Anotação UD x Google

Links úteis

- Manual de anotação Google (link) - válido para todas as línguas; não é específico para o pt
- Manual de anotação UD (link)
- Interface gráfica para testar análise sintática automática do Google (link)
- Treebank PUD (ud e github)

Trajetória

- Trabalho no Google (2015 a 2019)
  - Anotação de corpus + outros projetos

- Trabalho na Redação Nota 1000 (2019 até hoje)
  - Correção automática de redações
    - Uso de análise sintática automática em regras (chamadas via API)
    - Comparação de vários parsers/taggers/tokenizadores/lematizadores/NER
    - Script de tradução e conversão do formato Google para UD
    - Pequeno experimento de revisão manual de anotações automáticas para treinamento de modelos
    - Não fazemos anotação manual atualmente

Tokenização

- Contrações não são splitadas:
  - das, neste, pela, dela, ...
- Palavras compostas são splitadas
  - guarda-chuva, ex-presidente, jacaré-açu
- Verbo + clítico são splitados
  - sabe-se, fazendo-a, deu-lhe
- Particularidades de alfanuméricos
  - token único: G1, R7, 4Rs, m2, CO2
  - tokens separados: Lei n.26, qd13, lt11, 3l, 5m

Etiquetas morfossintáticas – POS

UD

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Google

UPOS	XPOS
coarse tags	fine tags
ADJ	ADJ
ADP	IN | INDT | INPDEM | INP
ADV	ADV
VERB	VBC | VBG | VBI | VBN
CONJ	CONJ
DET	DET
INTJ	INTJ
NOUN	NN | NNP | ADD
NUM	NUM
PRON	PRP | PDEM | POSS | WH
PUNCT	PUNCT
SYM	SYM
X	X

Etiquetas morfossintáticas – POS

UD

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Google

UPOS	XPOS
coarse tags	fine tags
ADJ	ADJ
ADP	IN | INDT | INPDEM | INP
ADV	ADV
VERB	VBC | VBG | VBI | VBN
CONJ	CONJ
DET	DET
INTJ	INTJ
NOUN	NN | NNP | ADD
NUM	NUM
PRON	PRP | PDEM | POSS | WH
PUNCT	PUNCT
SYM	SYM
X	X

POS: PROPN vs. NNP ou proper

- Google não tem POS PROPN, mas tem o XPOS NNP, usada apenas para substantivos.
- Outras classes de palavras mantêm sua categoria POS normal e adiciona feature morfológica: 'proper'

```
Adriana mora em Belo Horizonte.
```

TEXT	LEMMA	POS	TAG	DEP	HEAD
Adriana	adriana	NOUN	NNP__feminine__singular__proper	NSUBJ	mora
mora	morar	VERB	VBC__imperfective__indicative__singular__third__not_proper__present	ROOT	mora
em	em	ADP	IN__not_proper	PREP	mora
Belo	belo	ADJ	ADJ__masculine__singular__proper	AMOD	Horizonte
Horizonte	horizonte	NOUN	NNP__masculine__singular__proper	POBJ	em
.	.	PUNCT	PUNCT__not_proper	P	mora

POS: PROPN vs. NNP ou proper

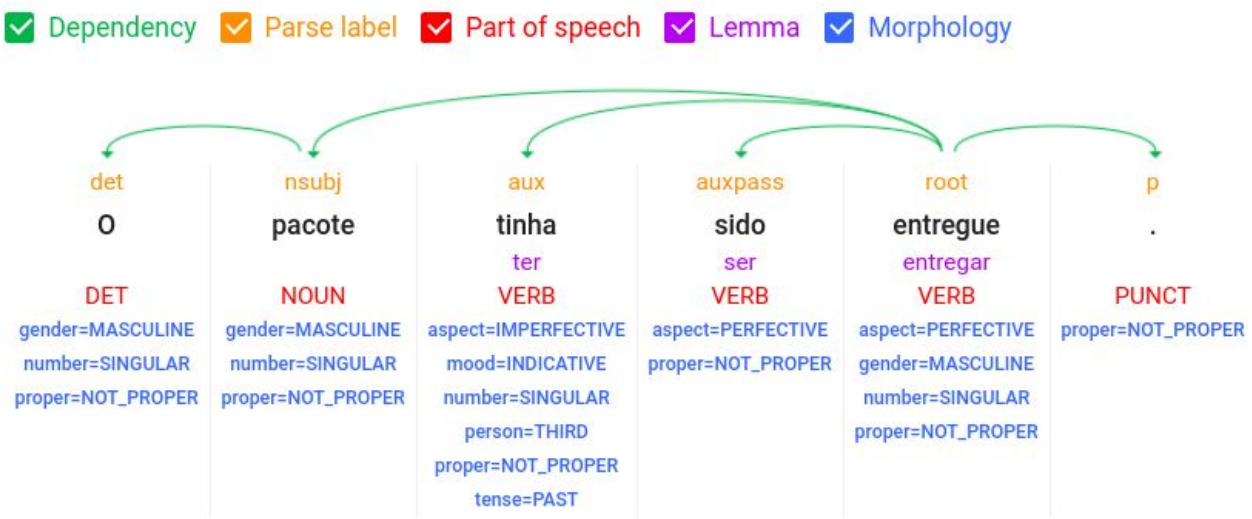
- TODOS os tokens precisam ser marcados com a tag morfológica de 'proper' ou not_proper. Por default, são 'not_proper'.
  - Decisões do que seria nome próprio foram tomadas seguindo critérios: antropônimo > topônimo > knowledge graph > por campo semântico > wikipedia (termo capitalizado ou não)

- Nomes que não existem como palavra conhecida da língua são taggeados como NNP_proper
  - Amanda, Thiago, Google, IBM, Araraquara

- Nomes próprios composicionais são anotados conforme seu POS de palavra comum e depois aplicada a tag morfológica 'proper'
  - Nova Granada, Belo Horizonte, Rio de Janeiro, Ministério da Saúde

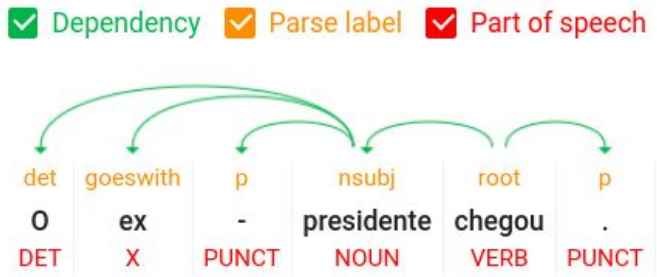
POS: AUX vs. VERB

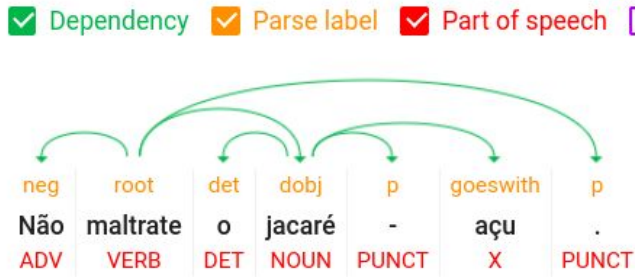
- Todo auxiliar é taggeado em POS como VERB pelo Google. A distinção só é feita no nível de dependências com deprel: aux ou auxpass

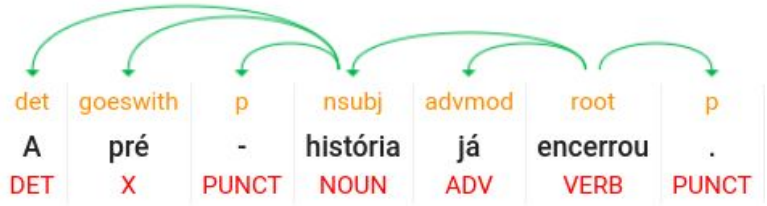


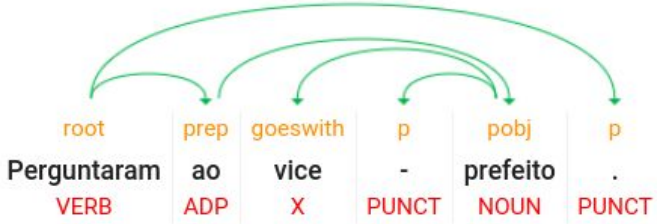
POS: PART vs. X?

- UD tem PART para o caso de prefixos, que o Google anota como X 😵









POS: CCONJ + SCONJ vs. CONJ + ADP

- O que UD chama de CCONJ o Google chama de CONJ.
- O que UD chama de SCONJ o Google chama de ADP.
- Google junta preposição e conjunção subordinativa na mesma tag.
- Justificativas:
  - As preposições acidentais também podem ser conjunções subordinativas:
  - Essas palavras são conjunções quando ligam verbos, mas, se o verbo está elíptico, ligam nomes
  - Essa diferença pode ser retomada no nível das dependências, como deprel prep (para preposições) ou mark (para conjunções subordinativas)
  - Cite (link): This treatment provides parallelism between different constructions across and within languages. A good result is that we now have greater parallelism between prepositional phrases and subordinate clauses, which are often introduced by a preposition in some languages (but note that the relation should be `mark` in those cases)

Etiquetas morfológicas

UD tem muito mais etiquetas morfológicas: morph features

Lexical features*	Inflectional features*	
	*Nominal**	*Verbal**
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	NounClass	Tense
Reflex	Number	Aspect
Foreign	Case	Voice
Abbr	Definite	Evident
Typo	Degree	Polarity
		Person
		Polite
		Clusivity

feature	valores
number	singular | plural
person	first | second | third
gender	feminine | masculine
case	accusative | dative | nominative | prepositional
tense	conditional | future | past | present | pluperfect
aspect	perfective | imperfective
proper	proper | not_proper

Etiquetas sintáticas – Dependências

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	MWE	Loose	Special	Other
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Deprel mais comuns: acomp, advcl, advmod, amod, appos, attr, aux, auxpass, cc, ccomp, conj, csubj, csubjpass, dep, det, discourse, dobj, goeswith, iobj, mark, mwe, nn, npadvmod, nsubj, nsubjpass, num, parataxis, predet, pobj, pcomp, p, prt, rcmod, root, vmod, xcomp

Outras menos frequentes: dislocated, expl, foreign, kw, list, remnant, reparandum, tmod, vocative

Etiquetas sintáticas – Dependências

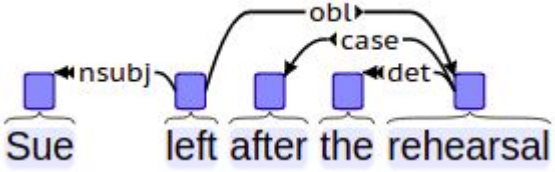
	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

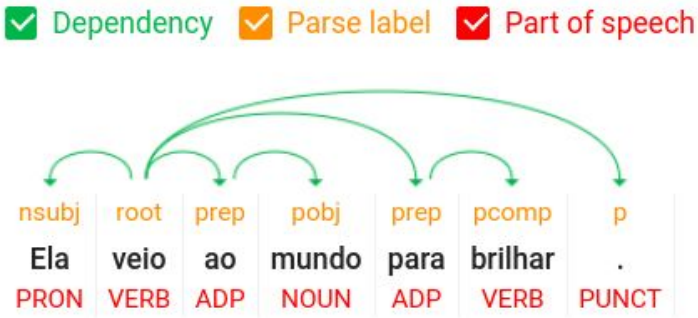
Deprel mais comuns: acomp, advcl, advmod, amod, appos, attr, aux, auxpass, cc, ccomp, conj, csubj, csubjpass, dep, det, discourse, dobj, goeswith, iobj, mark, mwe, nn, npadvmod, nsubj, nsubjpass, num, parataxis, predet, pobj, pcomp, p, prt, rcmod, root, vmod, xcomp

Outras menos frequentes: dislocated, expl, foreign, kw, list, remnant, reparandum, tmod, vocative

Deprel: obl + case vs. prep+pobj|pcomp

- Direção da seta:
  - Na UD: 'obl' é head e 'case' é dependente
  - No Google: 'prep' é head e 'pobj'|'pcomp' é dependente.
- Ud diferencia 'obl' de 'nmod'. Google não faz essa distinção
- 'Pobj' é o objeto da preposição, usado para nomes e adjetivos. 'Pcomp' é o complemento da preposição, usado para verbos.

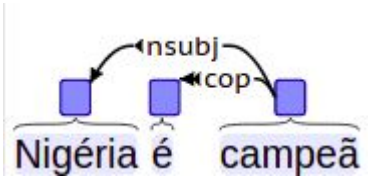


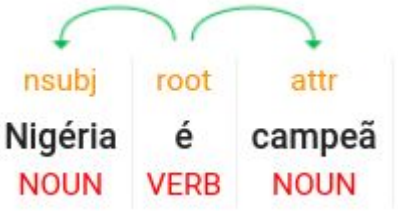


Deprel: cop + root vs. root + attr|acomp

- Google NÃO tem deprel cópula. Considera os verbos de cópula como verbos normais, podendo até ser a root, e só faz distinção quanto ao objeto, que é chamado de 'attr' ou 'acomp'.

- UD considera o predicativo como head e a copula como dependente.
- Google considera o verbo como head e o predicativo como dependente.



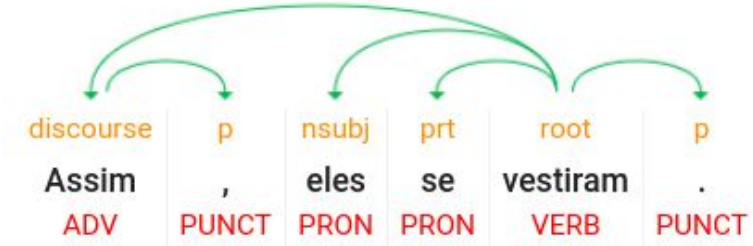


Expressões multipalavras

- UD tem 3 deprels para MWE: 'fixed', 'flat' e 'compound'.
- Google tem 3 deprels para MWE: 'mwe', 'nn' e 'goeswith'.
- A deprel 'fixed' parece corresponder à deprel 'mwe', que só pode marcar locuções prepositivas, conjuntivas e algumas adverbiais de uma lista fechada (rol finito). A diferença é que todos os filhos de 'fixed' saem do 1o elemento, enquanto na mwe as relações são subsequentes.
- A deprel 'flat' equivale à 'nn' exclusivamente para antropônimos.
- Google não tem deprel 'compound'. Casos específicos de palavras compostas ou qualquer sequência de tokens que poderia virar um token único é chamado de 'goeswith'.
- A deprel 'goeswith' da UD é usada igualmente no Google, mas o Google também usa essa deprel para casos de palavras compostas hifenizadas ou não (pré conceito, pré-conceito, pre conceito). No Google, essa deprel pode ter qualquer direção.

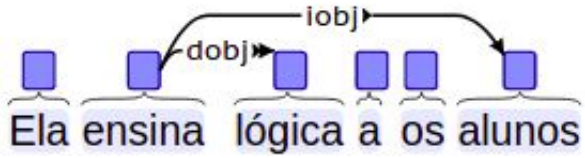
O que é considerado discourse?

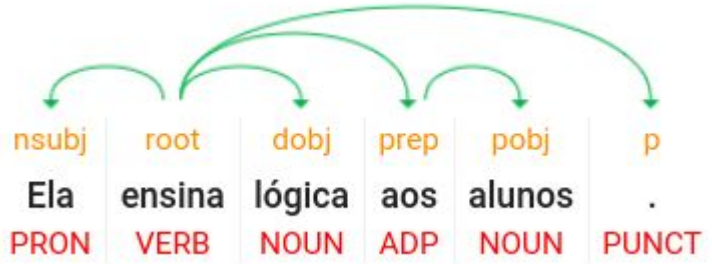
- Google chama de 'discourse' todos os casos da UD, mas também operadores discursivos:
  - conjunções subordinativas em frases mal escritas que não possuem oração principal. Daí considera-se a oração subordinada como root e marca-se 'discourse' naquilo que seria um 'mark'.
  - Advérbios e locuções adverbiais cujo escopo é a frase inteira.



Deprel: obj|iobj vs. dobj|iobj

- A deprel 'obj' parece corresponder exatamente à 'dobj'
- A deprel 'iobj' da UD só é usada pelo Google nos casos de pronomes dativos:
  - Karina **lhe** deu isso.
- Quando UD usa 'iobj' + 'case', o Google usa as relações de 'prep' e 'pobj', não fazendo distinção entre adjuntos e complementos verbais 😍





O que corresponde a 'rcmod'?

- Google chama de 'rcmod' as “relative clauses”, ou seja, as orações subordinadas adjetivas restritivas e explicativas. Como isso é feito pela UD?