Bastidores linguísticos e computacionais da construção de um dataset linguístico

Aluno: Elvis de Souza
Orientadora: Cláudia Freitas

1. Introdução

O presente relatório reúne as atividades realizadas entre agosto de 2019 e julho de 2020 no projeto "Construção de datasets para o PLN de língua portuguesa", por meio do qual o aluno Elvis de Souza foi bolsista de Iniciação Científica pelo CNPq (128693/2019-3). O projeto teve início em agosto de 2018, com a então bolsista Luísa Rocha, e os desdobramentos deste projeto em 2020 sem dúvida só foram possíveis por conta das realizações das etapas anteriores, relatadas no ano de 2019 [1].

As mais de 12 mil correções realizadas no corpus Bosque-UD [2] ao longo das suas versões 2.4, 2.5 e 2.6, durante 2018, 2019 e 2020, e os métodos de análise linguístico-computacional desenvolvidos para melhoria de *datasets* por meio de uma estação de trabalho desenvolvida durante o projeto [3] permitiram, além da melhoria do desempenho de um sistema baseado em aprendizado de máquina para a Língua Portuguesa, já demonstrado no relatório anterior, a construção de uma documentação minuciosa da anotação gramatical subjacente ao projeto Universal Dependencies [4] para a Língua Portuguesa. Na seção 2, explicaremos a relevância e a metodologia empregada na construção dessa documentação, traçando paralelos com questões de teoria gramatical que possam ser relevantes para a linguística.

Além disso, para que fosse possível a construção dessa documentação da anotação para língua portuguesa, precisamos tratar de algumas questões gramaticais/de linguística teórica que, embora já muito debatidas, não apresentam resultados que nos sejam conclusivos no âmbito da anotação gramatical de um corpus linguístico. Por isso, realizamos um estudo detalhado do caso das locuções verbais em Língua Portuguesa nas gramáticas e apresentamos um olhar computacional para resolver a questão, assunto do qual trataremos na seção 3 deste relatório.

Por fim, seguros de que fizemos as melhores decisões na anotação do corpus Bosque-UD e garantindo que deixamos para futuros anotadores de língua portuguesa uma extensa documentação sobre sua gramática subjacente, ensaiamos realizar a conversão de corpora no formato AC/DC [5] para o formato UD [4], e vice-versa, de tal modo que estamos possibilitando expandir, em grande quantidade e com qualidade, a infraestrutura para o PLN de língua portuguesa, objetivo deste projeto de pesquisa. Esse processo descrevemos na seção 4.

2. Diretivas e documentação de anotação UD em português (e para língua portuguesa)

O projeto Universal Dependencies ambiciona prover uma estrutura de anotação gramatical consistente e multilíngue, contando, atualmente, com corpora anotados sintaticamente (*treebanks*) de mais de 90 línguas humanas. Para que isso ocorra, eles construíram um formato de anotação padronizado, com o mesmo conjunto de etiquetas morfossintáticas, e, dentre outras especificidades, um modelo de dependências sintáticas que têm como "cabeças" palavras de conteúdo lexical, mantendo, assim, a comparabilidade mesmo entre línguas morfologicamente ricas que carecem de palavras funcionais, como o finlandês [6]. Além disso, o projeto conta com uma comunidade que colabora discutindo e modificando as diretivas de anotação gramatical, que se pretendem aplicáveis a todas as línguas do projeto.

Embora as diretivas gramaticais do projeto[1] sejam o norte principal de quem trabalha dentro do formato, questões específicas de certas línguas não podem ser tratadas sem considerar como as línguas se comportam na realidade, já que são dinâmicas e distintas. Nesse contexto, tendo sido responsáveis pela construção do principal *dataset* de língua portuguesa neste framework, o Bosque-UD, precisamos também documentar as decisões linguísticas e estruturais realizadas no corpus de modo que futuros anotadores em português possam manter-se consistentes não apenas com as diretivas mais gerais, mas também com as diretivas específicas de nossa língua.

A documentação que construímos está disponível em [7] e também é dinâmica, podendo ser atualizada de acordo com necessidades futuras. Seu objetivo é direcionar futuros anotadores no framework UD para língua portuguesa, no entanto, embora em conformidade com as diretivas gerais do projeto, optamos por realizar as explicações gramaticais partindo da nomenclatura gramatical de nossa língua, tendo em vista que assim atingiríamos um contingente maior de pessoas, falantes nativos do português, que passaram por esta nomenclatura durante a formação escolar.

Algumas questões específicas de UD não encontram correspondência em nossas gramáticas tradicionais (GT), e também pontuamos estas questões pois podem ser relevantes tanto para o aprendizado desta nova gramática quanto para um estudo linguístico-gramatical contrastivo do português.

2.1. A documentação como descrição

Nossa documentação se apresenta em um arquivo em PDF que conta, em sua versão 1.2, com 121 páginas. Ela se distribui principalmente em cinco capítulos, sendo quatro de níveis de análise gramatical (lema, classe gramatical, atributos morfológicos e dependência sintática), e um de "frases difíceis", casos mnemônicos que, embora a tradição gramatical costume ignorar, existem e precisam ser anotados como quaisquer outros, consistentemente.

O documento apresenta mais de 150 frases categorizadas morfossintaticamente, sendo em sua maioria retiradas do Bosque-UD, um corpus de textos jornalísticos em português europeu e brasileiro. Embora tenha sua base principalmente neste corpus, o documento funciona como uma documentação de qualquer outro corpus que se preste a ser anotado

[1] Disponíveis em <https://universaldependencies.org/guidelines.html>.

dentro do ambiente do projeto Universal Dependencies em língua portuguesa, portanto a necessidade de se manter a possibilidade de futura edição, afinal, cada nova frase pode nos confrontar com um novo desafio impensado, principalmente se de gêneros e épocas distintos.

De forma semelhante, em [8], Sampson produziu um livro de mais de 500 páginas que resultou da anotação de um *treebank*, o SUSANNE, de inglês moderno. Em seu preâmbulo, o autor deixa claro que seus objetivos são dois: apresentar uma forma de representar sentenças em sua língua que são formalizáveis – lembrando-se de que nem todas são facilmente categorizáveis –, e apresentar um novo corpus para ser processado por computadores, o SUSANNE, do qual derivaram suas categorias.

Em [9], os autores fazem algo ainda mais específico: descrevem como foi realizada a análise de papéis semânticos em seu novo corpus, o PropBank, ao longo de 66 páginas. A anotação de papéis semânticos não é comumente tratada nas gramáticas tradicionais, mas também apresenta dados relevantes para a compreensão da estrutura de uma língua, embora neste documento em especial os autores não tenham o objetivo de debater a língua em si, mas seu corpus.

Este panorama é importante pois é necessário entender que, assim como proposto por Sampson, a língua corresponde a "pessoas conversando e escrevendo" [10]. Sistematizar o que enunciamos, seja em gramática ou em uma documentação como essa que apresentamos, é uma tentativa que, embora pareça funcionar para um conjunto limitado de sentenças, usualmente se restringe a este conjunto apenas, portanto durante a tarefa de anotação estamos, também, descrevendo apenas uma parte muito reduzida de nossa língua.

Se, por um lado, ao descrever um corpus ou formato de anotação não estamos representando nossa língua em sua plenitude, pior seria se tentássemos fazê-lo a partir de fragmentos de língua que nunca foram enunciados por pessoas, escrevendo ou conversando. Nesse sentido, quanto mais perspectivas, frases e modelos de anotação colocarmos em contato, tanto melhor será nossa compreensão sobre as formas de se analisar uma língua. Para isso, na próxima seção colocamos luz sobre categorias gramaticais que se fazem presentes em nossa documentação e que são diferentes ou não são discutidas comumente em gramáticas do português na esperança de alimentar a quantidade de perspectivas pelas quais se pode observar a língua portuguesa.

2.2. Algumas questões para a teoria gramatical

Em primeiro lugar, como decorrência direta do fato de que a descrição de um corpus ou a documentação de um modelo de anotação gramatical se realiza somente em exemplos de língua em uso, precisamos deixar claro neste documento que a classificação de uma palavra é feita em contexto. Nesse sentido, uma mesma palavra pode, por exemplo, ser categorizada como advérbio ou adjetivo de acordo com o sintagma em que se insere. É o caso da figura 1, em que "para", comumente preposição, neste caso inicia uma oração subordinada, portanto se comporta como conjunção subordinativa:

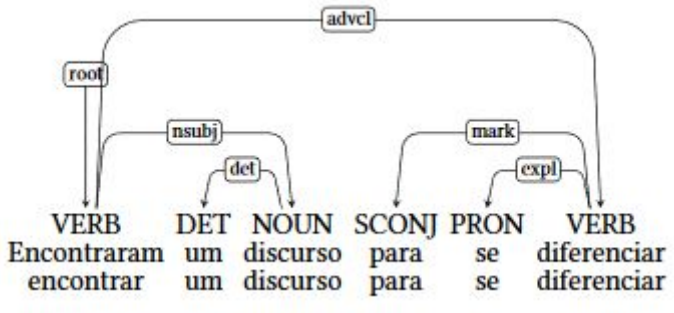

Figura 1: Encontraram um discurso *para* se diferenciar

Ainda em relação às preposições e todas as palavras funcionais, é interessante notar que, na contramão da tradição de descrição gramatical em português e até mesmo de certos modelos de anotação de corpus, as palavras funcionais (incluindo os verbos auxiliares e copulativos), em UD, nunca são cabeças de sintagma, como na figura 2. Nela, "é" e "de" são dependentes sintaticamente de "US$", que é raíz da sentença:

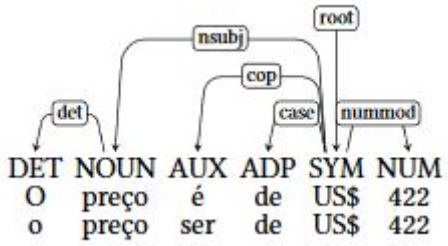

Figura 2: O preço *é de* US$ 422

Na figura acima, podemos verificar, ainda, que "422" é um número que modifica o símbolo "US$". Para UD, o adjunto adnominal receberá uma classificação sintática diferente a depender de sua classe gramatical: se, como na figura acima, for numeral, será "nummod" (modificador numeral); se, como na figura 3, for um adjunto adnominal que seja adjetivo, será "amod" (modificador adjetivo), ou, ainda, "nmod" (modificador substantivo) caso seja um substantivo, como na figura 4.

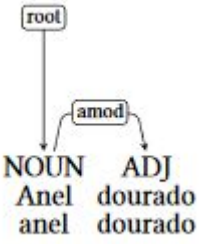

Figura 3: Anel dourado

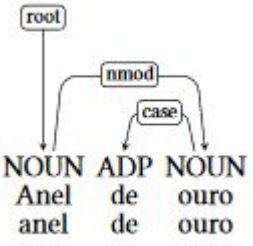

Figura 4: Anel de ouro

Outro dado interessante é que o modelo de anotação não diferencia adjunto adnominal (substantivo) de complemento nominal (de substantivo). Veja que, na figura 4 acima, "ouro" é

"nmod", assim como na figura 5 "represálias" também o é, pois ambos são dependentes de substantivo. No entanto, o modelo diferencia caso o complemento nominal seja de um adjetivo (figura 6), em que "revisão" é um "obl", ou complemento oblíquo.

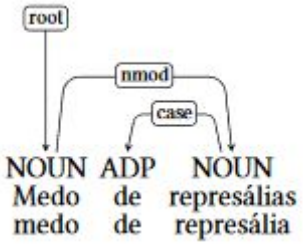

Figura 5: Medo de represálias

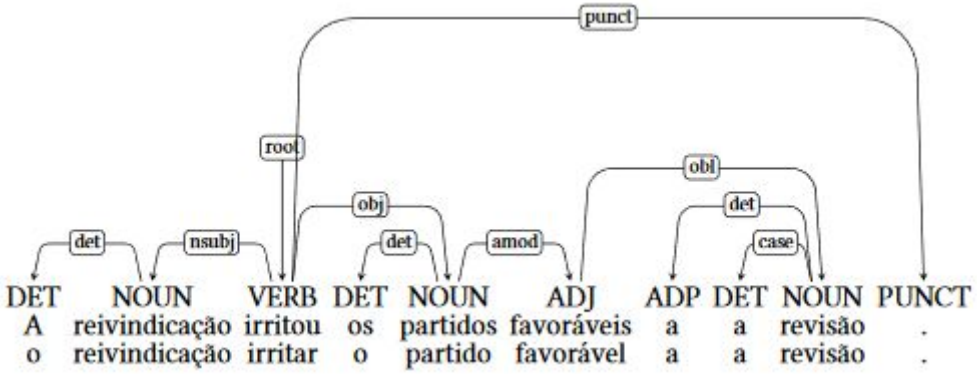

Figura 6: A reivindicação irritou os partidos favoráveis à revisão.

A classe "obl", por sua vez, é a tipicamente utilizada para os adjuntos adverbiais, que podem ser de dois tipos: "obl", caso a classe gramatical do adjunto adverbial não seja de advérbio ("mais" em figura 7), ou "advmod", caso a classe gramatical do adjunto adverbial seja a de advérbio ("entrevista" em figura 8):

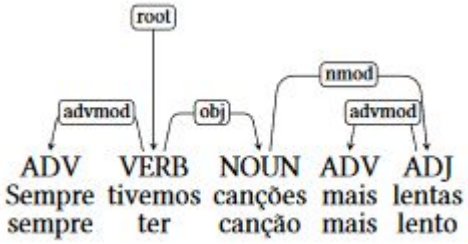

Figura 7: Sempre tivemos canções *mais* lentas

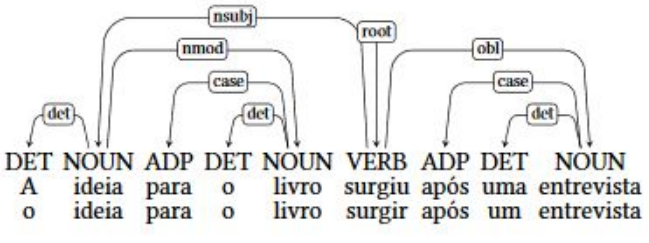

Figura 8: A ideia para o livro surgiu após uma *entrevista*

E, por fim, um dado muito discrepante da análise gramatical mais comum é o que se refere ao predicativo do objeto. Para o modelo UD, existe uma diferença crucial caso este

predicativo seja argumento do verbo, como na figura 9, em que "culpado" depende de "declarou", ou do nome, como na figura 10, em que "nua" depende de "modelo":

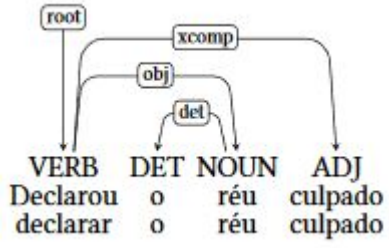

Figura 9: Declarou o réu culpado

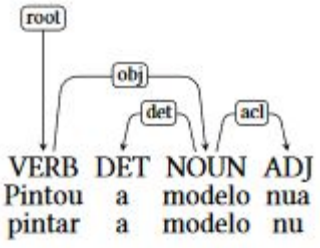

Figura 10: Pintou a modelo nua

3. Gramáticas em discussão: o caso das locuções verbais

Ao longo do processo de revisão de anotação linguística, um dos desafios com que nos deparamos foi o de anotar construções do tipo [estar + a + $V_{infinitivo}$]. Além da tarefa de decidir a classe gramatical e a função sintática dos dois verbos, precisamos realizar a anotação também da partícula "a" entre a forma "estar" e o verbo no infinitivo que a segue. Comparativamente, precisamos decidir se tais ocorrências devem receber um tratamento igual ao de casos como os das assim nomeadas locuções verbais aspectuais – [acabar + de + $V_{infinitivo}$] – e as locuções verbais modais – [querer + $V_{infinitivo}$]. No corpus Bosque-UD, em sua versão 2.4, as primeiras são tratadas como um caso de locução (*aux*) e, as segundas, como um caso de subordinação entre orações, em que a segunda é uma oração substantiva objetiva direta reduzida de infinitivo (*xcomp*) da segunda (figuras 11 e 12, respectivamente).

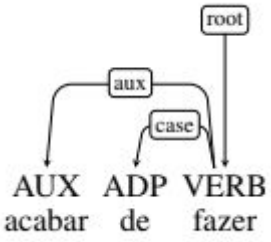

Figura 11. A anotação de "acabar de fazer" no Bosque-UD 2.4

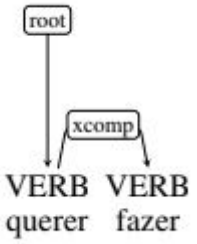

Figura 12. A anotação de "querer fazer" no Bosque-UD 2.4

Na gramática UD, a anotação sintática das partículas "de", "a" e "para" entre dois verbos pode variar drasticamente caso consideremos que 1) elas fazem parte de uma locução verbal, indicando que os dois verbos estariam sendo unidos pela partícula e as três palavras formam uma unidade, ou 2) elas iniciam uma oração, sendo que o segundo verbo é, via de regra, complemento do primeiro verbo, que não é auxiliar. No caso das locuções verbais de tempo composto – [ter/haver + $V_{particípio}$] – há consenso nas diretivas do projeto (e nas gramáticas) de que se trata de uma unidade verbal e o verbo ter/haver deve ser anotado como de função sintática "auxiliar", sendo dependente sintaticamente do segundo verbo. Por outro lado, se construções como "gostar de cantar" devem ser consideradas uma locução ou uma

subordinação entre orações não é uma resposta simples. Especialmente, é um ponto de discussão se a palavra "de" deve ser anotada como parte de uma locução verbal modal ou como conjunção subordinativa (iniciando uma oração reduzida de infinitivo que complementa o primeiro verbo).

A fim de sustentar nossa decisão, lançamos mão dos dicionários e gramáticas específicos do português de tal modo que o corpus seja anotado de maneira linguisticamente embasada e mantendo-se próxima a abordagens das chamadas gramáticas tradicionais. Buscamos em [11], [12] e [13] o que se postulou sobre o assunto, dados que sistematizaremos na seção 3.1, apontando também algumas lacunas e divergências entre os autores. Por fim, na seção 3.2, pensaremos em algumas soluções para nossos questionamentos a partir do que observamos nas gramáticas e tentaremos encarar a anotação de algumas sentenças complexas de nosso corpus com base nelas.

3.1. Revisão da literatura

As construções [estar + a + $V_{infinitivo}$] estão presentes apenas na parte portuguesa do corpus Bosque-UD, totalizando 80 sentenças, como em (1) e (2). Há também uma sentença do tipo [estar + para + $V_{infinitivo}$], na parte brasileira (3).

(1) CP3-4 – "**Estamos a dotar** os computadores de um novo sentido" disse Steve d'Averio, director de marketing para a Europa da Logitech.

(2) CP285-4 – António Guterres foi o primeiro convidado de uma série de debates com líderes políticos que o Inesc **está a promover**.

(3) CF835-12 – Cursando economia na Faap, Kiko espera ansioso o seu telefone celular, que **está para sair**.

Para dar direcionamento às questões, precisamos procurar referência em diferentes seções das gramáticas, sendo que alguns fenômenos são interpretados e nomeados diferentemente entre os autores. Entre as seções que contemplam o que procuramos estão *preposição, verbo auxiliar, locução verbal, verbo modal* e *conjunção*.

3.1.1. Sobre as locuções verbais

Na *Gramática da Língua Portuguesa* [12], os autores Vilela e Koch enquadram as construções [estar + a/para + $V_{infinitivo}$] em duas categorias simultaneamente, sem distinção: verbos copulativos e verbos auxiliares de aspecto. Ambas as categorias estão dentro da seção *Verbos plenos e verbos auxiliares*, de tal modo que o primeiro verbo se configura como um verbo auxiliar, e o segundo, pleno. Segundo os autores, portanto, nas construções acima (1)-(3), a forma "estar" é a forma:

> "(...) em que o peso gramatical é preponderante, ou porque o verbo se deslexicalizou e reforçou o seu peso gramatical (gramaticalizando-se) e necessita de um verbo pleno para poder funcionar como predicado ou porque o núcleo predicativo é constituído por um nome (*ter*

> *consideração por*), por um adjetivo (*ser inteligente*)" (Vilela e Koch, 2001, p. 72).

Por outro lado, o verbo no infinitivo nas construções (1)-(3) é verbo pleno, o que ocorre quando:

> "(...) o conteúdo se dirige diretamente para a configuração da processualidade existente no mundo extralinguístico e que gramaticalmente pode funcionar como predicado da frase sem qualquer apoio ou suporte" (Vilela e Koch, 2001, p. 72).

Dentro da categoria de auxiliares de aspecto, esta que contém a construção em foco [estar + a/para + $V_{infinitivo}$], há também as construções do tipo [começar/continuar + a + $V_{infinitivo}$]. É possível concluir, portanto, que, de um ponto de vista formal, as construções são análogas e devem receber o mesmo tratamento morfossintático. Ainda dentro da seção de verbos auxiliares, Vilela e Koch inserem os verbos auxiliares de tempos compostos [ter/haver + $V_{particípio}$] e auxiliares de modo [querer[2] + $V_{infinitivo}$].

Embora as construções com o "estar" sejam mais marcadamente auxiliares, em outras construções, entretanto, julgar se um verbo está mais ou menos gramaticalizado, isto é, se funciona ou não como um verbo auxiliar em um dado contexto não é tarefa simples. Depende, por exemplo, de uma análise que compare o uso/função do verbo em uma sentença com o uso/função, do mesmo verbo, em outros contextos que evidenciem o chamado sentido pleno desse verbo. A categorização como verbo pleno ou auxiliar pode, portanto, divergir para diferentes anotadores, assim como diverge para diferentes gramáticas.

Vejamos, como exemplo, a sentença (4) abaixo. Nela, o verbo "começar" está, indubitavelmente, exercendo sua função plena: é predicado verbal da sentença cujo sujeito é "A corrida sucessória". No entanto, nas frases (5) e (6), embora as gramáticas nos digam que as formas de "começar" indicam apenas o aspecto do segundo verbo – aspecto inceptivo –, aceitamos como absolutamente possível uma leitura que assume o sentido pleno de "começar", assim como em (4). Conseguimos, inclusive, parafrasear (5) e (6) como "começou a coordenação" e "começa a preparação", respectivamente.

É possível argumentar que a diferença é sintática: em (4) o verbo é intransitivo e, em (5)-(6), possivelmente transitivo. A observação da frase (7), porém, derruba a tese, pois trata-se de um "começar" transitivo com sentido pleno.

(4) CF288-3 – A corrida sucessória **começa** esta semana com um quadro mais claro e definido do que o da semana passada.

(5) CF28-1 – Pela segunda vez desde quando **começou** a coordenar as ações no Rio, há duas semanas, o Exército mudou o nome das operações.

(6) CF118-2 – O escritório de Júlio Neves já **começa** a preparar novos estudos para o prolongamento deste corredor além do shopping Morumbi, em direção à ponte do Socorro.

[2] São alguns dos poucos verbos inseridos na categoria *auxiliares de modo*, junto com [*ter de/que*, *dever* e *poder* + $V_{infinitivo}$].

(7) CF39-2 – Diniz **começou** sua carreira automobilística em 1989, no Brasileiro de Fórmula Ford, campeonato em que obteve a sexta posição na classificação final.

Mesmo considerando que haveria um certo consenso em dizer que [começar + a + $V_{infinitivo}$] é uma locução verbal aspectual, Vilela e Koch não deixam claro se a noção de aspecto inclui as partículas "a", "de" e "para" no escopo da locução verbal aspectual ou se elas não estão auxiliando na noção de aspecto junto ao verbo auxiliar. Como consequência, nós, na tarefa de anotação, precisamos decidir se tais partículas também podem ser consideradas de função auxiliar e qual sua classe gramatical.

3.1.2. Sobre as preposições

Sem muito pensar no assunto, costumamos associar as formas das palavras "a", "de" e "para" à de preposições, que, sendo (ilusoriamente) poucas, decoramos[3]. Suponhamos que se possa dizer que as partículas "a", "de" e "para", que aparecem nas construções [comecei + a + $V_{infinitivo}$], [acabei + de + $V_{infinitivo}$] e [estar + para + $V_{infinitivo}$], são preposições. Em seu *Dicionário de Linguística e Gramática* [11], Mattoso Câmara Jr. considera que preposições são:

> "vocábulos que servem de morfema de relação para subordinar um substantivo como: adjunto a outro substantivo ou como complemento a um verbo. Esse processo de subordinação tem o nome de regência" (Câmara Jr., 1978, p. 198).

No entanto, ao postular que a partícula "a" em [começar + a + $V_{infinitivo}$] é preposição, fazemos diferente do que Mattoso definiu pois no caso das locuções verbais não há um substantivo sendo relacionado, mas dois verbos. Ou seja, o segundo verbo, quando muito, seria complemento do primeiro, e preposições não fazem relação entre verbos. Não é à toa que, no verbete "Aspecto", Mattoso descreve as conjugações perifrásticas[4] com "estar" sem classificar a partícula "a": "[as conjugações perifrásticas se constituem de] o verbo auxiliar *estar*, conjugado com um gerúndio ou um infinitivo regido de *a*" (Câmara Jr., 1978, p. 61).

Ao lidar com verbos modais, em sua *Gramática pedagógica do português brasileiro* [9], Bagno afirma que os verbos modais são auxiliares e os verbos que os seguem, seu complemento:

> "a construção com os verbos modais se faz sempre com infinitivos na posição de verbo principal. Ao mesmo tempo, os verbos principais se constituem o complemento direto do verbo modalizador" (Bagno, 2012, p. 572).

Nesse ponto, a proposta vai ao encontro de Mattoso, pois, fazendo vista grossa e assumindo que preposições podem introduzir complementos que são verbos (orações

[3] Bagno [9] nos alerta para a inadequação da tradição gramatical no tocante às preposições: decoramos, em média, 17, entre as quais poucas ainda são utilizadas contemporaneamente, e deixamos de fora outras tantas que são mais usuais.

[4] Conjugações perifrásticas, como definidas em [7], têm definição idêntica à de locução verbal com que estamos lidando.

subordinadas substantivas objetivas indiretas reduzidas de infinitivo), o segundo verbo, nessas construções, é complemento do primeiro.

Essa proposta, no entanto, traz algumas incongruências. Em primeiro lugar, do ponto de vista da anotação no ambiente UD, há uma contradição se levarmos ao cabo a observação de Bagno: não podemos considerar que o segundo verbo em [querer/poder/precisar + $V_{infinitivo}$] é, ao mesmo tempo, complemento do primeiro verbo *e* verbo principal de uma locução verbal, pois, em UD, verbos auxiliares não podem ter complemento. Se encaramos que o segundo verbo é complemento do primeiro, ambos devem ser plenos.

Em segundo lugar, porque Câmara Jr. argumenta, no verbete de conjugações perifrásticas, que:

> "É má técnica de descrição gramatical considerar formas perifrásticas a combinação de dois verbos numa única oração em que ambos guardam a sua significação verbal e a significação total é uma das significações (**quero sair** – **vamos conversando** até a casa – já **tenho** uma carta **escrita**) e não houve a gramaticalização do primeiro verbo" (Câmara Jr., 1978, p. 80).

Com a afirmação, além de partir do pressuposto de que seja fácil identificar quando um verbo guarda sua significação total – já vimos que não o é –, o autor utiliza o mesmo exemplo prototípico de "locuções verbais modais" – *quero sair* – para dizer que não concorda que se trate de uma locução verbal, na contramão tanto de [12] quanto de [13], que consideram tais construções como locuções verbais.

3.2. Desbravando o Bosque-UD

Para lidar com as partículas "a", "de" e "para" no centro das locuções verbais, nenhuma consulta a gramática nos foi especialmente relevante. No entanto, uma exposição de verbos na *Gramática Pedagógica do Português brasileiro* [13] lançou luz sobre um dado que nos parece esclarecedor (Tabela 1).

Verbo auxiliar	Exemplo
acabar	Ana acabou desistindo de viajar em julho.
acabar por	Ana acaba de desistir de viajar em julho.
acabar por	Ana acabou por viajar em julho.
andar	Ana anda pensando em viajar em julho.
cessar de	Ana ainda não cessou de sofrer com a separação.
começar	Ana começou falando dos pais

Tabela 1: Verbos auxiliares (6 primeiras entradas) em [13], p. 604

A opção que Bagno faz por colocar as partículas “de” e “por” junto ao verbo na primeira coluna, mesmo que sem qualquer comentário sobre essa colocação, nos diz algo. A noção de auxiliaridade, de fato, comparece quando o verbo é acompanhado por tais partículas, evidenciando assim, por exemplo, que “acabar” seria diferente de “acabar de”; “vir” seria diferente de “vir a”, e, do mesmo modo, “começar a” seria diferente de “começar”.

A colocação de partículas (ou preposições) próximas ao verbo nos faz lembrar do conceito de *phrasal verbs* em inglês, quando uma preposição se junta a um verbo, criando uma entrada diferente tanto do verbo de origem quanto da preposição originária. Ainda que tenhamos dificuldade em chamar tais partículas de preposição, indicar que estamos diante de fenômenos semelhantes ao de *phrasal verbs* nos parece adequado. Para lidar com a anotação morfossintática das partículas “a”, “de”, “para” e “por”, portanto, assumiremos que elas são exigidas pelos verbos auxiliares, quando queremos torná-los auxiliares.

Continuando com a análise, no contexto UD, diríamos então que “começar a”, “vir a” e “acabar de” são expressões multi-palavras (MWEs) e, nesses casos, a partícula associada se une ao verbo auxiliar pela relação de dependência *compound* (relação usada para os *phrasal verbs* do inglês). Como consequência, tem-se uma MWE do tipo *verbo auxiliar* e dependente do verbo principal, como anotado no formato UD nas figuras (13)-(17).

Figura 13. CP3-4 – «Estamos a dotar os computadores de um novo sentido» disse Steve d’Averio, director de marketing para a Europa da Logitech.

Figura 14. CF835-12 – Cursando economia na Faap, Kiko espera ansioso o seu telefone celular, que está para sair».

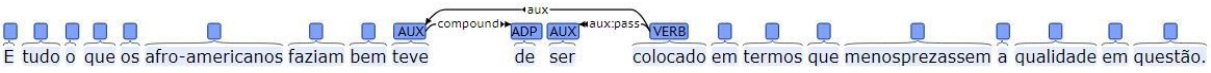

Figura 15. CF27-5 – E, assim, tudo o que os afro-americanos faziam bem teve de ser colocado em termos que menosprezassem a qualidade em questão.

Figura 16. CF28-1 – Pela segunda vez desde quando começou a coordenar as ações no Rio, há duas semanas, o Exército mudou o nome das operações.

Figura 17. CF118-2 – O escritório de Júlio Neves já começa a preparar novos estudos para o prolongamento deste corredor além do shopping Morumbi, em direção à ponte do Socorro.

4. Conversão entre formatos de anotação linguística: o formato AC/DC e o UD

A disponibilização de corpora em diferentes formatos é um fator importante para o desenvolvimento do PLN em uma determinada língua pois diferentes sistemas podem utilizar diferentes formatos e, independentemente de formato, um mesmo corpus linguístico pode alimentar diferentes sistemas. Por este motivo, montamos um conversor de formato AC/DC (Acesso a Corpora, Disponibilização de Corpora) [5] para o formato UD (Universal Dependencies) [4] e vice-versa, e doravante descrevemos os desafios que foram encontrados durante a conversão.

O formato AC/DC segue uma codificação em XML, e o UD, em CoNLL-U. Neste relatório tratamos apenas da conversão da estrutura de corpora codificados nos dois formatos. Para o futuro, ainda, será necessário pensar em formas de tradução da gramática subjacente aos formatos, nomeadamente a gramática do UD e a gramática do PALAVRAS [14], anotador morfossintático dos corpora no formato AC/DC.

Para realizar a conversão do formato, escrevemos o script *ACDC-UD.py*, em Python, disponível em repositório no GitHub[5]. O script pode ser utilizado em três ocasiões distintas.

O primeiro uso é para converter um corpus no formato AC/DC para o formato do UD, processo descrito na seção 4.1. Essa conversão mantém todos os metadados necessários para que, posteriormente, esse corpus convertido, já no formato CoNLL-U, seja convertido de volta para o formato AC/DC sem nenhuma perda de informação. Por fim, o script pode também ser utilizado na conversão de um corpus originalmente no formato UD para o formato AC/DC. O próprio código trata de reconhecer se o corpus inicial está no formato AC/DC ou UD, e, no caso de UD, se ele é fruto de uma conversão do AC/DC ou não.

4.1. Conversão de AC/DC para UD

A conversão do formato AC/DC para o UD ocorre mantendo todos os metadados necessários para que, posteriormente, seja possível retornar do UD para o AC/DC sem perda de informação. Por conta disso, ao converter o corpus OBRAS [15], no formato AC/DC, para o UD, obtemos as características de tamanho descritos na tabela 2. Como a volta ocorre de forma que o resultado final seja igual ao inicial, estes tamanhos também próximos, como seria de se esperar.

[5] Disponível em <https://github.com/alvelvis/ACDC-UD/blob/master/ACDC-UD.py>.

	AC/DC original	**UD**	**AC/DC convertido do UD (processo de volta)**
OBRAS	2.047 mb	2.076 mb	2.190 mb

Tabela 2: Tamanho do corpus OBRAS nos diferentes formatos

Em primeiro lugar, é importante notar que todas as etiquetas do XML do arquivo AC/DC original serão repassadas para o arquivo UD, seja no metadado "xml_tags", ou no campo *misc* (décima coluna) dos tokens. Essa memória das tags do XML é imprescindível para que, ao retornar do arquivo UD para o AC/DC, nenhuma informação seja perdida.

Todas as etiquetas do XML que tenham o atributo *id* (no OBras, "autor", "obra" e "tituloobra"), se tornam metadados do arquivo CoNLL-U de todas as sentenças, facilitando, assim, o reconhecimento da origem de uma dada frase no corpus. Já o metadado "sent_id", tradicionalmente utilizado no UD para identificar as sentenças, é composto utilizando a etiqueta mais importante do XML seguida de uma numeração progressiva. Por exemplo, no OBras, essa etiqueta é a "obra", de modo que uma "sent_id" possível no CoNLL-U poderia ser "O_Homem_que_Sabia_Javanês_e_Outros_Contos-Prosa:contos-LB-1997-masc--2059", sendo "2059" o número da frase dentro da determinada obra.

Cabe considerar que, no AC/DC, os tokens têm muito mais colunas (ou atributos) que no UD, podendo ultrapassar 20 colunas, enquanto no UD a quantidade é de 10 colunas, invariavelmente. Além disso, enquanto no AC/DC a organização do arquivo se dá por obras, e, dentro das obras, há as sentenças (com as etiquetas "<u>", "<s>" e "<t>" separando-as, embora nem sempre todas as etiquetas estejam presentes em todas as sentenças), no arquivo UD a organização se dá somente por sentenças, de modo que a cada linha em branco ("\n\n") se inicia uma nova sentença.

4.2. Atributos estruturais e morfossintáticos

Tanto em um quanto em outro formato a delimitação de colunas é feita com tabulação ("\t"), mas não há uma correspondência imediata entre as colunas de ambos. Por isso, o alinhamento foi feito da seguinte forma:

* O que, no UD, corresponde à primeira coluna, *id*, no AC/DC corresponde à coluna 18, já que nela há as informações de id do token e qual seu pai sintático, por exemplo, "1->2" (leia-se: token número 1, cujo pai é o token número 2); desse modo, o que está à direita da seta "->", na 18ª coluna do AC/DC, corresponde à 7ª coluna do UD, onde se insere o pai sintático do token.

* A segunda coluna do UD corresponde à *word*, o que, no AC/DC, encontra-se na primeira coluna.

* Já a terceira coluna do UD, correspondente ao *lemma*, encontra-se somente na 9ª coluna do AC/DC.

* A quarta coluna do UD, *upostag*, que corresponde à classe gramatical, pode ser alinhada com a décima coluna do AC/DC.

* Já a quinta coluna do UD, *xpostag*, foi deixada em branco, com um underline.

* A sexta coluna do UD, de *features* (ou atributos morfológicos), foi encontrada ao concatenar três valores do AC/DC: a coluna 11, que corresponde ao tempo e categoria verbal; a coluna 12, que corresponde a número e pessoa; e a coluna 13, que corresponde ao gênero. Como de costume em UD, esses três valores foram separados por uma barra vertical.

* A oitava coluna do UD, de *deprel* (relação de dependência), é alinhável com a 14ª coluna do AC/DC;

* e, por fim, colocamos, na nona coluna do UD, o valor da 16ª coluna do AC/DC, que corresponde ao *sema*, embora esse não seja o uso padrão do UD para esta coluna, mas assim o fizemos por tratar-se de uma informação relevante. Todas as demais colunas do AC/DC que não foram contempladas nas colunas do UD foram colocadas na décima coluna do UD, *misc*, separadas por uma barra em pé, e com o número da coluna antecedendo o valor, como no exemplo:

3=LimBar|5=desc|7=orig|14=0|16=0|17=9->7|18=correr|19=__UNDEF__|20=__UNDEF__|21=__UNDEF__

Deve-se notar também que, embora a 18ª coluna do AC/DC, de id e dependência sintática, esteja presente nas colunas do UD, ela também foi colocada no *misc* para garantir que, no retorno do UD para o AC/DC, não haverá conflitos já que, por conta de MWEs e contrações, o id dos tokens, em UD, pode sofrer alteração. O que nos leva a pensar na próxima questão: como lidar com MWEs e contrações?

4.3. MWEs e contrações

As MWEs no AC/DC, como "de=quando=em=quando", não são contraídas, isto é, cada palavra se apresenta como um token independente: "de/quando/em/quando". Para indicar que se trata de uma MWE, no XML há a etiqueta "<mwe lema=X pos=Y>", que adicionamos ao *misc* dos tokens para recuperá-la na volta do UD para o AC/DC. Mas cabe uma alteração cuidadosa na conversão para o UD: os 4 tokens de "de quando em quando", no AC/DC, têm o mesmo id, mas com a indicação, ao lado, de a qual "parte" da MWE o token se refere. Por exemplo, se a MWE começar com um token de id 5, e a primeira palavra for o head da MWE:

de 5-1->X
quando 5-2->5-1
em 5-3->5-1
quando 5-4->5-1

Esta numeração com hífen é inexistente em UD, por isso precisamos realizar a conversão para que a MWE fique da seguinte forma:

de 5->X
quando 6->5

em 7->5
quando 8->5

Essa mudança nos ids, é claro, faz com que toda a frase se reconfigure com base nos novos ids e, consequentemente, os *depheads* também devem ser alterados. Por isso, antes de executar as mudanças de MWE, realizamos um "mapeamento de dephead", indicando quais tokens dependem sintaticamente de quais tokens. Então, após realizar a mudança de ids, refazemos a dependência sintática com base nos novos ids.

Já em relação às contrações, o AC/DC não as descontrai, de modo que o exemplo (1) do AC/DC deve se tornar (i) em UD, e (2) deve-se tornar (ii).

(1) ao a+o PRP+DET_artd <ADVL+>N 35->34+36->38
(i)
35-36	ao	_	_	_	_	_	_	_	_
35	a	a	PRP	34	<ADVL				
36	o	o	DET_artd	38	>N				

(2) ouvia-me ouvir+eu 4->29+5->4
(ii)
4-5	ouvia-me	_	_	_	_	_	_	_	_
4	ouvia	ouvir	29						
5	me	eu	4						

Cabe notar que, no AC/DC, em uma palavra como "da", não há nenhuma indicação de que os tokens, quando descontraídos, devem ter word "de" e "a", apenas que os lemas devem ser "de" e "o". Em razão disso, foi construído um dicionário de contrações e como devem ser descontraídas, do qual retiramos alguns exemplos na tabela 3.

Forma contraída	**Tokens descontraídos**
No	Em+o
pela	por+a
dos	de+os
da	de+a
do	de+o
na	em+a
desse	de+esse
ao	a+o
à	a+a
num	em+um

Tabela 3: 10 das 253 contrações encontradas no corpus OBRAS

5. Considerações finais

Com este relatório, o projeto "Construção de datasets para o PLN de língua portuguesa" chega ao fim. Neste ano, cumprimos o objetivo de melhorar a qualidade e a quantidade de material disponível publicamente para o processamento automático de linguagem natural em língua portuguesa, tendo, na primeira etapa (2018-2019), realizado correções sistemáticas e bem documentadas no principal corpus de textos de jornal em língua portuguesa com revisão humana e utilizado como dataset no projeto Universal Dependencies, o Bosque-UD. Em seguida, documentamos, na segunda etapa do projeto (2019-2020), as decisões linguísticas relativas à gramática UD para língua portuguesa que foram tomadas levando em conta as diretivas do projeto e as principais gramáticas do português; e, finalmente, também nesta etapa, tornamos possível a conversão de corpora em dois formatos distintos e dos mais relevantes na área: o AC/DC e o UD.

Tratou-se, portanto, de um projeto cujo foco foi o desenvolvimento da infraestrutura da área de pesquisa relativa ao PLN. No futuro, tendo em vista os resultados relatados neste e no relatório anterior, é possível uma vasta gama de aplicações práticas que visem desenvolver o processamento do português e/ou pavimentar o uso dessa tecnologia para aplicações na pesquisa e na indústria por meio do processamento automático de textos em áreas diversas, como educação, estudos sociais, tradução etc.

Referências

1 - DE SOUZA, Elvis; FREITAS, C. "Relatório anual (2018-2019) do projeto 'Construção de datasets para o PLN de língua portuguesa'". (Relatório de pesquisa). 2019. Disponível em: http://comcorhd.letras.puc-rio.br/relatorio-anual-2018-2019-do-projeto-construcao-de-datasets-para-o-pln-de-lingua-portuguesa/

2 - RADEMAKER, Alexandre et al. "Universal dependencies for Portuguese". In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017). 2017. p. 197-206.

3 - DE SOUZA, Elvis; FREITAS, Cláudia. "ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente". In VI Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana (TILic 2019). TILic 2019, Salvador, BA, Brazil, Outubro, 15-18, 2019.

4 - NIVRE, Joakim, et al. "Universal dependencies v1: A multilingual treebank collection." **Proceedings of the Tenth International Conference on Language Resources and Evaluation** (LREC'16). 2016.

5 - SANTOS, Diana; SARMENTO, Luís. "O projecto AC/DC: acesso a corpora/disponibilização de corpora". In Amália Mendes; Tiago Freitas (ed) **Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística** (APL 2002)(Porto Portugal 2-4 de Outubro de 2002) Lisboa: APL, 2003.

6 - NIVRE, Joakim; FANG, Chiao-Ting. Universal dependency evaluation. In: **Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies** (UDW 2017). 2017. p. 86-95.
7 - DE SOUZA, E.; CAVALCANTI, T.; SILVEIRA, A.; EVELYN, W.; FREITAS, C. **Diretivas e documentação de anotação UD em português (e para língua portuguesa)**. Departamento de Letras, PUC-Rio, 2020. Disponível em: http://comcorhd.letras.puc-rio.br/documenta-o-ud-pt/
8 - SAMPSON, Geoffrey. **English for the computer: The SUSANNE corpus and analytic scheme**. 2002.
9 - BONIAL, Claire et al. Propbank annotation guidelines. **Center for Computational Language and Education Research, CU-Boulder**, 2010.
10 - SAMPSON, Geoffrey. **Empirical linguistics**. A&C Black, 2002.
11 - CÂMARA, Joaquim Mattoso. **Dicionário de lingüística e gramática: referente à língua portuguesa**. Vozes, 1978.
12 - VILELA, Mario; KOCH, Maria Ingedore Vilaça. **Gramática da Língua Portuguesa: gramática de palavra, gramática de frase e gramática de texto/discurso**. Coimbra: Almedina, 2001.
13 - BAGNO, Marcos. **Gramática pedagógica do português brasileiro**. Parábola Ed., 2012.
14 - BICK, Eckhard. "The parsing system palavras: Automatic grammatical analysis of Portuguese in a constraint grammar famework". Aarhus Universitetsforlag, 2000.
15 - SANTOS, Diana; FREITAS, Cláudia; BICK, Eckhard. "OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain." In CorLex 24 de setembro de 2018, 2018.