

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 detalhada 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
1	Não apagar: seção de códigos que podem ser úteis de copiar . .	3
2	Formato UD	4
	2.1 Colunas/anotações	4
	2.2 Manipulação em Python	5
3	Classe gramatical (upos)	5
	3.1 Primeiro lugar: ADJ ou NUM?	5
	3.2 Verbos de ligação	7
	3.3 Verbo <i>ser</i> como verbo pleno	7
	3.4 Verbo <i>ser</i> como voz passiva	8
4	Atributos morfológicos (feats)	8
5	Estruturas comparativas	10
	5.1 Frases do Working Group	11
	5.2 Frases do Bosque-UD	11

1 Não apagar: seção de códigos que podem ser úteis de copiar

Phasellus maximus erat ligula, accumsan rutrum augue facilisis in. Proin sit amet pharetra nunc, sed maximus erat. Duis egestas mi eget purus venenatis vulputate vel quis nunc. Nullam volutpat facilisis tortor, vitae semper ligula dapibus sit amet. Suspendisse fringilla, quam sed laoreet maximus, ex ex placerat ipsum, porta ultrices mi risus et lectus. Maecenas vitae mauris condimentum justo fringilla sollicitudin. Fusce nec interdum ante. Curabitur tempus dui et orci convallis molestie (Chomsky1957). Nordhoff2018 is useful for compiling bibliographies

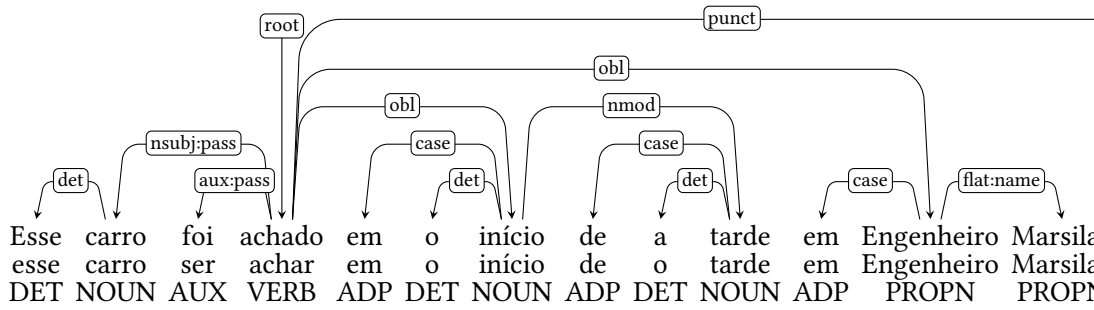


Figura 1: The sentence ‘Esse carro foi achado no início da tarde em Engenheiro Marsilac/This car was found in the beginning of the evening at ‘Engenheiro Marsilac’ (location)’ annotated with the parser PALA-VRAS and UD scheme.

Tabela 1: Frequencies of word classes

	nouns	verbs	adjectives	adverbs
absolute	12	34	23	13
relative	3.1	8.9	5.7	3.2

2 Formato UD

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 2: Colunas do formato UD 2.0**, com uma tabulação (*Tab*) separando as colunas. Colunas sem nenhum valor devem, necessariamente, ser preenchidas com *underline*.

Tabela 2: Colunas do formato UD 2.0

id word lemma upos xpos feats dephead deprel deps misc

2.1 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.

2.2 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 `estrutura_ud.py`.

3 Classe gramatical (upos)

As classes gramaticais em UD podem ser consultadas na [Tabela 3: As classes gramaticais do UD em português](#).

3.1 *Primeiro* lugar: ADJ ou NUM?

Numerais ordinais escritos por extenso devem ser anotados como *ADJ*, e recebem a feature “NumType=Ord”, como na [Figura 2: Anotação do sintagma primeira tentativa](#).

Tabela 3: As classes gramaticais do UD em português

upos	Observações
ADJ	adjetivos e numerais ordinais
ADP	preposições
PUNCT	pontuação
ADV	advérbio
AUX	auxiliar - “ser”, “estar” (Subseção 3.2: 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

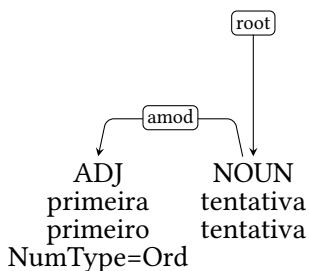


Figura 2: Anotação do sintagma *primeira tentativa*

3.2 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 3: O preço é de US\$ 422.**

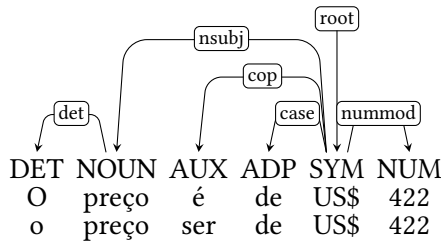


Figura 3: O preço é de US\$ 422

3.3 Verbo *ser* como verbo pleno

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

1) Como na **Figura 4: 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*).

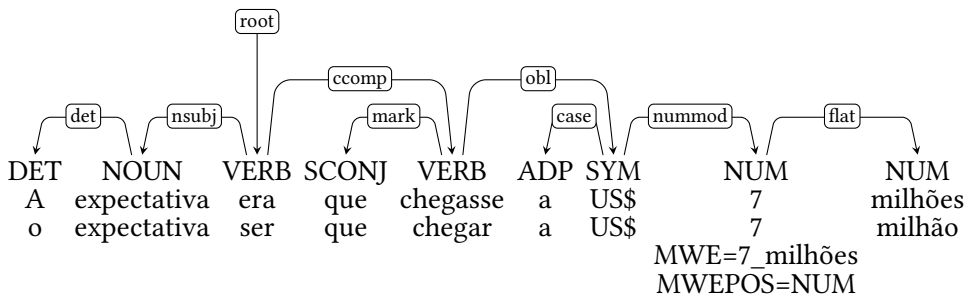


Figura 4: A expectativa *era* que chegasse a US\$7 milhões

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 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.

5.1 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 7: 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	-	Gender=Fem Number=Sing	4	amod	-	-
7	de	de	ADP	-	-	8	case	-	-
8	todos	todo	PRON	-	Gender=Masc Number=Plur Prontype=Tot	6	obl	-	SpaceAfter=No
9	.	.	PUNCT	-	-	4	punct	-	SpaceAfter=No

Figura 8: Martin é o cara *mais inteligente de todos*.

5.2 Frases do Bosque-UD

Abreviações

Agradecimentos

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.