Linguística Computacional, Corpus e Humanidades Digitais

# 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

# 2 Formato UD

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

Tabela de conteúdos

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

Tabela de conteúdos

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 Classes gramaticais (upos)

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 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 |

## 1 *Primeiro* lugar: adjetivo ou numeral?

Tabela de conteúdos

Numerais ordinais escritos por extenso devem ser anotados como *ADJ*, e recebem a feature *NumType=Ord*, como na Figura 1: Anotação do sintagma *primeira tentativa*.

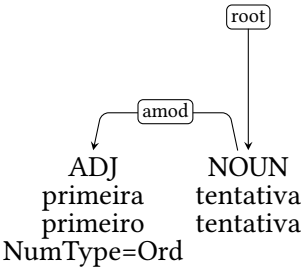


Figura 1: Anotação do sintagma *primeira tentativa*

## 2 Verbos de ligação

Tabela de conteúdos

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 2: O preço *é* de US$ 422.

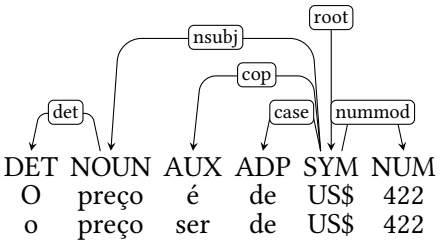


Figura 2: O preço *é* de US$ 422

# 3 Verbo *ser* como verbo pleno

Tabela de conteúdos

Atenção para casos em que o "ser" deve ser *VERB*.

1) Como na Figura 3: 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*).

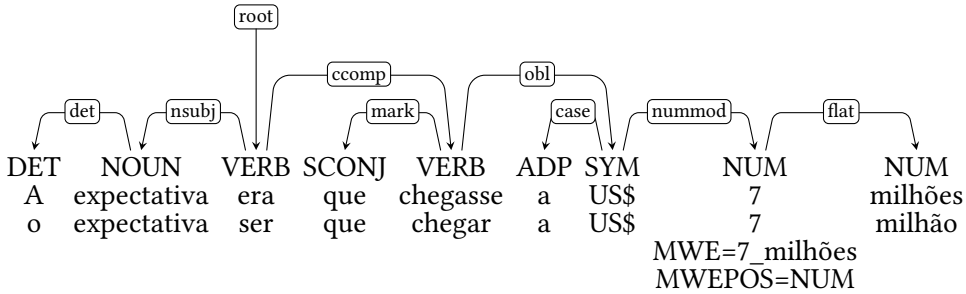


Figura 3: A expectativa *era* que chegasse a US$7 milhões

2) "ser" verbo intransitivo (verbo pleno) também deve ter a anotação *VERB* (Figura 4: Isso *foi* nos Estados Unidos).

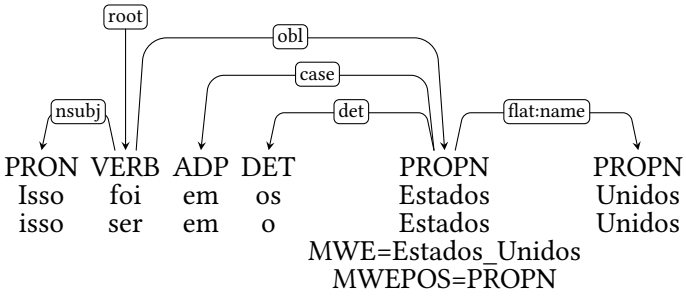


Figura 4: Isso *foi* nos Estados Unidos

## 4 Verbo *ser* como voz passiva

Tabela de conteúdos

A anotação de "ser" como voz passiva é diferente da anotação do verbo de ligação (Seção 2: Verbos de ligação) e da anotação de "ser" como verbo pleno (Seção 3: Verbo *ser* como verbo pleno).

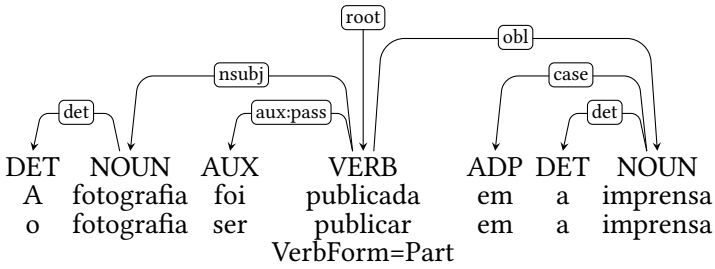


Figura 5: A fotografia *foi* publicada na imprensa

# 4 Atributos morfológicos (feats)

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

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

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

# 1 Estruturas comparativas

Tabela de conteúdos

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

Tabela de conteúdos

```
Eu coloquei tanta farinha quanto a receita pedia.
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.

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

## 1.2 Frases do Bosque-UD

Tabela de conteúdos

# Abreviações

Tabela de conteúdos

# Agradecimentos

Tabela de conteúdos

## Referências

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-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.