

Documentação UD em português (e para língua portuguesa)

Elvis de Souza

PUC-Rio, Brasil

Tatiana Cavalcanti

Aline Silveira

Wograine Evelyn

Cláudia Freitas

O projeto Universal Dependencies ([McDonald et al. 2013](#)) apresenta um tagset & uma gramática. Isso significa dizer que, para além de um conjunto de etiquetas que correspondem às classes da Gramática Tradicional (objeto, sujeito etc.), o UD também faz diversas escolhas que diferem da GT. Nesse documento, apresentamos a documentação detalhadas e as escolhas linguísticas relativas ao processo de revisão do material UD em Português. Considerando que UD funciona como uma espécie de segunda língua gramatical, partimos, sempre que possível, das categorias e análises de GT, e não de UD.

Conteúdo

Documentação UD em português

(e para língua portuguesa)

Elvis de Souza, Tatiana Cavalcanti, Aline Silveira, Wograine Evelyn,
Cláudia Freitas

1

2 Formato UD

5

1 Colunas/anotações 5

2 Manipulação em Python 6

3 Classes gramaticais (upos)

7

1 Verbos de ligação 7

1.1 Verbo *ser* como verbo pleno 8

1.2 Verbo *ser* como voz passiva 9

2 Numerais 9

2.1 *Primeiro* lugar: adjetivo ou numeral? 9

4 Atributos morfológicos (feats)

11

5 Dependências (dephead e deprel)

15

1 Estruturas comparativas 15

1.1 Frases do Working Group 15

1.2 Frases do Bosque-UD 15

2 Formato UD

[Ir para tabela de conteúdos](#)

Os treebanks adaptados para a gramática UD são disponibilizados no formato CoNLL, em que há um token por linha. Cada anotação de cada token, por sua vez, é disposta em uma coluna, sendo 10 colunas ao todo. Cada token tem a configuração conforme a [Tabela 1: Colunas do formato UD 2.0](#), com uma tabulação (*Tab*) separando as colunas. Colunas sem nenhum valor devem, necessariamente, ser preenchidas com *underline*.

Tabela 1: Colunas do formato UD 2.0

id	word	lemma	upos	xpos	feats	dephead	deprel	deps	misc
----	------	-------	------	------	-------	---------	--------	------	------

1 Colunas/anotações

1. “id” corresponde ao número do token, em ordem crescente;
2. “word”, à palavra tal como aparece na frase (exceto no caso de contração, como “da”, em que a palavra será desmembrada nos tokens “de” e “a”);
3. “lemma” se refere à palavra tal como aparece no dicionário: em no singular e em masculino ou infinitivo;
4. “upos” (classe gramatical “universal”) se refere à classe gramatical;
5. No corpus Bosque-UD, a coluna “xpos” (classe gramatical específica) é preenchida com a saída do sistema PALAVRAS para a mesma frase;
6. “feats” (atributos morfológicos) é preenchida com as características morfológicas do token;
7. “dephead” (dependência sintática), com o id do token de quem é filho;

8. “deprel” (relação de dependência), com a relação sintática que o conecta ao seu pai;
9. “deps” (dependência específica) não é utilizado no Bosque-UD;
10. “misc” (miscelânea) se refere a quaisquer informações extras que desejemos adicionar ao token.

2 Manipulação em Python

Para manipular arquivos no formato UD em Python, com as classes `Corpus`, `Sentence` e `Token` (e suas respectivas anotações), desenvolvemos e utilizamos o `es-estrutura_ud.py`.

3 Classes gramaticais (upos)

Ir para tabela de conteúdos

As classes gramaticais em UD podem ser consultadas na [Tabela 2: As classes gramaticais do UD em português](#).

Tabela 2: As classes gramaticais do UD em português

upos	Observações
ADJ	adjetivos e numerais ordinais
ADP	preposições
PUNCT	pontuação
ADV	advérbio
AUX	auxiliar - “ser”, “estar” (Seção 1: Verbos de ligação), e locuções verbais
SYM	símbolos
INTJ	interjeição
CCONJ	conjunção coordenativa
NOUN	substantivo
DET	determinante - artigos e pronomes adjetivos
PROPN	nomes próprios, apenas se com inicial maiúscula
NUM	numeral - exceto os ordinais, que são adjetivos
PART	partícula
VERB	verbo
PRON	apenas pronomes substantivos
SCONJ	conjunções subordinativas
X	no Bosque-UD, palavras estrangeiras

1 Verbos de ligação

Apenas os verbos “ser” e “estar” são considerados verbos de ligação, e portanto serão sempre anotados como *AUX*. Os demais verbos que a GT costuma elencar como verbo de ligação (parecer, permanecer, etc.) são anotados como *VERB*. Os

verbos de ligação *AUX* terão relação sintática “cop”, e nunca poderão ser núcleo de uma oração (Xx) nem conter dependentes. **Figura 1: O preço é de US\$ 422.**

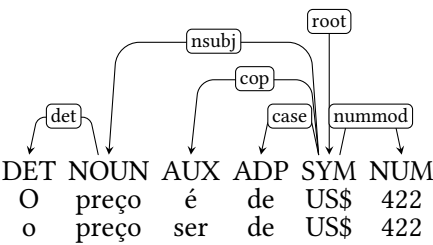


Figura 1: O preço é de US\$ 422

1.1 Verbo *ser* como verbo pleno

Atenção para casos em que o “ser” deve ser *VERB*.

1) Como na **Figura 2: A expectativa era que chegasse a US\$7 milhões**, o “ser” deve manter a relação de núcleo da oração caso o predicado (que seria não-verbal, por se tratar de um verbo de ligação) seja uma oração (*ccomp*, *xcomp*).

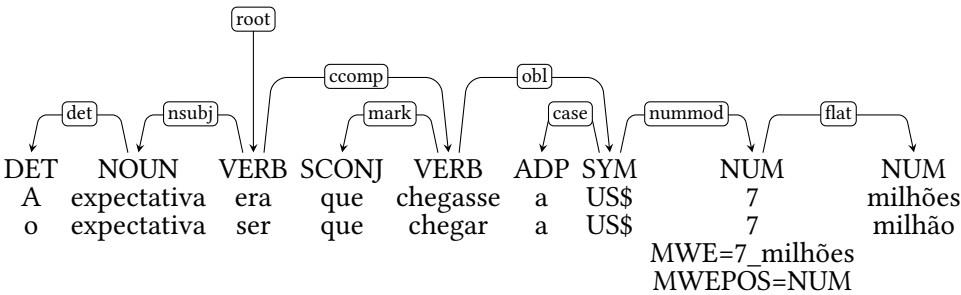


Figura 2: A expectativa *era* que chegasse a US\$7 milhões

2) “ser” verbo intransitivo (verbo pleno) também deve ter a anotação *VERB* (**Figura 3: Isso foi nos Estados Unidos**).

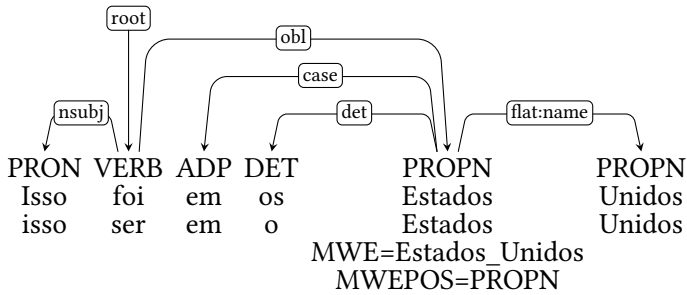


Figura 3: Isso *foi* nos Estados Unidos

1.2 Verbo *ser* como voz passiva

A anotação de “ser” como voz passiva é diferente da anotação do verbo de ligação (Seção 1: Verbos de ligação) e da anotação de “ser” como verbo pleno (Subseção 1.1: Verbo *ser* como verbo pleno).

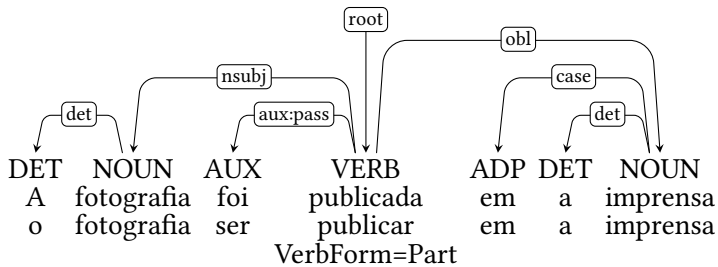


Figura 4: A fotografia *foi* publicada na imprensa

2 Numerais

2.1 Primeiro lugar: adjetivo ou numeral?

Numerais ordinais escritos por extenso devem ser anotados como *ADJ*, e recebem a feature *NumType=Ord*, como na Figura 5: Anotação do sintagma *primeira tentativa*.

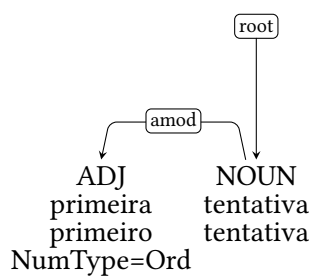


Figura 5: Anotação do sintagma *primeira tentativa*

4 Atributos morfológicos (feats)

Ir para tabela de conteúdos

Temos a seguinte distribuição de atributos morfológicos por classe gramatical (**Tabela 3: Atributos morfológicos (feats)**). É importante notar que os atributos morfológicos devem constar em ordem alfabética e são separados por uma barra reta.

upos	features
ADJ	Gender=[Fem, Masc, Unsp] NumType=[Ord] Number=[Plur, Sing]
ADP	–
ADV	Polarity=[Neg] –
AUX	Gender=[Fem, Masc] Mood=[Cnd, Imp, Ind, Sub] Number=[Plur, Sing] Person=[1, 2, 3] Tense=[Fut, Imp, Past, Pqp, Pres] VerbForm=[Fin, Ger, Inf, Part]
CCONJ	–
DET	Definite=[Def, Ind] Gender=[Fem, Masc, Unsp] Number=[Plur, Sing, Unsp] PronType=[Art, Dem, Emp, Ind, Int, Neg, Prs, Rel, Tot]

INTJ	–
NOUN	Foreign=[Yes] Gender=[Fem, Masc, Unsp] NumType=[Ord] Number=[Plur, Sing, Unsp]
NUM	Gender=[Fem, Masc, Unsp] NumType=[Card, Frac, Mult, Ord, Range, Sets] Number=[Plur, Sing]
PART	Gender=[Masc] Number=[Sing]
PRON	Case=[Acc, Dat, Nom] Definite=[Def, Ind] Gender=[Fem, Masc, Unsp] Number=[Plur, Sing, Unsp] Person=[1, 2, 3] PronType=[Art, Dem, Ind, Int, Neg, Prs, Rel, Tot] Reflex=[Yes] VerbForm=[Ger]
PROPN	Gender=[Fem, Masc, Unsp] Number=[Plur, Sing]
PUNCT	–
SCONJ	Gender=[Fem, Masc] Number=[Plur, Sing] PronType=[Ind, Rel]
SYM	–

VERB	Gender=[Fem, Masc] Mood=[Cnd, Imp, Ind, Sub] Number=[Plur, Sing] Person=[1, 2, 3] Tense=[Fut, Imp, Past, Pqp, Pres] VerbForm=[Fin, Ger, Inf, Part] Voice=[Pass]
------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

X	—
---	---

5 Dependências (dephead e deprel)

Ir para tabela de conteúdos

1 Estruturas comparativas

Estruturas comparativas são de anotação complexa, o que se verifica pela existência de um **working group (WG) em UD** dedicado especialmente a elas. A seguir, listamos as frases utilizadas no WG, traduzidas em português, e com a anotação adequada, além de algumas frases de anotação complexa no Bosque-UD.

1.1 Frases do Working Group

1	Eu	eu	PRON	—	Case=Nom Gender=Fem Number=Sing Person=1 Prontype=Prs	2	nsubj	—	—
2	coloquei	colocar	VERB	—	Mood=Ind Number=Sing Person=1 Tense=Past VerbForm=Fin	0	root	—	—
3	tanta	tanto	DET	—	Gender=Fem Number=Sing Prontype=Ind	4	det	—	—
4	farinha	farinha	NOUN	—	Gender=Fem Number=Sing	2	obj	—	—
5	quanto	quanto	SCONJ	—	—	8	mark	—	—
6	a	a	DET	—	Definite=Def Gender=Fem Number=Sing Prontype=Art	7	det	—	—
7	receita	receita	NOUN	—	Gender=Fem Number=Sing	8	nsubj	—	—
8	pedia	pedia	VERB	—	Mood=Ind Number=Sing Person=3 Tense=Imp VerbForm=Fin	2	advcl	—	SpaceAfter=No
9	.	.	PUNCT	—	2	punct	—	SpaceAfter=No	—

Figura 6: Eu coloquei *tanta farinha quanto* a receita pedia.

1	Martin	Martin	PROPN	—	Gender=Masc Number=Sing	4	nsubj	—	—
2	é	ser	AUX	—	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	4	cop	—	—
3	o	o	DET	—	Definite=Def Gender=Masc Number=Sing Prontype=Art	4	det	—	—
4	cara	cara	NOUN	—	Gender=Fem Number=Sing	0	root	—	—
5	mais	mais	ADV	—	6	advmod	—	—	—
6	inteligente	inteligente	ADJ	—	8	case	—	—	—
7	de	de	ADP	—	Gender=Fem Number=Sing	4	amod	—	—
8	todos	todo	PRON	—	Gender=Masc Number=Plur Prontype=Tot	6	obl	—	SpaceAfter=No
9	.	.	PUNCT	—	4	punct	—	SpaceAfter=No	—

Figura 7: Martin é o cara *mais inteligente de todos*.

1.2 Frases do Bosque-UD

Abreviações

[Ir para tabela de conteúdos](#)

Agradecimentos

[Ir para tabela de conteúdos](#)

Referências

McDonald, Ryan, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, T Oscar et al. 2013. Universal dependency annotation for multilingual parsing. Em *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 92–97.