativo
quem	quem	PRON	PronType=Int
			Pronome Interrogativo
mim	mim	PRON	Case=Dat|Number=Sing|Person=1|PronType=Prs
			Pronome Pessoal
vossos	vosso	PRON	Gender=Masc|Number=Plur|Person=2|Poss=Yes|PronType=Prs
			Pronome Possessivo

3.12. Nomes Próprios (PROPN)

Classe aberta.

O lema de um nome próprio é a repetição da palavra na exata grafia respeitando maiúsculas e minúsculas.

3.12.1. Atributos Possíveis para PROPN:

- "_" Nenhum atributo admitido

Exemplos de POPN:

Termo	Lema	PoS	Atributos
Obama	Obama	PROPN	_
Tribunal	Tribunal	PROPN	_
iPhone	iPhone	PROPN	_
FGTS	FGTS	PROPN	_
PoS	PoS	PROPN	_

3.13. Pontuações (PUNCT)

Classe aberta.

O lema de uma pontuação é a repetição da token na exata grafia.

3.13.1. Atributos Possíveis para PUNCT:

- "_" Nenhum atributo admitido

Exemplos de PUNCT:

Termo	Lema	PoS	Atributos
!	!	PUNCT	_
.	.	PUNCT	_
...	...	PUNCT	_
(	(	PUNCT	_
"	"	PUNCT	_

3.14. Conjunções Subordinativas (SCONJ)

Classe fechada.

O lema de uma conjunção subordinativa é a repetição da palavra em minúsculas, exceto se abreviada.

3.14.1. Atributos Possíveis para SCONJ:

- **Abbr=Yes** - atributo opcional

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de SCONJ:

Termo	Lema	PoS	Atributos
pois	pois	SCONJ	_
Como	como	SCONJ	_
q	que	SCONJ	Abbr=Yes
conquanto	conquanto	SCONJ	_
Porque	porque	SCONJ	_

3.15. Símbolos (SYM)

Classe aberta.

O lema de um símbolo é a repetição do token na exata grafia, inclusive respeitando maiúsculas e minúsculas.

3.15.1. Atributos Possíveis para SYM:

- "_" Nenhum atributo admitido

Exemplos de SYM:

Termo	Lema	PoS	Atributos
R$	R$	SYM	_
%	%	SYM	_
!@#	!@#	SYM	_
(r)	(r)	SYM	_
#	#	SYM	_

3.16. Verbos (VERB)

Classe aberta.

O lema de um verbo é a palavra no infinitivo impessoal, sempre em minúsculas.

3.16.1. Atributos Possíveis para VERB:

- **VerbForm=Inf/Ger/Part/Fin** (obrigatório)
- **Mood=Ind/Sub/Cnd/Imp** - Não utilizado para os tempos verbais Infinitivo, Gerúndio e Particípio, sendo obrigatório para os demais
- **Tense=Pres/Past/Fut/Imp/Pqp** - Não utilizado para os tempos verbais Infinitivo, Gerúndio, Particípio e Futuro do Pretérito, sendo obrigatório para os demais
- **Gender=Fem/Masc** - Obrigatório para verbos no particípio, sendo ausente nos demais tempos verbais
- **Number=Sing/Plur** - Não utilizado para os tempos verbais Infinitivo impessoal e Gerúndio, sendo obrigatório para os demais
- **Person=1/2/3** - Não utilizado para os tempos verbais Infinitivo impessoal, Gerúndio e Particípio, sendo obrigatório para os demais
- **Voice=Pass** - Opcional só para verbos no particípio, sendo ausente nos demais tempos verbais
- **Abbr=Yes** - atributo opcional que só deve ser utilizado se o verbo estiver abreviado e, neste caso, o lema deve ser a forma não abreviada

Exemplos de VERB:

Termo	Lema	PoS	Atributos
Conjugar	conjugar	VERB	VerbForm=Inf
			Infinitivo Impessoal
quereres	querer	VERB	Number=Sing|Person=2|VerbForm=Inf
			Infinitivo Pessoal
lutando	lutar	VERB	VerbForm=Ger
			Gerúndio
aposentado	aposentar	VERB	Gender=Masc|Number=Sing|VerbForm=Part
			Particípio

Escrevo	escrever	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
			Presente do Indicativo
jogava	jogar	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin
			Pretérito Imperfeito do Indicativo
andava	andar	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin
			Pretérito Imperfeito do Indicativo
subiras	subir	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pqp|VerbForm=Fin
			Pretérito Mais-que-Perfeito do Indicativo
contereis	conter	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Fut|VerbForm=Fin
			Futuro do Presente do Indicativo
fugirias	fugir	VERB	Mood=Cnd|Number=Sing|Person=2|VerbForm=Fin
			Presente do Subjuntivo
componha	compor	VERB	Mood=Sub|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
			Presente do Subjuntivo
estudássemos	estudar	VERB	Mood=Sub|Number=Plur|Person=1|Tense=Past|VerbForm=Fin
			Pretérito Imperfeito do Subjuntivo
forem	ir	VERB	Mood=Sub|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin
			Futuro do Subjuntivo
Corramos	correr	VERB	Mood=Imp|Number=Plur|Person=1|VerbForm=Fin
			Imperativo Afirmativo

3.17. Outros (X)

Classe aberta (frequentemente usada para palavras em língua estrangeira, mas não somente).

O lema de um token marcado como X é a repetição do token na exata grafia respeitando maiúsculas e minúsculas.

3.17.1. Atributos Possíveis para X:

- **Abbr=Yes** - atributo opcional
- **Foreign=Yes** - atributo opcional

Como todo token em UD, quando não houver atributos, a indicação "_" deve ser utilizada.

Exemplos de SYM:

Termo	Lema	PoS	Atributos
habitué	habitué	X	Foreign=Yes
Italy	Italy	X	Foreign=Yes
ão	ão	X	_
paeja	paeja	X	Foreign=Yes
time	time	X	Foreign=Yes

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Articial da USP (C4AI - http://c4ai.inova.usp.br/), com o apoio da IBM e da FAPESP (processo 2019/07665-4). Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei nº 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

de Marneffe, M.-C.; Manning, C.D.; Nivre, J.; Zeman, D. (2021). Universal Dependencies. Computational Linguistics, Vol. 47, N. 2, pages 255-308.

Duran, M.S. (2021). Manual de anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em língua portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Technical Report 434, ICMC-USP.

Duran, M.S.; Nunes, M.G.V.; Lopes, L.; Pardo, T.A.S. (2022). Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. Domínios de Lingu@gem, Vol. 16, N. 4, pages 1608-1643.

Duran, M.S.; Lopes, L.; Nunes, M.G.V.; Pardo, T.A.S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In the Proceedings of the 14th Symposium in Information and Human Language Technology (STIL). Em publicação.

Lopes, L.; Duran, M.S.; Fernandes, P.; Pardo, T.A.S. (2022). PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In the Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 6635-6643, Marseille, France.

Lopes, L.; Duran, M.S.; Pardo, T.A.S. (2023). Verifica-UD: a Verifier for Universal Dependencies Annotation for Portuguese. In the Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival (UDFest-BR). Em publicação.

Nivre, J.; de Marneffe, M.-C.; Ginter, F.; Hajic, J.; Manning, C. D.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In the Proceedings of the 12th Language Resources and Evaluation Conference, pages 4034-4043, Marseille, France.

Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In the Proceedings of the XIII Symposium in Information and Human Language (STIL), pages 1-10.