

RELATÓRIO ANUAL PIBIC (2018-2019) CONSTRUÇÃO DE DATASETS PARA O PLN DE LÍNGUA PORTUGUESA

Aluno: Elvis de Souza
Orientadora: Cláudia Freitas

1. Apresentação

O presente relatório corresponde às atividades realizadas entre agosto de 2018 e junho de 2019 no projeto Construção de datasets para o PLN de língua portuguesa. De agosto de 2018 a janeiro de 2019 atuei como PIBIC voluntário, em colaboração com a então bolsista Luísa Rocha. Desde fevereiro de 2019, sou PIBIC bolsista do projeto. Deste modo, o relatório descreve atividades realizadas por Luisa Rocha e por mim (seção 2), e a partir de então, apenas por mim (seções 3, 4 e 5).

O projeto “Construção de datasets para o PLN de língua portuguesa” tem como objetivos a revisão e ampliação do corpus Bosque segundo o modelo de anotação do projeto Universal Dependencies. Atualmente com 9 mil frases, pretendemos ampliar o material até chegarmos a 15 mil frases. Como objetivo secundário está o teste de uma metodologia relativa à otimização do processo de revisão da anotação linguística. Esboçada em [1], a metodologia se sustenta na hipótese de que dois sistemas, sobretudo se elaborados segundo princípios distintos, não farão análises erradas idênticas, isto é, não errarão da mesma maneira. Assim, a estratégia de revisão da anotação consiste na revisão de frases cuja análise – realizada por dois (ou mais) sistemas - foi divergente, o que é observado a partir da análise da matriz de confusão. Como resultado da criação do dataset, espera-se uma melhoria dos resultados dos sistemas baseados em aprendizado automático, como o UDPipe [4] (sistemas open source e gratuitos), bem como (b) a criação de um cenário para o estudo sobre o impacto de diferentes tagsets em tarefas de PLN.

2. Análise do impacto da correção de part-of-speech no corpus Bosque-UD

Em colaboração com a bolsista Luisa Rocha, realizamos esse experimento com o intuito de verificar o possível impacto das correções feitas no corpus UD_Portuguese-Bosque no aprendizado do parser automático UDPipe. Essas correções são o resultado da análise das Matrizes de Confusão de Part-Of-Speech (já documentadas em [2]), que como resultado final teve 303 tokens com PoS alterados (0.14% dos tokens do corpus).

Além dessas correções, também há as correções das convergências: ao analisar matriz de confusão, que retorna divergências entre modelo treinado e golden, supõe-se que as convergências estão certas. Contudo essa suposição não é completamente verdadeira, como a análise das convergências mostrou. 47 frases apresentaram erros na PoS, e 55 tokens tiveram

PoS corrigidas. Ao total, as correções derivadas da análise de convergências foram de 358 tokens, o que corresponde a 0,16% do corpus.

Primeiramente, precisamos treinar no UDPipe dois modelos diferentes do Bosque-UD: o primeiro, anterior às correções, e o segundo, posterior.

O arquivo de treino, chamado "mais-antiga.conllu", referente ao corpus anterior às correções de PoS, foi obtido do commit de número: 7a65972e77e153ad24ddd3f761f0976ff52da063 do Bosque-UD, no GitHub¹. O arquivo "depois-convergencia.conllu", referente ao corpus posterior às correções, é o commit de número: 5ccf8baa7810c297271c3f560215e64ebccd5b93, no GitHub. A única diferença entre os dois arquivos são as 358 correções. Uma vez baixados os repositórios, juntamos os arquivos da pasta "documents/" em arquivos de treino, desenvolvimento e teste, utilizando o script "generate_release.py"², desenvolvido por mim.

Outra etapa importante, para otimizar o processo de treinamento do UDPipe, foi o de tratamento dos arquivos de treino (train + dev) gerados na etapa acima. A partir do código "tratar_conllu.py", removemos as colunas XPOS e Misc dos arquivos de treino "mais-antiga" e "depois-convergencia", para que o UDPipe não tivesse que se preocupar em aprender dados irrelevantes para o nosso propósito e, assim, otimizar nosso tempo.

O treino dos dois modelos, simultaneamente, durou cerca de 5 horas, utilizando-se o seguinte comando:

```
$ ./udpipe-1.2.0 --train --tokenizer=none --tagger --parser
```

Não treinamos o tokenizador para nenhum dos modelos uma vez que, em momento algum, o utilizaríamos, pois sempre damos ao UDPipe materiais já previamente tokenizados.

Já com os modelos "mais-antiga.udpipe" e "depois-convergencia.udpipe" treinados, o segundo passo é avaliar qual modelo aprendeu melhor. Utilizamos o script tokenizar_conllu.py para remover a anotação da partição de teste do Bosque "depois-convergencia". Com o arquivo cru (sem anotação), mas já tokenizado, utilizamos o programa udpipe_vertical.py para anotar a partição teste do Bosque com os modelos "mais-antiga.udpipe" e, posteriormente, "depois-convergencia.udpipe". Uma vez com os resultados das anotações, utilizamos o programa de avaliação oficial do UDPipe "conll18_ud_eval.py" com o parâmetro "-v" ativado para gerar a comparação entre as anotações de cada modelo e o golden do Bosque depois-convergencia, partição teste.

Os resultados completos podem ser conferidos nas imagens a seguir:

¹ http://github.com/UniversalDependencies/UD_Portuguese-Bosque

² <http://github.com/alvelvis/ACDC-UD>

Mais-antiga

```

elvis@elvis-PC024:/mnt/mmcblk2/Dropbox/PIBIC/ACDC-UD$ python3
conll18_ud_eval.py -v Experimento\ Revisão\ de\ POS/
pt-ud-test-depois-convergencia.conllu Experimento\ Revisão\ de\
POS/sistema_teste_mais-antiga.conllu
Metric | Precision | Recall | F1 Score | AligndAcc
-----+-----+-----+-----+-----
Tokens | 100.00 | 100.00 | 100.00 |
Sentences | 100.00 | 100.00 | 100.00 |
Words | 100.00 | 100.00 | 100.00 |
UPOS | 95.85 | 95.85 | 95.85 | 95.85
XPOS | 5.88 | 5.88 | 5.88 | 5.88
UFeats | 94.98 | 94.98 | 94.98 | 94.98
AllTags | 4.79 | 4.79 | 4.79 | 4.79
Lemmas | 96.71 | 96.71 | 96.71 | 96.71
UAS | 84.66 | 84.66 | 84.66 | 84.66
LAS | 81.09 | 81.09 | 81.09 | 81.09
CLAS | 73.92 | 73.45 | 73.68 | 73.45
MLAS | 65.07 | 64.66 | 64.87 | 64.66
BLEX | 70.40 | 69.95 | 70.18 | 69.95
    
```

Depois-convergencia

```

elvis@elvis-PC024:/mnt/mmcblk2/Dropbox/PIBIC/ACDC-UD$ python3
conll18_ud_eval.py -v Experimento\ Revisão\ de\ POS/
pt-ud-test-depois-convergencia.conllu Experimento\ Revisão\ de\
POS/sistema_teste_depois-convergencia.conllu
Metric | Precision | Recall | F1 Score | AligndAcc
-----+-----+-----+-----+-----
Tokens | 100.00 | 100.00 | 100.00 |
Sentences | 100.00 | 100.00 | 100.00 |
Words | 100.00 | 100.00 | 100.00 |
UPOS | 96.01 | 96.01 | 96.01 | 96.01
XPOS | 5.88 | 5.88 | 5.88 | 5.88
UFeats | 94.98 | 94.98 | 94.98 | 94.98
AllTags | 4.79 | 4.79 | 4.79 | 4.79
Lemmas | 96.92 | 96.92 | 96.92 | 96.92
UAS | 84.86 | 84.86 | 84.86 | 84.86
LAS | 81.10 | 81.10 | 81.10 | 81.10
CLAS | 73.55 | 73.09 | 73.32 | 73.09
MLAS | 65.21 | 64.80 | 65.00 | 64.80
BLEX | 70.24 | 69.79 | 70.02 | 69.79
    
```

Modelo	UPOS	UFeats	Lemmas	ULAS	LAS
Antes da revisão de POS	95.85	94.98	96.71	84.66	81.09
Depois da revisão de POS	96.01	94.98	96.92	84.86	81.10

Tabela 1: métricas relevantes

A tabela 1 mostra as métricas importantes para nós e as compara de maneira mais fácil. Como os números foram iguais para precision, recall e F1 score para as métricas na tabela, só foi colocado um número.

As métricas importantes para nós são: UPOS, UFeats, Lemmas, UAS e LAS. Todas as mudanças entre um treino e outro foram abaixo de 1%, ou seja, não houve mudança

significativa em nenhum aspecto. As diferenças entre os testes foram, em UPOS, uma melhoria de 0,16%; em UFeats não houve melhora; em Lemmas, de 0,21%; em UAS, de 0,20%; e em LAS, só de 0,01%.

A correção derivada da matriz de confusão foi de 2% das POS do Bosque, o que pode ter relação com as pequenas melhorias. Como a correção foi na POS era esperado que essa porcentagem tivesse a maior melhora, contudo foi em Lemmas que os números aumentaram mais (0,21%), seguido de UAS (0,20%). Com as baixas mudanças, não achamos que se possa dizer que houve um impacto importante em UAS nem em LAS.

3. Sobre o desenvolvimento de uma Estação de Trabalho

Após o exercício com classes de palavras, passamos para a revisão da anotação sintática. Desde a experiência anterior com a revisão de POS, ficou clara a necessidade de desenvolver uma ferramenta para auxiliar na revisão/edição e na posterior avaliação das correções. Começou a ser desenvolvida uma Estação de Trabalho (ET) para auxílio nessas tarefas.

3.1. Introdução

Sistemas de anotação morfossintática, ou parsers, que utilizam tecnologia de aprendizado de máquina, demandam corpora volumosos e bem anotados para aprender a anotar textos adequadamente. De modo geral, visando melhorar a qualidade dos anotadores automáticos, já se fez muito em relação à tecnologia subjacente aos sistemas de parsing, o que elevou drasticamente, no decorrer dos anos, a qualidade da anotação gramatical.

No entanto, há um gargalo, por exemplo, em relação à anotação de classes de palavras — por mais que se avance na tecnologia, os resultados para POS (part-of-speech, ou classes gramaticais) não conseguem ultrapassar o marco de 97% de acertos para nenhuma língua utilizando nenhum sistema em específico. No corpus Bosque-UD [3], de textos jornalísticos em Língua Portuguesa adaptados para o projeto Universal Dependencies [4], esse gargalo é relativamente confortável pois as métricas de POS são altas: 96,46% de acertos na sua versão 2.3. Contudo, em outros níveis de análise linguística, como o de relações sintáticas, e em língua portuguesa, os resultados indicam que ainda há um grande espaço para melhorias.

Sugerimos, junto com Manning [5], que um caminho para superar esses gargalos deve ser pela via linguística: melhorando a anotação dos corpora que servem de treino para os sistemas de parsing, tornando-os mais consistentes e eliminando possíveis erros humanos.

Então relatamos a construção de uma estação de trabalho desenhada a partir da perspectiva linguística, com o objetivo de facilitar a revisão, a edição e a avaliação de corpora anotados, alinhando o trabalho feito pelos especialistas em língua, de um lado, e os resultados práticos, isto é, o desempenho de sistemas de PLN (Processamento (automático) de Linguagem Natural), de outro. Desse modo, quaisquer discussões teóricas sobre as categorias gramaticais podem ser embasadas não apenas quanto à adequação linguística a certas teorias, como se tem feito na tradição gramatical, mas também nos resultados empíricos de sistemas de aprendizado artificial.

3.2. Arquitetura da ET

Nossa Estação de Trabalho (ET) compreende dois eixos centrais: 1) a avaliação e a interpretação dos resultados da análise automática, e 2) a revisão do corpus anotado que lhe serve de treino e avaliação. Apresentamos a arquitetura das ferramentas participantes da nossa ET tendo em vista que foram desenvolvidas objetivando duas tarefas complexas, em momentos distintos: primeiro, a contabilização de sujeitos ocultos no corpus Bosque-UD, e posteriormente, o lançamento de uma nova versão do mesmo corpus.

É importante ressaltar, no entanto, que, apesar de estarmos disponibilizando as ferramentas publicamente, nosso objetivo, no lugar de simplesmente compartilhar os códigos empacotados (disponíveis em [6]) e descrever como os implementamos, é relatar o que nos levou a desenvolvê-los, destacando a importância de estruturar os fenômenos nos corpora sob diferentes perspectivas que possam motivar o trabalho linguístico no PLN, fomentando o diálogo entre as duas áreas.

De modo geral, as métricas de avaliação de sistemas de PLN são informativas do ponto de vista computacional, mas pouco nos dizem sobre quais categorias linguísticas podem ser melhoradas no corpus com que trabalhamos. Pensando na integração entre as duas áreas, desenvolvemos algumas maneiras de nos auxiliar no processo de correção de um corpus ao direcionar nosso olhar para as categorias que apresentam maior quantidade de problemas. Primeiro, visualizamos as porcentagens de erros da anotação automática para cada categoria linguística: classe gramatical, função sintática, etc. Depois, utilizamos matrizes de confusão para verificar, dentre os erros, de que forma os sistemas erraram e quais são as tais frases em que o sistema cometeu equívocos.

Em posse das frases que o sistema errou, podemos generalizar regras de conversão que nos auxiliem na correção do corpus que serve de treino ao sistema, ou podemos pensar em sintaxes de busca que nos retornem todas as frases com possíveis erros para que possamos corrigi-las manualmente. A etapa de busca das sentenças e correção foi realizada utilizando o Interrogatório [5], um ambiente de busca em arquivos no formato CoNLL-U escrito para o projeto de pesquisa em Python e Javascript e em franca expansão.

Com o Interrogatório, hoje podemos fazer pesquisas nos corpora a partir de 5 diferentes critérios de busca. Dentro das buscas, ainda, podemos realizar pesquisas paralelas que funcionam como filtros para a pesquisa inicial. No contexto de correção de um corpus, filtros de pesquisa são de extrema relevância porque, caso a correção seja feita manualmente — e, em alguns casos, elaborar regras de conversão pode ser mais humanamente custoso do que corrigir as frases manualmente —, é importante que possamos reunir todas as frases com problemas similares em uma só página. Assim, conseguimos alterar frases parecidas utilizando uma só ou poucas estratégias de revisão, tornando o processo mais eficiente.

Uma vez tendo sido feitas as alterações pretendidas no corpus, podemos realizar o treinamento e a avaliação do sistema novamente. Voltando à etapa primeira da ET, comparamos os resultados das versões e, então, começamos novos experimentos, de tal modo que as duas etapas da ET se retroalimentam.

4. Primeira utilização da ET: identificação de frases com sujeito oculto

Tendo desenvolvido nossa Estação de Trabalho, foi possível fazer pesquisas mais complexas dentro do projeto, como a busca por sujeitos ocultos.

4.1. Motivação

O sujeito oculto é um fenômeno muito comum na língua portuguesa para evitar repetição de palavras ou criar um estilo de escrita. Contudo, isso dificulta a extração automática de informação, pois deixa dependente do contexto e da semântica a informação de quem está realizando ou sofrendo a ação do verbo.

Por isso, avaliar a quantidade de ocorrências do fenômeno sujeito oculto na língua portuguesa é importante. Para a extração de informação automática, é importante saber como lidar com esse fenômeno.

Não fizemos distinções entre sujeito oculto e sujeito indeterminado - ou seja, nossa busca/contagem pelas frases com "sujeito oculto" considera ambos os casos, já que, em ambos, não há sujeito explícito, ainda que exista um sujeito. A gramática de Cunha & Cintra [7], por exemplo, considera o primeiro "sujeito oculto (determinado)" e o segundo "sujeito indeterminado", sendo a diferença entre eles a possibilidade de determinação do sujeito pela desinência do verbo. Deixamos de fora as orações sem sujeito como frases com verbos *haver* impessoais (Exemplo 3), frases sem verbo (Exemplo 4) e fenômenos da natureza.

EXEMPLOS:

Exemplo 1: CP31-3 *Sempre que surge um problema, **chamam-na**.*

Exemplo 2: CP22-3 *Eu tentei, o senhor Vance tentou, se for respeitado, **urrah!**», **comentou**.*

Exemplo 3: CP23-8 ***Há**, no ar, uma certa ideia de invasão.*

Exemplo 4: CP1-1 *Um revivalismo refrescante.*

4.2. Pesquisa

A pesquisa foi realizada nos corpora Bosque, composto por 9366 sentenças de textos jornalísticos, DHBB, composto por sentenças do Dicionário Histórico Biográfico Brasileiro, publicado pelo CPDOC/FGV, e uma fração do OBras, composto por sentenças de obras da literatura brasileira. Para realizar as pesquisas, utilizamos o ambiente de pesquisa em corpora anotados Interrogatório. Os corpora são adaptações dos respectivos corpora originais para o formato Universal Dependencies, que contém informações de número do token, word, lemma, UPOS, XPOS, feats, deprel, deprel e misc em colunas, totalizando 10 colunas para cada token de cada sentença.

Primeiro, queríamos achar todas as orações principais que não têm um token com relação de sujeito, como o exemplo 2, no qual *comentou* é o *root* da frase e não possui um sujeito. Foi adotado o critério de pesquisa 2 do Interrogatório (usado para buscas que envolvem ausência, isto é, não existe um B que aponta para A), com a seguinte expressão:

2 root#8#nsubj|csubj|nsubj:pass#8

que significa “um token A marcado como *root* na coluna oito, que não tenha nenhum token B como *nsubj* ou *csubj* ou *nsubj:pass* na coluna oito também, apontando para este token A” (exemplo 2). A pesquisa retornou 2774 frases, para o Bosque. Dentro deste resultado, outros tipos de frases, além do fenômeno procurado, apareceram, como frases com verbo *haver* impessoal (exemplo 3) -- construções que envolvem, necessariamente, um verbo sem sujeito, e *roots* que não eram verbos (e nem tinham relação de *cópula (cop)*), como manchetes (exemplo 4). Foi então necessário excluir também esses tipos de frases.

Para retirar as ocorrências de “*Haver*” impessoal, a expressão foi:

1 `haver.*Number=Sing|Person=3.*root`

que significa, usando o critério 1, Expressão Regular (Regex), procurar frases com lema *haver* e features de terceira pessoa do singular.

Depois utilizamos outra expressão para retirar as frases sem verbos como as do exemplo 4:

4 `!tAUXt.*cop :: ^(?!.*VERB).*root`

A expressão utiliza o critério 4 de busca (Pais e filhos), criado para procurar relações de dependência. A expressão acima retira todas as frases cujo *root*, ou raiz da sentença, não é um *VERB* (`^(?!.*VERB).*root`) e não tem uma relação de *cop*, isto é, verbo de cópula (`!tAUXt.*cop`).

Fenômenos da natureza não foram encontrados no DHBB; no Bosque, houve apenas 1 ocorrência e, no OBRas, 27. Utilizamos a seguinte busca:

`\t(chover|ventar|anoitecer|amanhecer|entardecer|relampejar|trovejar|escurecer|clarear)t.*root`

Depois de aplicados os três filtros, tivemos o resultado final de 1503 sentenças, 16.04% do Bosque-UD. Uma porcentagem significativa para justificar um jeito melhor de lidar com esse fenômeno nas tarefas de extração de informação.

No DHBB, utilizando os mesmos critérios e expressões de busca, a busca retornou 127.741 frases. Isso significa que 39.5% do corpus DHBB apresenta sujeitos ocultos.

No OBRas, por sua vez, a porcentagem de sujeitos ocultos ficou em 28.4% do corpus.

Os resultados estão resumidos na tabela 1.

	Bosque-UD	DHBB-UD	Machado de Assis-UD Romances Crônicas Contos
--	------------------	----------------	--

Tokens	244.628	12.214.414	2.699.842 726.016 721.715 1.252.111
Sentenças	9.366	323.301	145.756 38.256 33.409 74.091
Sujeitos ocultos	16,04% (1.503)	39,5% (127.741)	28.42% (41.424) 16,54% (12.012) 28,74% (9.603) 26,73% (19.809)

Tabela 1: distribuição dos sujeitos ocultos entre os corpora

5. Revisão geral do Bosque-UD

Com a ET como interface para o corpus, foi possível também enfrentar de maneira sistemática a revisão do corpus. Ao longo de 5 meses (de fevereiro a junho), foram corrigidos mais de 12 mil palavras no corpus Bosque-UD, processo resumido a seguir.

Revisão do aposto #242

Tipo: 1. opção de anotação, 2. correção da conversão & 3. revisão do corpus.

Total de tokens modificados no corpus: 989

1. Foi uma **opção de anotação** nossa criar o deprel “appos:parataxis” para designar os apostos cuja relação com o elemento anterior não é explicitada na frase, mas conseguimos deprender com conhecimento de mundo. Encontramos, no corpus, tokens desse tipo que estavam como “nmod” e como “appos”, a partir de algumas estratégias que originaram 615 alterações por lote e com revisão manual:

a) Procuramos por preposições (deprel “case”) com “N<PRED” na coluna do PALAVRAS. O token que é pai dessas preposições deveria, com raras exceções, se tornar “appos:parataxis”:

47	1	o	o	DET	<artd> ART M S @>N	Definite=Def Gender=Masc Number=Sing PronType=Art	2		
48	2	modelo	modelo	NOUN	<np-def> N M S @SUBJ>	Gender=Masc Number=Sing	11	nsubj:pass	
49	3	Lx	Lx	PROPN		Gender=Masc Number=Sing	2	appos	ChangedBy=Issue119 MWE=Lx_810 M
50	4	810	810	PROPN		Number=Sing	3	flat:name	ChangedBy=Issue119 ChangedBy=Is
51	5	,	,	PUNCT	PU @PU		8	punct	
52	6-7	da							
53	6	de	de	ADP	<sam-> PRP @N<PRED		8	case	
54	7	a	o	DET	<-sam> <artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art			
	- 8	Epson	Epson	PROPN	PROP F S @P<	Gender=Fem Number=Sing	2	nmod	ChangedBy=Issue165 Spac
	+ 8	Epson	Epson	PROPN	PROP F S @P<	Gender=Fem Number=Sing	2	appos:parataxis	ChangedBy=Issue:
56	9	,	,	PUNCT	PU @PU		8	punct	
57	10	é	ser	AUX	<aux> V PR 3S IND @FS-STA	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin			
58	11	vendido	vender	VERB	<pass> <mv> V PCP M S @ICL-AUX<	Gender=Masc Number=Sing VerbForm=Part Voice=Pass			

b) Procuramos por expressões entre parênteses, casos em que, salvo poucas exceções, o núcleo do sintagma dentro do parênteses deve ser “appos:parataxis”:

128	26	segundo	segundo	ADP	PRP @<ADVL		28	case	
129	27	um	um	DET	<arti> ART M S @>N	Definite=Ind Gender=Masc Number=Sing PronType=Art	28		d
130	28	júri	júri	NOUN	<np-idf> N M S @P<	Gender=Masc Number=Sing	16	obl	ChangedBy=Issue13
131	+ 29	em	em	ADP	PRP @<ADVL		30	case	
132	30	Anchorage	Anchorage	PROPN	PROP M S @P<	Gender=Masc Number=Sing	28	nmod	
133	31	((PUNCT	PU @PU		32	punct	ChangedBy=Issue165 SpaceAfter=No
	- 32	Alaska	Alaska	PROPN	PROP M S @N<PRED	Gender=Masc Number=Sing	30	appos	ChangedBy=Issue16
	+ 32	Alaska	Alaska	PROPN	PROP M S @N<PRED	Gender=Masc Number=Sing	30	appos:parataxis	ChangedBy
135	33))	PUNCT	PU @PU		32	punct	ChangedBy=Issue165 SpaceAfter=No
136	34	.	.	PUNCT	PU @PU		16	punct	
137									

c) Procuramos por números entre vírgulas, caso em que o número deve ser “appos:parataxis” do elemento anterior:

128	26	segundo	segundo	ADP	PRP @<ADVL		28	case	
129	27	um	um	DET	<arti> ART M S @>N	Definite=Ind Gender=Masc Number=Sing PronType=Art	28		d
130	28	júri	júri	NOUN	<np-idf> N M S @P<	Gender=Masc Number=Sing	16	obl	ChangedBy=Issue13
131	+ 29	em	em	ADP	PRP @<ADVL		30	case	
132	30	Anchorage	Anchorage	PROPN	PROP M S @P<	Gender=Masc Number=Sing	28	nmod	
133	31	((PUNCT	PU @PU		32	punct	ChangedBy=Issue165 SpaceAfter=No
	- 32	Alaska	Alaska	PROPN	PROP M S @N<PRED	Gender=Masc Number=Sing	30	appos	ChangedBy=Issue16
	+ 32	Alaska	Alaska	PROPN	PROP M S @N<PRED	Gender=Masc Number=Sing	30	appos:parataxis	ChangedBy
135	33))	PUNCT	PU @PU		32	punct	ChangedBy=Issue165 SpaceAfter=No
136	34	.	.	PUNCT	PU @PU		16	punct	
137									

Após testes nomeando esses casos de “appos:parataxis” de diferentes formas, tal como “nmod”, “appos” e “parataxis”, chegamos à conclusão, ao avaliar as métricas e guidelines UD, de que a melhor estratégia é chamá-los todos, indiscriminadamente, de “nmod”.

2. Foi uma estratégia nossa de **correção da conversão** do PALAVRAS olhar para as sentenças em que havia ocorrência de “, o que”, em que o verbo que sucede a expressão tinha a etiqueta “FS-S<” na coluna do PALAVRAS, casos em que “o” é pronome relativo e aposto:

47	7	um	um	DET	<arti> ART M S @>N	Definite=Ind Gender=Masc Number=Sing PronType=Art	8	d
48	8	caráter	caráter	NOUN	<np-idf> N M S @<ACC	Gender=Masc Number=Sing 3 obj		
49	9	excessivamente	excessivamente	ADV	ADV @>A	10 advmod		
50	10	defensivo	defensivo	ADJ	ADJ M S @N<	Gender=Masc Number=Sing 8 amod		ChangedBy
51	11	,	,	PUNCT	PU @PU	3 punct		
	- 12	o	o	PRON		Gender=Masc Number=Sing PronType=Dem 13 det		
52	+ 12	o	o	PRON		Gender=Masc Number=Sing PronType=Dem 3 appos		
53	13	que	que	PRON	<rel> INDP M S @SUBJ>	Gender=Masc Number=Sing PronType=Rel 14 nsubj		
	- 14	desagrada	desagradar	VERB	<mv> V PR 3S IND @FS-S<	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm		
54	+ 14	desagrada	desagradar	VERB	<mv> V PR 3S IND @FS-S<	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm		
55	15-16	aqueles						

Foram 151 tokens corrigidos manualmente deste jeito.

3. Durante os processos descritos acima, realizamos, também, 223 correções manuais fruto de **revisão do corpus**.

Eliminação de dep #243

Tipo: correção da conversão do PALAVRAS.

Total de tokens modificados no corpus: 497.

“Dep” é um deprel que, a rigor, deve poder ser substituído por outras categorias sintáticas. No entanto, como eram 954 tokens com deprel “dep” na versão 2.3 do Bosque-UD (após as correções, dispomos de apenas 457 na versão 2.4), foi necessário desenvolver estratégias para corrigi-los de maneira eficaz.

Em primeiro lugar, buscamos por ocorrências de tokens com “ICL-<PRED” ou “ICL-PRED>” na coluna do PALAVRAS que tivessem o deprel “dep”. Estes poderiam ser tanto “acl” quanto “advcl”, ao que nos coube fazer filtros específicos e manuais para discriminá-los e aplicar as alterações necessárias:

5	1	A	o	DET	<artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art	2	d
6	2	exclusão	exclusão	NOUN	<np-def> N F S @SUBJ>	Gender=Fem Number=Sing 14 nsubj		
7	3	de	de	ADP	PRP @N<	4 case		
8	4	Maradona	Maradona	PROPN	PROP M S @P<	Gender=Masc Number=Sing 2 nmod		ChangedBy
9	5	,	,	PUNCT	PU @PU	8 punct		
10	6	por	por	ADP	PRP @ADVL>	8 mark		
11	7	ter	ter	AUX	<aux> V INF @ICL-P<	VerbForm=Inf 8 aux		
12	8	jogado	jogar	VERB	<mv> V PCP M S @ICL-AUX<	VerbForm=Part 14 advcl		
	- 9	dopado	dopar	VERB	<mv> V PCP M S @ICL-<PRED	Gender=Masc Number=Sing VerbForm=Part 8 dep		
13	+ 9	dopado	dopar	VERB	<mv> V PCP M S @ICL-<PRED	Gender=Masc Number=Sing VerbForm=Part 4 acl		
14	10	contra	contra	ADP	PRP @ADVL	12 case		
15	11	a	o	DET	<artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art	12	d
16	12	Nigéria	Nigéria	PROPN	PROP F S @P<	Gender=Fem Number=Sing 9 obl		ChangedBy=Issue137 Change

23	um	um	DET	<arti> ART M S @>N	Definite=Ind Gender=Masc Number=Sing PronType=Art	24	det	_
24	facto	facto	NOUN	<np-idf> N M S @P<	Gender=Masc Number=Sing 17	nmod	_	_
25	isolado	isolar	VERB	<mv> V PCP M S @ICL-N<	Gender=Masc Number=Sing VerbForm=Part	24	acl	_
26	,	,	PUNCT	PU @PU	25	punct	_	_
27	porque	porque	SCONJ	KS @SUB	30	mark	_	_
28	ele	ele	PRON	PERS M 3S NOM @SUBJ>	Case=Nom Gender=Masc Number=Sing Person=3 PronType=Prs	30	nsubj	_
29	só	só	ADV	ADV @ADVL>	30	advmod	_	_
30	faz	fazer	VERB	<mv> V PR 3S IND @FS-<ADVL	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	14	advcl	_
31	sentido	sentido	NOUN	<np-idf> N M S @ACC	Gender=Masc Number=Sing 30	obj	_	_
- 32	integrado	integrar	VERB	<mv> V PCP M S @ICL-<PRED	Gender=Masc Number=Sing VerbForm=Part	30	dep	_
+ 32	integrado	integrar	VERB	<mv> V PCP M S @ICL-<PRED	Gender=Masc Number=Sing VerbForm=Part	30	advcl	_
33-34	no	-	-	-	-	-	-	-

Buscamos, também, por casos de “ICL-<SC” na coluna do PALAVRAS e “dep” em seu deprel, tokens cujo deprel deveria se tornar “ccomp” ou, de acordo com nossas novas diretivas, “ccomp:parataxis”:

114	7	Clarismino	Clarismino	PROPN	_	Gender=Masc Number=Sing 2	appos	_	ChangedBy=Issue11
115	8	Júnior	Júnior	PROPN	_	Number=Sing 7	flat:name	_	ChangedBy=Issue119 ChangedBy=Issu
116	9	,	,	PUNCT	PU @PU	7	punct	_	_
117	10	disse	dizer	VERB	<mv> V PS 3S IND @FS-STA	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin	0	_	_
118	11	que	que	SCONJ	KS @SUB	17	mark	_	_
119	12	um	um	NUM	<card> NUM M S @SUBJ>	NumType=Card 17	nsubj	_	ChangedBy=Issue168
120	13-14	dos	-	-	-	-	-	-	-
121	13	de	de	ADP	<sam-> PRP @N<	15	case	_	_
122	14	os	o	DET	<-sam> <artd> ART M P @>N	Definite=Def Gender=Masc Number=Plur PronType=Art	1	_	_
123	15	objetivos	objetivo	NOUN	<np-def> N M P @P<	Gender=Masc Number=Plur 12	nmod	_	_
124	16	é	ser	AUX	<mv> V PR 3S IND @FS-ACC	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	1	_	_
- 17	17	identificar	identificar	VERB	<mv> V INF @ICL-<SC	VerbForm=Inf 10	dep	_	_
125	+ 17	identificar	identificar	VERB	<mv> V INF @ICL-<SC	VerbForm=Inf 10	ccomp	_	_
126	18	as	o	DET	<artd> ART F P @>N	Definite=Def Gender=Fem Number=Plur PronType=Art	19	c	_
127	19	macrozonas	macrozona	NOUN	<np-def> N F P @ACC	Gender=Fem Number=Plur 17	obj	_	_
79	1	A	o	DET	<artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art	2	d	_
80	2	idéia	idéia	NOUN	<np-def> N F S @SUBJ>	Gender=Fem Number=Sing 4	nsubj	_	_
81	3	é	ser	AUX	<mv> V PR 3S IND @FS-ACC	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	4	_	_
- 4	4	privilegiar	privilegiar	VERB	<mv> V INF @ICL-<SC	VerbForm=Inf 15	dep	_	_
82	+ 4	privilegiar	privilegiar	VERB	<mv> V INF @ICL-<SC	VerbForm=Inf 15	ccomp:parataxis	_	_
83	5	os	o	PRON	<first-cjt> <dem> DET M P @ACC	Gender=Masc Number=Plur PronType=Dem 4	obj	_	_
84	6	de	de	ADP	PRP @N<	8	case	_	_
85	7	maior	grande	ADJ	<KOMP> <COMP> ADJ F S @>N	Gender=Fem Number=Sing 8	amod	_	_
86	8	qualidade	qualidade	NOUN	<np-idf> N F S @P<	Gender=Fem Number=Sing 5	nmod	_	_
87	9	e	e	CCONJ	<co-acc> KC @CO	11	cc	d2d:#106	_
88	10	as	o	DET	<artd> ART F P @>N	Definite=Def Gender=Fem Number=Plur PronType=Art	11	d	_
89	11	feiras	feira	NOUN	<cjt> <np-def> N F P @ACC	Gender=Fem Number=Plur 5	conj	_	_
90	12	estaduais	estadual	ADJ	ADJ F P @N<	Gender=Fem Number=Plur 11	amod	_	ChangedBy
91	13	»	»	PUNCT	PU @PU	15	punct	_	ChangedBy=Issue165 SpaceAfter=No
92	14	,	,	PUNCT	PU @PU	4	punct	_	_
93	15	diz	dizer	VERB	<mv> V PR 3S IND @FS-STA	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	0	_	_
94	16	Souza	Souza	PROPN	PROP M S @SUBJ	Gender=Masc Number=Sing 15	nsubj	_	ChangedBy=Issue165 SpaceA
95	17	.	.	PUNCT	PU @PU	15	punct	_	_

Correções de relato direto #247

Tipo: conversão

Total de tokens modificados no corpus: 249.

Com o auxílio da anotação de relato direto no Bosque que está disponível no AC/DC (Linguatca), anotamos alguns verbos no Bosque-UD, que estavam com o deprel “ccomp”,

como “ccomp:parataxis”. Além disso, segundo as guidelines, *reported speech direto* deve ser parataxis, e é assim que está, hoje, na versão 2.4.

Correções de obl que deveriam ser obj ou iobj #248

Tipo: correção da conversão do PALAVRAS

Total de tokens modificados no corpus: 2068

Observamos no corpus Bosque-UD, fruto da conversão da versão PALAVRAS para UD, um excesso de “obl” em detrimento de objetos, diretos ou indiretos. Em decorrência disso, sistematizamos a seguinte correção:

- Ao procurar por ocorrências de preposições com “PIV” na coluna do PALAVRAS, e “case” na coluna deprel, o token que é pai da preposição deve ser objeto direto, caso não haja nenhum outro objeto direto, ou objeto indireto, caso já haja um objeto direto. As correções foram em lote, mas houve espaço também para poucas revisões dos casos modificados pela regra.

30	16	a	o	DET	<artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art	18	d
31	17	esmagadora	esmagador	ADJ	ADJ F S @>N	Gender=Fem Number=Sing	18	amod
32	18	maioria	maioria	NOUN	<np-def> N F S @SUBJ>	Gender=Fem Number=Sing	26	nsubj
33	19	((PUNCT	PU @PU		21	punct
34	20	77	77	NUM	<card> NUM M P @>N	NumType=Card	21	nummod
35	21	%	%	SYM	<np-def> N M P @N<PRED	Gender=Masc Number=Plur	18	appos
36	22))	PUNCT	PU @PU		21	punct
37	23-24	dos						
38	23	de	de	ADP	<sam-> PRP @N<		25	case
39	24	os	o	DET	<-sam> <artd> ART M P @>N	Definite=Def Gender=Masc Number=Plur PronType=Art		2
40	25	eleitores	eleitor	NOUN	<np-def> N M P @P<	Gender=Masc Number=Plur	18	nmod
41	26	quer	querer	VERB	<mv> V PR 3S IND @FS-N<PRED	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin		8
42	27	o	o	DET	<artd> ART M S @>N	Definite=Def Gender=Masc Number=Sing PronType=Art		28
43	28	PT	PT	PROPN	PROP M S @<ACC	Gender=Masc Number=Sing	26	obj
44	29	participando	participar	VERB	<mv> V GER @ICL-<OC	VerbForm=Ger	26	xcomp
45	30-31	do						
46	30	de	de	ADP	<sam-> PRP @<PIV		32	case
47	31	o	o	DET	<-sam> <artd> ART M S @>N	Definite=Def Gender=Masc Number=Sing PronType=Art		3
-	32	Governo	governo	NOUN	<prop> <np-def> N M S @P<	Gender=Masc Number=Sing	29	obl
+	32	Governo	governo	NOUN	<prop> <np-def> N M S @P<	Gender=Masc Number=Sing	29	obj
103	14	pediu	pedir	VERB	<mv> V PS 3S IND @FS-STA	Mood=Ind Number=Sing Person=3 Tense=Past VerbForm=Fin		6
104	15	mais	mais	ADV	<setop> ADV @A		16	advmod
105	16	um	um	NUM	<card> NUM M S @>N	NumType=Card	17	nummod
106	17	dia	dia	NOUN	<np-def> N M S @<ACC	Gender=Masc Number=Sing	14	obj
107	18	de	de	ADP	PRP @N<		19	case
108	19	folga	folga	NOUN	<np-idf> N F S @P<	Gender=Fem Number=Sing	17	nmod
109	20-21	ao						
110	20	a	a	ADP	<sam-> PRP @<PIV		22	case
111	21	o	o	DET	<-sam> <artd> ART M S @>N	Definite=Def Gender=Masc Number=Sing PronType=Art		2
-	22	treinador	treinador	NOUN	<np-def> N M S @P<	Gender=Masc Number=Sing	14	obl
+	22	treinador	treinador	NOUN	<np-def> N M S @P<	Gender=Masc Number=Sing	14	iobj
113	23	.	.	PUNCT	PU @PU		14	punct

Correção de numerais #249

Tipo: correção da conversão do PALAVRAS

Total de tokens modificados no corpus: 1093

A fim de nos aproximar mais das guidelines UD, fizemos diversas alterações em lote, com direito a revisões manuais, tratando dos numerais no Bosque-UD:

1. Tokens com lema “mil”, “milhão”, “bilhão”, “trilhão” e “milhar”, quando podem ser substituídas por algarismos, têm pos “NUM”, feats “NumType=Card”, e apontam para o primeiro número da sequência, com o deprel “flat”;
2. O primeiro número da sequência de números deve apontar para a palavra que está modificando com o deprel “nummod”, e todos os tokens que apontavam para o número, passam a apontar para o token modificado pelo número.

54	8	receber	receber	VERB	<mv> V INF @ICL-<SC	VerbForm=Inf	0	root	_	_	
-	9	7	7	NUM	<card> NUM M P @>N	NumType=Card	10	nummod	_	ChangedBy=Issue168	
-	10	milhões	milhão	NOUN	<np-def> N M P @<ACC	Gender=Masc Number=Plur	8	obj	_		
55	+ 9	7	7	NUM	<card> NUM M P @>N	NumType=Card	12	nummod	_	ChangedBy=Issue168	
56	+ 10	milhões	milhão	NUM	<np-def> N M P @<ACC	NumType=Card	9	flat	_		
57	11	de	de	ADP	PRP @N<ARG		12	case	_		
-	12	declarações	declaração	NOUN	<np-idf> N F P @P<	Gender=Fem Number=Plur	10	nmod	_		
58	+ 12	declarações	declaração	NOUN	<np-idf> N F P @P<	Gender=Fem Number=Plur	8	obj	_		
59	+ 13	.	.	PUNCT	PU @PU		8	punct	_		

Eliminação de nmod e tmod #251

Tipo: conversão de tagset (seguindo guidelines UD 2.0)

Total de tokens modificados no corpus: 794

Segundo as guidelines de UD 2.0, “nmod:npm” e “nmod:tmod” são duas categorias sintáticas que deixaram de existir na nova versão. Por isso, eliminamos todas as suas ocorrências, transformando-as ora em “nmod”, ora em “appos” (aposto especificativo).

61	5	desde	desde	ADP	PRP @<ADVL		6	case	_	
62	6	30	30	NUM	<card> NUM M S @P<	NumType=Card	4	nmod	_	ChangedBy=Issue168
63	7	de	de	ADP	PRP @N<		8	case	_	
-	8	julho	julho	NOUN	<np-idf> N M S @P<	Gender=Masc Number=Sing	6	nmod:tmod	_	
64	+ 8	julho	julho	NOUN	<np-idf> N M S @P<	Gender=Masc Number=Sing	6	nmod	_	
65	9	de	de	ADP	PRP @N<		10	case	_	
66	10	93	93	NUM	<card> NUM M P @P<	NumType=Card	8	nmod	_	ChangedBy=Issue165 Change
67	11	.	.	PUNCT	PU @PU		4	punct	_	
39	20	o	o	DET	<artd> ART M S @>N	Definite=Def Gender=Masc Number=Sing PronType=Art	21	c		
40	21	ex-lider	ex-lider	NOUN	<np-def> N M S @<ACC	Gender=Masc Number=Sing	15	obj	_	
41	22	religioso	religioso	ADJ	ADJ M S @N<	Gender=Masc Number=Sing	21	amod	_	
-	23	Jonas	Jonas	PROPN	_	Gender=Masc Number=Sing	21	nmod:npm	_	ChangedBy=Issue119 MWE=Jc
42	+ 23	Jonas	Jonas	PROPN	_	Gender=Masc Number=Sing	21	appos	_	ChangedBy=Issue119 MWE=Jonas_Rubi
43	24	Rúbio	Rúbio	PROPN	_	Number=Sing	23	flat:name	_	ChangedBy=Issue119 ChangedBy=Issu

Correção de orações relativas com "o que" #252

Tipo: opção de anotação

Total de tokens modificados no corpus: 453

Buscando por ocorrências de “o que” em frases em que há uma oração relativa (acl:relcl), encontramos um grande número de erros nos quais o pronome “o” estava com o deprel “det”, típico dos artigos. Fizemos as alterações em lote, assistidas por uma revisão manual.

250	5	um	um	DET	<-sam> <arti> ART M S @>N	Definite=Ind Gender=Masc Number=Sing PronType=Art		
251	6	ambiente	ambiente	NOUN	<np-idf> N M S @P<	Gender=Masc Number=Sing	15	advcl
252	7	como	como	ADP	<com> PRP @N<		8	case
-	8	o	o	PRON	<dem> DET M S @P<	Gender=Masc Number=Sing PronType=Dem	6	det
-	9	que	que	PRON	<rel> INDP M S @ACC>	Gender=Masc Number=Sing PronType=Rel	11	obj
253	+ 8	o	o	PRON	<dem> DET M S @P<	Gender=Masc Number=Sing PronType=Dem	6	nmod
254	+ 9	que	que	PRON	<rel> INDP M S @ACC>	Gender=Masc Number=Sing PronType=Rel	11	obl
255	10	eu	eu	PRON	PERS F 1S NOM @SUBJ>	Case=Nom Gender=Fem Number=Sing Person=1 PronType=Prs	11	
256	11	vivia	viver	VERB	<mv> V IMPF 1S IND @FS-N<	Mood=Ind Number=Sing Person=1 Tense=Imp VerbForm=Fin		
257	12	,	,	PUNCT	PU @PU		11	punct

Uniformização de AUX #255

Tipo: revisão do corpus

Total de tokens modificados no corpus: 149

A dificuldade em caracterizar as formas verbais como AUX (caso consideremos parte de locução verbal com um verbo que se segue) ou VERB (caso consideremos que forme uma oração distinta daquela do segundo verbo) nos fez olhar detidamente para estes casos. Consideramos, neste momento, não realizar mudanças bruscas, mas aprofundar as escolhas já feitas pelo PALAVRAS: se um verbo X seguido de outro, na maior parte dos casos, forma uma locução verbal, em todas as outras ocorrências ele também deverá formar locução verbal.

191	15	a	o	DET	<artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art	16	
192	16	peseta	peseta	NOUN	<np-def> N F S @SUBJ>	Gender=Fem Number=Sing	17	nsubj
-	17	poderia	poder	VERB	<cjt> V COND 3S @FS-PAUX	Mood=Cnd Number=Sing Person=3 VerbForm=Fin	12	
-	18	tomar	tomar	AUX	<aux> <cjt> V INF @ICL-AUX<	VerbForm=Inf	29	aux
193	+ 17	poderia	poder	AUX	<cjt> V COND 3S @FS-PAUX	Mood=Cnd Number=Sing Person=3 VerbForm=Fin	18	
194	+ 18	tomar	tomar	VERB	<aux> <cjt> V INF @ICL-AUX<	VerbForm=Inf	10	ccomp
195	19	o	o	DET	<artd> ART M S @>N	Definite=Def Gender=Masc Number=Sing PronType=Art	20	
196	20	caminho	caminho	NOUN	<np-def> N M S @ACC	Gender=Masc Number=Sing	29	obj
197	21-22	da	de	ADP	<first-cjt> <sam-> PRP @N<		23	case
198	21	de	de	ADP	<first-cjt> <sam-> PRP @N<		23	case
199	22	a	o	DET	<-sam> <artd> ART F S @>N	Definite=Def Gender=Fem Number=Sing PronType=Art		
200	23	libra	libra	NOUN	<first-cjt> <np-def> N F S @P<	Gender=Fem Number=Sing	20	nmod

Preposições entre verbos #250

Correções de preposições #258

Tipo: 1. opção nossa de anotação e 2. correção de conversão do PALAVRAS

Total de tokens modificados no corpus: 2208

As correções de preposições foram sobretudo em lote e visando sistematizar, segundo guidelines UD e algumas poucas decisões nossa, a disposição das preposições no Bosque-UD.

1. Foi opção nossa de anotação nomear todas as preposições entre locuções verbais como elementos sintáticos auxiliares, parte da mesma locução verbal (ADP_aux). Esse tipo de decisão alterou 664 tokens no corpus:

20	14	grandes	grande	ADJ	ADJ M P @>N	Gender=Masc Number=Plur	15	amod	_	_
21	15	bancos	banco	NOUN	<np-def> N M P @SUBJ>	Gender=Masc Number=Plur	19	nsubj	_	_
22	16	passaram	passar	AUX	<cjt> <aux> V PS/MQP 3P IND @FS-STA	Mood=Ind Number=Plur Person=3 VerbForm=F				
	- 17	a	a	ADP	PRP @PRT-AUX<		19	obl	_	_
	+ 17	a	a	ADP	PRP @PRT-AUX<		19	aux	_	_
24	18	não	não	ADV		Polarity=Neg	19	advmod	_	_
25	19	fornecer	fornecer	VERB	<mv> V INF @ICL-AUX<	VerbForm=Inf	3	conj	_	_
26	20	recursos	recurso	NOUN	<np-idf> N M P @<ACC	Gender=Masc Number=Plur	19	obj	_	_

2. Por conta do script de validação, precisamos posteriormente chamar essas preposições de ADP_case.

3. Por outro lado, foi fruto da correção da conversão do PALAVRAS que alteramos os outros 1544 tokens, garantindo sistematicidade nas demais preposições do corpus. De forma geral, as preposições, caso estejam introduzindo verbos, são “mark” (exceto, como visto acima, elas sejam parte de locução verbal). Nos demais casos, preposições serão sempre “case”.

Documento descrevendo o caso das preposições:
https://drive.google.com/open?id=1jhYXKOP6so-XGKK8kqQkIR0O3my0K9bfP-DMFX_s mBk

3. Posteriormente, e isso não entrou na versão 2.4, decidimos chamar as preposições no meio de locuções verbais de PART_case, para passar no script de validação UD.

Correção de MWEs #257

Tipo: correção de conversão do PALAVRAS e revisão do corpus

Total de tokens modificados no corpus: 2516

Fizemos esforços para desmembrar as MWEs que estavam como “flat:name” e como “compound”, ao mesmo tempo que juntamos como “fixed” outras tantas MWEs cuja análise seria dificultada sem essa relação. Além disso, garantimos que todos os “fixed”, “compound” e “flat” tinham o campo MWE no misc, razão do número elevado de mudanças, e verificamos também por outros erros de conversão do PALAVRAS, tais como preposições e conjunções apontando para elementos anteriores e conjunções coordenativas sem um pai que seja “conj”:

Foram MWEs sistematizadas como fixed:

1. é como se/foi como se
2. sendo assim
3. sendo que
4. do que

1	«	«	PUNCT	PU @PU	_	2	punct	_	ChangedBy=Issue165 SpaceAfter=No
2	Que	Que	PROPN	PROP M S @SUBJ>		13	nsubj	_	MwE=Que_futuro_para_o_livro_em_a_E
3	futuro	futuro	NOUN	PROP @N<		2	flat:name	_	
- 4	para	para	ADP	PRP @N<		2	flat:name	_	
- 5	o	o	DET	PROP @P<		2	flat:name	_	
+ 4	para	para	ADP	PRP @N<		5	case	_	
+ 5	o	o	DET	PROP @P<		2	nmod	_	
6	livro	livro	NOUN	PROP @N<		2	flat:name	_	
7-8	na	_	_	_		_	_	_	
- 7	em	em	ADP	PRP @N<		2	flat:name	_	
- 8	a	a	DET	PROP @P<		2	flat:name	_	
+ 7	em	em	ADP	PRP @N<		8	case	_	
+ 8	a	a	DET	PROP @P<		5	nmod	_	
9	Europa	Europa	PROPN	PROP @N<		2	flat:name	_	ChangedBy=Issue165 SpaceAf
10	»	»	PUNCT	PU @PU		2	punct	_	
72	7	não	não	ADV	_	8	advmod	_	
73	8	refletem	refletir	VERB	<mv> V PR 3P IND @FS-N<PRED		Mood=Ind Number=Plur Person=3 Tense=Pres		
74	9	apenas	apenas	ADV	ADV @>N	10	advmod	_	
75	10	conjeturas	conjetura	NOUN	<first-cjt> <np-idf> N F P @<ACC		Gender=Fem Number=Plur	8	o
76	11	teóricas	teórico	ADJ	ADJ F P @N<		Gender=Fem Number=Plur	10	amod
77	12	,	,	PUNCT	PU @PU	8	punct	_	
- 13	mas	mas	CCONJ	<co-vfin> KC @CO		8	cc	_	
78	+ 13	mas	mas	CCONJ	<co-vfin> KC @CO		15	cc	_
79	14	a	o	DET	<artd> ART F S @>N		Definite=Def Gender=Fem Number=Sing PronType=Art	15	d
- 15	experiência	experiência	NOUN	<np-def> N F S @FS-N<PRED		8	Gender=Fem Number=Sing	8	dep
80	+ 15	experiência	experiência	NOUN	<np-def> N F S @FS-N<PRED		Gender=Fem Number=Sing	8	conj
81	16	duramente	duramente	ADV	ADV @ADVL>		17	advmod	_
82	17	vivida	viver	VERB	<mv> V PCP F S @ICL-N<		Gender=Fem Number=Sing VerbForm=Part Voice=Pass	15	acl

Correções obl matriz de confusão #253

Tipo: revisão manual do corpus

Total de tokens modificados no corpus: 391

Analisando os erros da matriz de confusão entre obl e nmod, conseguimos revisar, manualmente, 391 tokens que estavam anotados erroneamente, buscando, por exemplo, “nmod” que fossem filhos de verbos (erro!):

19	14	casado	casado	ADJ	<mv> V PCP M S @ICL-N<PRED		Gender=Masc Number=Sing	2	acl	_	ChangedBy
20	15	com	com	ADP	PRP @<PIV		17	case	_		
21	16	um	um	NUM	<card> NUM M S @>N		NumType=Card	17	nummod	_	ChangedBy=Issue168
- 17	filho	filho	NOUN	<np-def> N M S @P<		14	Gender=Masc Number=Sing	14	obj	_	ChangedBy=Issue16
22	+ 17	filho	filho	NOUN	<np-def> N M S @P<		Gender=Masc Number=Sing	14	obl	_	ChangedBy=Issue16

Correções xcomp matriz de confusão #259

Tipo: revisão manual do corpus

Total de tokens modificados no corpus: 164

Analisando os erros da matriz de confusão entre “xcomp” e outras categorias sintáticas, conseguimos revisar, manualmente, 164 tokens que estavam anotados erroneamente na partição “teste” do corpus:

104	26	alianças	aliança	NOUN	<np-idf> N F P @<OC	Gender=Fem Number=Plur	25	obj	_	_
105	27	capazes capaz	ADJ	ADJ F P @N<		Gender=Fem Number=Plur	26	amod	_	_
106	28	de de	ADP	PRP @A<ARG			29	mark	_	_
-	29	sustentar	sustentar	VERB	<mv> V INF @ICL-P<	VerbForm=Inf	27	advcl	_	_
+	29	sustentar	sustentar	VERB	<mv> V INF @ICL-P<	VerbForm=Inf	27	xcomp	_	_
108	30	Lula Lula	PROPN	PROP M S @<ACC		Gender=Masc Number=Sing	29	obj	_	_
109	31	quando quando	ADV	<rel> ADV @ADVL>			32	advmod	_	_
110	32	virasse virar	VERB	<mv> V IMPF 3S SUBJ @FS-<ADVL		Mood=Sub Number=Sing Person=3 Tense=Imp VerbForm=Fin				

Atenção especial para casos limítrofes em que o verbo pode ser considerado tanto “xcomp” como “advcl”, dependendo do ponto de vista empregado. Nesses casos, optamos por utilizar “advcl”, uma vez que a semântica das orações adverbiais, nesses casos, é mais generalizável.

32	15	com com	ADP	PRP @N<PRED			17	case	_	_
33	16	cinco cinco	NUM	<card> NUM F P @>N		NumType=Card	17	nummod	_	ChangedBy=Issue168
34	17	telas tela	NOUN	<np-def> N F P @P<		Gender=Fem Number=Plur	13	nmod	_	_
35	18	que que	PRON	<rel> INDP F P @SUBJ>		Gender=Fem Number=Plur PronType=Rel	19	nsubj	_	_
36	19	servem servir	VERB	<mv> V PR 3P IND @FS-N<		Mood=Ind Number=Plur Person=3 Tense=Pres VerbForm=Fin	17			
37	20	para para	ADP	PRP @<PIV			21	mark	_	_
-	21	ilustrar	ilustrar	VERB	<mv> V INF @ICL-P<	VerbForm=Inf	19	xcomp	_	_
+	21	ilustrar	ilustrar	VERB	<mv> V INF @ICL-P<	VerbForm=Inf	19	advcl	_	_
39	22	sua seu	DET	<poss> <si> DET F S @>N		Gender=Fem Number=Sing PronType=Prs	24	det	_	_
40	23	« «	PUNCT	PU @PU			24	punct	_	ChangedBy=Issue165 SpaceAfter=No
41	24	evolução	evolução	NOUN	<np-idf> N F S @<ACC	Gender=Fem Number=Sing	21	obj	_	_
42	25	» »	PUNCT	PU @PU			24	punct	_	_

Correções após validador UD #254

Tipo: revisão manual do corpus

Total de tokens modificados no corpus: 518, inicialmente.

Posteriormente, próximo do lançamento, mais 811 tokens foram alterados.

São correções efetuadas a partir dos erros apontados pelo script oficial de validação do projeto UD de 2018. As frases foram buscadas a partir de regras e filtros complexos e modificadas manualmente. Atenção especial para a pesquisa por adjuntos adnominais filhos de verbo: são confusões comuns nas matrizes de confusão, pois vai depender de para onde o adjunto adnominal estaria apontando – para o complemento do verbo (um nome, portanto deve ser nmod), ou para o verbo (portanto deve ser obl).

134	13-14	elevou-se	_	_	_	_	_	_	_	_
135	13	elevou elevar	VERB	<mv> <hyphen> <se-passive> V PS 3S IND @FS-STA		Mood=Ind Number=Sing Person=3 Tense=Past				
136	14	se se	PRON	PERS M 3S ACC @<ACC-PASS		Case=Acc Gender=Masc Number=Sing Person=3 PronType=Prs				
137	15	a a	ADP	PRP @<SA			17	case	_	_
138	16	298 298	NUM	<card> NUM M P @>N		NumType=Card	19	nummod	_	ChangedBy=Issue168
139	17	milhões milhão	NUM	<np-def> N M P @P<		NumType=Card	16	flat	_	_
140	18	de de	ADP	PRP @N<ARG			19	case	_	_
-	19	dólares dólar	NOUN	<np-idf> N M P @P<		Gender=Masc Number=Plur	13	nmod	_	_
+	19	dólares dólar	NOUN	<np-idf> N M P @P<		Gender=Masc Number=Plur	13	obj	_	_

6. Considerações finais e próximos passos

O presente relatório detalhou as atividades desenvolvidas ao longo de um ano, no projeto Construção de datasets para o PLN de língua portuguesa. Assim que o processo de revisão terminar, os próximos passos são: (i) tornar pública toda a documentação linguística do corpus; (ii) estudar outras abordagens para detecção de inconsistências em treebanks, como [8] [9] [10] [11] [12] [13]; (iii) testar outras abordagens para detecção de inconsistências em treebanks.

Referências

- 1 - Trugo, Luiza Frizzo; de Freitas, M. C. (2016). Classes de palavras—da Grécia Antiga ao Google: Um estudo motivado pela conversão de tagsets (Doctoral dissertation, PUC-Rio).
- 2 - Rocha, L., Soares-Bastos, I., Freitas, C., & Rademaker, A. Scavenger hunt: what do we find when look for confusions.
- 3 - https://github.com/UniversalDependencies/UD_Portuguese-Bosque
- 4 - McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., ... & Bedini, C. (2013). Universal dependency annotation for multilingual parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (Vol. 2, pp. 92-97).
- 5 - MANNING, Christopher D. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In: International conference on intelligent text processing and computational linguistics. Springer, Berlin, Heidelberg, 2011. p. 171-189.
- 6 - <https://github.com/alvelvis/Interrogat-rio>
- 7 - Cunha, Celso, and Lindley Cintra. "Nova gramática do português contemporâneo: terceira edição revista." Nova apresentação 3 (2001).
- 8 - Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, Maria Simi, and Giulia Venturi. 2018. Assessing the impact of incremental error detection and correction. a case study on the Italian universal dependency treebank. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 1–7, Brussels, Belgium, November. Association for Computational Linguistics.
- 9 - Adriane Boyd, Markus Dickinson, and W. Detmar Meurers. 2008. On detecting errors in dependency treebanks. Research on Language and Computation, 6(2):113–137, Oct.
- 10 - Marie-Catherine de Marneffe, Matias Gioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the universal dependencies corpora. In Proceedings

of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 108–115, Pisa, Italy, September. Linköping University Electronic Press.

11 - Markus Dickinson and W. Detmar Meurers. 2003. Detecting inconsistencies in treebanks. In Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT-03), pages 45–56, Växjö, Sweden.

12 - Markus Dickinson. 2015. Detection of annotation errors in corpora. *Language and Linguistics Compass*, 9(3):119–138.

13 - Alexander Volokh, Günter Neumann (2011): Automatic Detection and Correction of Errors in Dependency Treebanks. *ACL (Short Papers) 2011*: 346-350