Explorando a revisão de corpora por meio da comparação de regras gramaticais em padrões sintáticos

Wellington José Leite da Silva[1], **Alexandre Rademaker**[1,3], **Leonel Figueiredo de Alencar**[1,2]

[1]Escola de Matemática Aplicada da FGV (EMAP), Brazil

[2]Universidade Federal do Ceará (UFC), Brazil

[3]IBM Research, Brazil

Abstract. *Language resources, such as corpora, are fundamental for the development of text processing tools. A resource currently considered fundamental for NLP in Portuguese is the corpus UD Bosque, part of the corpora collection in the Universal Dependencies (UD) project. Despite UD Bosque being originated from a manually revised (golden) corpus, several annotation consistency problems are encountered in its current version. In this work, we present the methodology to correct the problems of morphological annotations in the corpus; in particular, we correct morphological agreements of adjectives, determinants, and nouns. We discuss the errors, exceptions, or non-trivial cases, corrections that we made, and the impact of changes on the corpus on the training of statistical parsers.*

Resumo. *Recursos linguísticos, como corpora, são fundamentais para o desenvolvimento de ferramentas para processamento de textos. No processamento de textos em português, um recurso atualmente considerado fundamental é o corpus UD Bosque, parte da coleção de corpora no projeto 'Universal Dependencies' (UD). A despeito do corpus UD Bosque ter sido convertido para as anotações de UD de um corpus originalmente revisado, ainda são vários os problemas de consistência das anotações encontrados na atual versão do corpus. Neste trabalho, apresentamos a metodologia usada para corrigir os problemas de anotações morfológicas nos corpus UD Bosque, em particular, identificamos erros nas anotações morfológicas de determinantes e adjetivos que deveriam concordar com os substantivos que modificam. Discutimos como os erros foram identificados, as exceções ou casos não triviais, as correções realizadas e o impacto das mudanças no corpus no treinamento de analisadores sintáticos estatísticos.*

1. Introdução

O aprendizado de máquina (ML) [Mitchell et al. 1997] evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial e hoje é largamente utilizado no processamento de linguagem natural (NLP). Em sua essência, tais métodos demandam dados, manualmente anotados ou não, para o treinamento de sistemas na realização de tarefas específicas.

Na linguística computacional [1] ou no processamento de linguagem natural, um analisador sintático é o primeiro componente de muitos sistemas que pretendem processar e interpretar texto. Dada uma frase como entrada, o analisador sintático marca cada palavra com uma classe gramatical (POS) e determina as relações sintáticas entre as palavras na frase. Na análise de dependências, estas relações são representadas na árvore de análise de dependência e estão diretamente relacionadas ao significado subjacente da frase em questão [Jurafsky and Martin 2009]. Um dos principais problemas que torna a análise sintática uma tarefa desafiadora é que as linguagens humanas apresentam níveis notáveis de ambiguidade. Não é incomum que frases de comprimento moderado, 20 ou 30 palavras, tenham centenas, milhares ou mesmo dezenas de milhares de possíveis estruturas sintáticas [Oepen et al. 2004]. Um analisador sintático de linguagem natural deve de alguma forma pesquisar todas essas alternativas e encontrar a estrutura mais plausível de acordo com o contexto. Os analisadores sintáticos podem utilizar uma gramática computacional implementada em formalismos como HPSG [Sag et al. 2003] ou serem baseados no aprendizado de máquina, usando grande volumes de textos (corpora), segmentados em sentenças que são associadas a sua respectiva análise sintática.

O 'Universal Dependencies' (UD) [2] é um projeto para o desenvolvimento de um esquema de anotação multilinguagem, com o objetivo de facilitar o desenvolvimento de analisador multilíngue, aprendizagem multilíngue e pesquisa na área de análise sintática com a perspectiva da tipologia linguística. O desenvolvimento de analisadores sintáticos era, antes de UD, em grande parte limitado pela falta de um grande volume de dados anotados de forma consistente seguindo um mesmo esquema de anotações (marcações e diretrizes). Isto é, vários corpora eram criados por diferentes grupos de pesquisa usando diferentes esquemas de anotação. Atualmente o projeto UD conta com três corpora em português. O corpus mais consistente e mais utilizado no treinamento dos analisadores sintáticos mais populares e livremente distribuídos é o UD Bosque [Rademaker et al. 2017].

Infelizmente, embora seja o mais importante corpus do português no projeto UD, o corpus UD Bosque tem limitações. Em primeiro lugar, é um corpus considerado pequeno, com aproximadamente nove mil sentenças apenas. Em segundo lugar, embora tenha sido convertido para o formato UD a partir de um corpus previamente elaborado em outro formalismo e manualmente revisado [Afonso et al. 2002], ainda apresenta várias inconsistências e erros de anotações, eventualmente introduzidas inadvertidamente durante suas revisões ou na conversão para o formato UD [Rademaker et al. 2017]. Motivados em resolver tais problemas, buscamos uma metodologia que nos ajude a evitar novas inconsistências e ofereça escalabilidade para a manutenção de um novo corpus para o português que deverá contar com aproximadamente 320 mil sentenças (35 vezes maior que o UD Bosque) [Ribeiro et al. 2020]. Este trabalho começou com a exploração da validação cruzada do corpus com o léxico de formas plenas do português MorphoBr [de Alencar et al. 2018]. A partir desta etapa, documentada em outro artigo de autoria dos dois últimos autores do presente trabalho, artigo esse submetido a um periódico e ainda em revisão, identificamos casos de anotações morfológicas inconsistentes de acordo com as regras da gramática do português.

[1] O termo linguística computacional é usado aqui para designar a área de pesquisa que utiliza sistemas computacionais para estudos linguísticos e se contrasta com o processamento de linguagem natural, mais aplicado, a área interessada no desenvolvimento de sistemas para processamento automático de textos.

[2] `https://universaldependencies.org`

Neste artigo, documentamos a metodologia que adotamos para correção dos erros das anotações de adjetivos e determinantes que, segundo a gramática do português, devem concordar em gênero e número com os substantivos que introduzem ou modificam. Diferentemente de uma gramática computacional, onde a regra de concordância entre adjetivos, determinantes e substantivos é explicitamente codificada, permitindo assim que um analisador sintático que utilize tal gramática seja preciso na análise de uma sentença como gramatical ou agramatical, um sistema baseado no aprendizado de máquina deve aprender as regras de concordância do português a partir dos dados. Se os dados apresentam anotações inconsistentes, espera-se que o analisador sintático não será capaz de aprender corretamente a regra de concordância.

Este trabalho está organizado da seguinte forma. Na seção 2 apresentamos o projeto UD e o corpus Bosque. Na seção 3 discutimos os erros encontrados no corpus diretamente relacionados à concordância de marcações morfológicas, indiretamente relacionados ao desvio de concordância entre palavras ou relativos a erros de concordância nos textos originais. Na seção 4, avaliamos o impacto das mudanças descritas neste trabalho no treinamento de dois analisadores sintáticos. Finalmente, apresentamos nossas conclusões na seção 5.

2. Universal Dependencies e o corpus UD Bosque

As 'Universal Dependencies' (UD) [de Marneffe et al. 2021a] são um esquema de anotação morfossintática usado para criar corpora para mais de 100 idiomas. O UD especifica um conjunto de etiquetas (tagset) e diretrizes de uso destas etiquetas para a codificação de análises sintáticas de sentenças. As relações gramaticais entre palavras são usadas para explicar como as estruturas de argumento-predicado são codificadas morfossintaticamente em diferentes idiomas, enquanto as características morfológicas e as classes gramaticais fornecem as propriedades das palavras. [de Marneffe et al. 2021b] sustenta que esta teoria é uma boa base para a anotação consistente de línguas tipologicamente diversas de uma forma a apoiar a implementação computacional da linguagem natural, bem como estudos linguísticos mais amplos. O projeto é um esforço de uma comunidade aberta com mais de 300 contribuidores, que produziram quase 200 treebanks em mais de 100 idiomas.

O UD teve sua primeira versão de corpus do português em 2015 e hoje conta com 3 corpora em português (UD Bosque, GSD e PUD), sendo o Bosque [Rademaker et al. 2017] o mais usado para treinamento de analisadores sintáticos, com cerca de 210 mil tokens e 9,3 mil sentenças. Tudo que será apresentado neste trabalho foi realizado no corpus UD Bosque, mas será, em trabalhos futuros, aplicado nos demais corpora do português.

As anotações do UD são armazenadas em arquivos CoNLL-U [3] em que sentenças são codificadas com um token por linha e cada linha com 10 colunas, a saber, (i) o índice do token na sentença (ID), (ii) a forma original da palavra na sentença ou pontuação (form), (iii) o lema ou radical da palavra (lemma), (iv) a classe gramatical (UPOS), (v) a classe gramatical específica do idioma (XPOS), (vi) a lista de características morfológicas como gênero, número e flexões verbais (feats), (vii) o id do token governante na estrutura

[3] http://universaldependencies.org/format.html

de dependências (HEAD), (viii) a relação de dependência com o token governante (DEPREL), (ix) o gráfico de dependências expandido (DEPS) e (x) outras anotações (MISC). Na Figura 1 temos um exemplo da estrutura.

Figura 1. exemplo de anotações morfossintáticas de uma sentença do corpus UD Bosque no formato CoNLL-U

```
# text = Maradona negou veementemente as críticas da mãe de Franco.
# sent_id = CF388-2
1   Maradona        Maradona        PROPN   _   Gender=Masc|Number=Sing                                  2   nsubj   _   _
2   negou           negar           VERB    _   Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin   0   root    _   _
3   veementemente   veementemente   ADV     _   _                                                       2   advmod  _   _
4   as              o               DET     _   Definite=Def|Gender=Fem|Number=Plur|PronType=Art        5   det     _   _
5   críticas        crítica         NOUN    _   Gender=Fem|Number=Plur                                   2   obj     _   _
6-7 da              _               _       _   _                                                       _   _       _   _
6   de              de              ADP     _   _                                                       8   case    _   _
7   a               o               DET     _   Definite=Def|Gender=Fem|Number=Sing|PronType=Art       8   det     _   _
8   mãe             mãe             NOUN    _   Gender=Fem|Number=Sing                                   5   nmod    _   _
9   de              de              ADP     _   _                                                       10  case    _   _
10  Franco          Franco          PROPN   _   Gender=Masc|Number=Sing                                  8   nmod    _   SpaceAfter=No
11  .               .               PUNCT   _   _                                                       2   punct   _   _
```

3. Erros de concordância no UD Bosque

Em trabalho anterior dos dois últimos autores deste artigo,[4] a comparação do dicionário de formas plenas Morpho-Br [de Alencar et al. 2018] com o corpus UD Bosque [Rademaker et al. 2017] foi usada para verificação de inconsistências e omissões no primeiro. Durante a comparação dos recursos, a inspeção de análises morfológicas incompletas no corpus revelou também inconsistências entre anotações no próprio corpus que ignoravam as regras de concordância do português. Dentre os casos encontrados decidimos concentrar nossa análise nos artigos e adjetivos que devem concordar com os substantivos que introduzem e modificam:

> ...quer seja definido ou indefinido, o artigo caracteriza-se por ser a palavra que introduz o substantivo indicando-lhe o gênero e número. [Cunha and Cintra 1985, página 225]

> ...o adjetivo toma a forma de singular ou plural do substantivo que ele qualifica. ...o substantivo tem sempre um gênero, o que não ocorre com o adjetivo, que assume o gênero do substantivo que ele qualifica. [Cunha and Cintra 1985, páginas 264-265]

Em UD, os artigos pertencem à classe dos determinantes, termo difundido sobretudo pela gramática gerativa de Chomsky. É uma noção distribucional, como está claro na própria definição das dependências universais.[5] Pelo critério distribucional, são determinantes em português também os pronomes demonstrativos, além dos artigos, entre outras classes de palavras. Estendemos assim nossa verificação para todos os determinantes além de apenas artigos.

Para identificar os casos de desvio de concordância, utilizamos a biblioteca Udapi [Popel et al. 2017]. A consulta 1 pesquisa por tokens anotados como adjetivos (campo `UPOSTAG` com valor `ADJ`) cujo token governante seja um substantivo (`NOUN`) e cujas etiquetas morfológicas de número e gênero ou não estejam especificadas ou sejam diferentes do seu token (substantivo) governante. Em UD, a relação de dependência `amod` é usada para ligar um adjetivo ao substantivo ou pronome que ele modifica de forma composicional ou idiomática.[6]

[4] O trabalho foi submetido para publicação e encontra-se em avaliação.

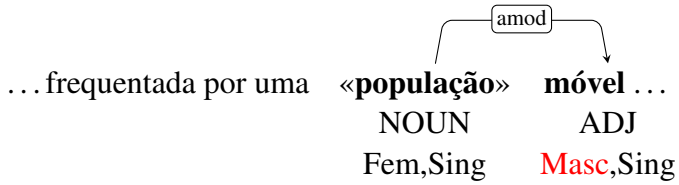
[5] `https://universaldependencies.org/u/pos/DET.html`

[6] `https://universaldependencies.org/u/dep/amod.html`

Listing 1. Exemplo de consulta no corpus usando Udapi

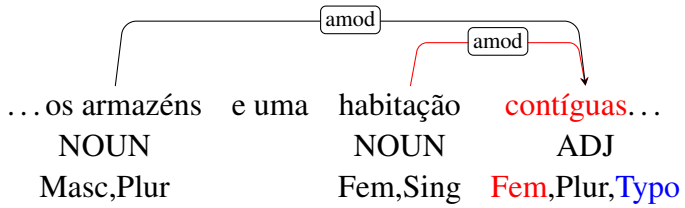
```
if ((node.feats["Gender"] != node.parent.feats["Gender"]
     or node.feats["Number"] != node.parent.feats["Number"]
     or node.feats["Gender"] == ""
     or node.feats["Number"] == "")
    and node.upos == "ADJ" and node.deprel == "amod"
    and node.parent.upos == "NOUN"):
 print(node)
```

A consulta 1, quando submetida ao corpus UD Bosque, retornou 191 casos suspeitos de falta de concordância. Os casos mais simples foram trivialmente corrigidos. No exemplo 1, o adjetivo uniforme 'móvel' [Cunha and Cintra 1985] teve a marcação de gênero corrigida de `Masc` para `Fem`.[7]



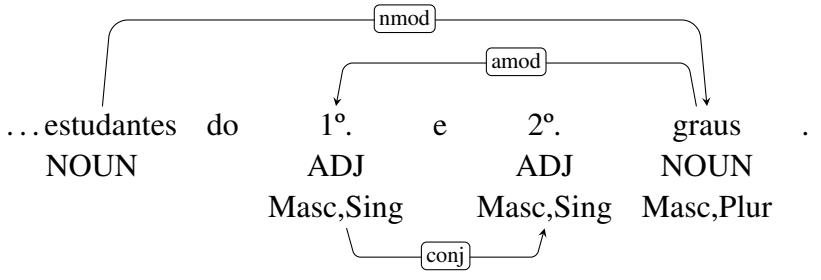
Example 1. CP52-12

Também identificamos erros gramaticais cometidos pelos autores dos textos. No exemplo 2, o adjetivo 'contíguas' deveria ter sido flexionado no masculino plural. Neste caso, UD determina que a marcação `Typo=Yes` deve ser incluída nas 'features' do token e no campo `misc`, a indicação da forma correta com `CorrectForm=contíguos`. Nota-se ainda que o erro de concordância revelou um erro na anotação sintática onde o governante do adjetivo deveria ser 'armazém', o token 'head' da coordenação.



Example 2. CP103-2

Embora a concordância de um adjetivo com uma coordenação de substantivos seja detalhadamente descrita em [Cunha and Cintra 1985], o mesmo não ocorre para construções como a apresentada no exemplo 3. Este caso consideramos como um falso positivo de nossa consulta (um falso erro) dado que entendemos a construção como gramatical.



Example 3. CF66-4

[7]Nos exemplos, marcações em vermelho sinalizam os erros encontrados cuja correção é descrita no texto, marcações introduzidas estão em azul. Os números `CFXX-XX` ou `CPXX-XX` são identificadores das sentenças do corpus.

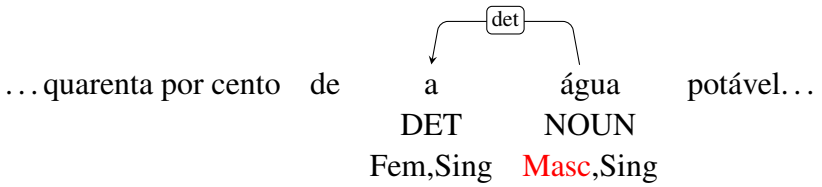
Nota-se que a consulta 1, embora produtiva na identificação dos erros mais frequentes de anotação, listou falsos erros, os quais tiveram que ser analisados manualmente. Os principais falsos erros foram casos de coordenação entre substantivos modificada por um adjetivo ou como no exemplo anterior, coordenação de adjetivos modificando um substantivo. No futuro, pretendemos explorar outros padrões como as construções de adjetivos em função predicativa. Não obstante, alguns casos não triviais também foram revelados. Em "...muitos senadores **antiaborto**..." (CP786-2) temos a palavra 'antiaborto' anotada como adjetivo e sua flexão não acompanha 'senadores', o substantivo que modifica, revelando uma possível exceção à regra de concordância do português ou uma possível necessidade de anotação mais elaborada onde apenas o afixo 'anti' seria o adjetivo (compare com 'senadores contrários ao aborto') ou seria tratado como preposição (compare com 'senadores contra o aborto'). Em "...interpretadas por bandas **cover**..." (CF840-5) temos um termo estrangeiro que optamos por anotar com a morfologia adequada à concordância, embora adjetivos no inglês não flexionem.

Para tratar os desvios de concordância entre determinante e substantivos, partimos da consulta 2, uma variação simples da consulta 1. Encontramos 1226 casos de desvios de concordância entre determinantes e substantivos. Diferentemente dos casos de concordância com adjetivos, neste caso foram possíveis algumas correções em lote. Restringindo a consulta 2 para os artigos definidos do português sem anotações morfológicas de gênero ou número, encontramos um número expressivo de casos. Como os artigos definidos do português têm gênero e número regulares, fizemos um script para adicionar as anotações morfológicas em todos os artigos definidos sem anotações morfológicas de gênero e número. Com isto, o número de casos da consulta 2 diminuiu para 282.

Listing 2. Consulta no corpus de DET/NOUN usando Udapi

```
if ((node.feats["Gender"] == ""
     or node.feats["Number"] == ""
     or node.feats["Gender"] != node.parent.feats["Gender"]
     or node.feats["Number"] != node.parent.feats["Number"])
     and node.upos == "DET" and node.deprel == "det"
     and node.parent.upos == "NOUN"):
  print(node)
```

No exemplo 4, temos um caso de simples correção, onde o token 'água' estava erroneamente[8] com gênero `Masc`.

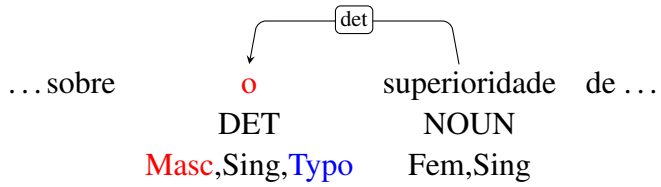


Example 4. CP452-3

Casos de erros gramaticais e ortográficos também foram encontrados, como na verificação dos adjetivos. No exemplo 5, onde autor do texto deveria ter usado o artigo definido singular feminino 'a', seguindo novamente as diretrizes de UD, utilizamos as marcações de `Typo=Yes` e `CorrectForm=a` para indicar o erro do autor do texto e

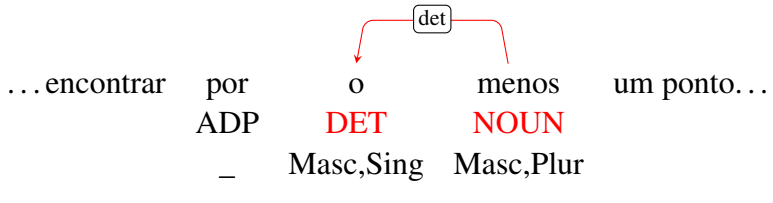
[8] E surpreendentemente para um corpus que já passou por tantas revisões.

prover uma anotação consistente para um analisador estatístico.



Example 5. CF27-7

Também encontramos casos como o exemplo 6, envolvendo a locução adverbial 'pelo menos' [9]. Neste caso, decidimos adiar o tratamento das expressões multipalavras para uma etapa futura embora nos pareça que uma análise possível seja como apresentada no exemplo 6 pelas relações de dependência.



Example 6. CF685-1

Considerando os casos aqui discutidos, no total, foram modificadas 1.094 sentenças (12% das sentenças do corpus) e 1.875 tokens (aproximadamente 1.7 tokens por sentença).

4. Avaliação

Conforme descrito na seção 1, nossa intenção é desenvolver um processo robusto e escalável para revisão de corpora. Para tal, entendemos que modificações não devem ser feitas de forma *ad hoc*. Para a avaliação do impacto de nossas alterações na qualidade do corpus, utilizamos o analisador sintático UDPipe [Straka and Straková 2017]. O UDPipe em sua versão 1.2 não é baseada em modelos de redes neurais, mas na última CoNLL 2018 Shared Task [Zeman et al. 2018] sua versão 'Future' ficou entre os três melhores colocados em todos os ranks.[10]

Os resultados obtidos são apresentados na tabela 1. Para avaliação, inicialmente, treinamos dois modelos do UDPipe com os arquivos `pt_bosque-ud-train.conllu` da versão 2.8 do corpus (de Maio 2021) e do 'branch workbench' (corrigido conforme descrito na seção 3). Usando os respectivos modelos, processamos os arquivos `pt_bosque-ud-test+dev.conllu` (juntamos os arquivos de teste e desenvolvimento que somam 1.036 sentenças) também das duas versões obtendo um arquivo `system-test+dev.conllu` para cada `pt_bosque-ud-test+dev.conllu`. Finalmente, usamos o script de avaliação desenvolvido para a CoNLL 2018 Shared Task,[11] para comparar os respectivos arquivos `pt_bosque-ud-test+dev.conllu` e `system-test+dev.conllu`. Os números da terceira coluna indicam que em todas as métricas usadas na Shared Task, o sistema melhorou seu resultado com a nova versão dos corpus.

[9] O token 'pelo' é a contração da preposição 'por' com o pronome 'o'.
[10] `https://universaldependencies.org/conll18/results.html`
[11] `https://universaldependencies.org/conll18/evaluation.html`

Existem várias métricas para quantificar a diferença entre anotações sintáticas de dependências, geralmente usadas para avaliar quão próximo é o resultado de um sistema em relação às anotações humanas ('golden') do mesmo dado. Diferentes métricas de avaliação avaliam diferentes aspectos das anotações. Na tabela 1 são apresentadas as três métricas principais usadas pelo script de avaliação da CoNLL 2018 Shared Task. As métricas LAS, MLAS e BLEX são métricas conhecidas para avaliação. A métrica 'labeled attachment score' (LAS) é uma métrica de avaliação padrão na análise de dependência: a porcentagem de tokens que são atribuídos ao token governante sintático correto e ao rótulo de dependência correto. A métrica 'Morphology-Aware Labeled Attachment Score' (MLAS) é uma extensão do CLAS (publicado experimentalmente em 2017), combinada com a avaliação de marcações de PoS tags e marcações morfológicas. A parte central é idêntica ao LAS descrito acima mas, ao contrário do LAS, certos tipos de relações não são avaliados diretamente. Palavras anexadas por meio de tais relações não são contadas como palavras independentes, sendo tratadas como características das palavras de conteúdo a que pertencem. Finalmente, a 'Bilexical dependency score' (BLEX) é semelhante ao MLAS no sentido de que se concentra nas relações entre as palavras do conteúdo. Em vez de anotações morfológicas, incorpora a lematização na avaliação. Ele está, portanto, mais próximo do conteúdo semântico e avalia dois aspectos da anotação UD que são importantes para a compreensão da linguagem: dependências e lemas.

Tabela 1. Comparação das métricas de avaliação entre as versões antiga e nova do corpus

	2.8	workbench	diferença
LAS	81.90	82.66	0.76
MLAS	67.08	67.74	0.66
BLEX	70.60	71.26	0.66

Analisadores sintáticos mais modernos têm utilizado cada vez mais pipelines baseados em redes neurais e modelos neurais pré-treinados. Um destes sistemas é o Stanza [Qi et al. 2020], a reimplementação em Python da biblioteca CoreNLP [Manning et al. 2014] de Stanford. Infelizmente, o treinamento do Stanza não terminou a tempo para que pudéssemos realizar uma comparação com a avaliação feita com o UDPipe. No entanto, na tabela 2 apresentamos a comparação entre os números parciais produzidos durante o treinamento do Stanza e os números publicados pelos autores do Stanza para os modelos pré-treinados (versão 2.5 do corpus de novembro de 2019). Obviamente esta diferença não pode ser diretamente comparada com os resultados do UDPipe, que comparam a versão atual com a versão 2.8 do corpus, que já acumula várias melhorias em relação à versão 2.5.

Tabela 2. Comparação entre os números parciais durante treinamento do Stanza e números publicados no site para o UD Bosque release 2.5

	2.5	parciais	diferença
LAS	87.57	91.11	3.54
MLAS	76.78	86.05	9.27
BLEX	80.3	87.12	7.09

5. Conclusão

Neste artigo apresentamos os primeiro passos para uma metodologia de revisão de um corpus motivada pela formalização de regras gramaticais em consultas por padrões no corpus. Em particular, revisamos o corpus UD Bosque segundo a concordância de gênero e número dos adjetivos e determinantes com os substantivos que modificam e introduzem. Conseguimos demonstrar, pela avaliação realizada, que anotações morfológicas são efetivamente usadas pelos analisadores sintáticos existentes e que houve melhora efetiva na qualidade dos dados após nossas revisões. Em trabalhos futuros, pretendemos expandir e refinar as consultas explorando talvez abordagens complementares como de [Passos 2018]. Também vale destacar métodos ligados à avaliação extrínseca dos dados, em tarefas aplicadas, como reportado em [Iwamoto et al. 2021].

Referências

Afonso, S., Bick, E., Haber, R., and Santos, D. (2002). Floresta sintá (c) tica: a treebank for portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*. ELRA.

Cunha, C. and Cintra, L. (1985). *Nova gramática do português contemporâneo*. LEXIKON Editora Digital ltda.

de Alencar, L. F., Cuconato, B., and Rademaker, A. (2018). Morphobr: An open source large-coverage full-form lexicon for morphological analysis of portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021a). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021b). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Iwamoto, R., Kanayama, H., Rademaker, A., and Ohko, T. (2021). A Universal Dependencies corpora maintenance methodology using downstream application. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 23–31, Online. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing*. Prentice-Hall, Inc., USA, 2 edition.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mitchell, T. M. et al. (1997). *Machine learning*. McGraw-hill New York.

Oepen, S., Flickinger, D., Toutanova, K., and Manning, C. D. (2004). Lingo redwoods. *Research on Language and Computation*, 2(4):575–596.

Passos, G. P. (2018). A formal specification for syntactic annotation and its usage in corpus development and maintenance: a case study in universal dependencies. Master's thesis, Universidade Federal do Rio de Janeiro.

Popel, M., Žabokrtský, Z., and Vojtek, M. (2017). Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Depen-*

dencies (UDW 2017), pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *CoRR*, abs/2003.07082.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva Universal Dependencies for Portuguese, V. (2017). Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy.

Ribeiro, L., Zulini, J. P., and Rademaker, A. (2020). The construction of a corpus from the brazilian historical-biographical dictionary. In Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., and Gonçalves, T., editors, *Computational Processing of the Portuguese Language*, pages 109–117, Cham. Springer International Publishing.

Sag, I. A., Wasow, T., and Bender, E. M. (2003). *Syntactic Theory: a formal introduction*. University of Chicago Press, Chicago, second edition edition.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.