Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1033-1058, 2021

Sujeito oculto às claras: uma abordagem descritivo-computacional

Omitted subjects revealed: a quantitative-descriptive approach

Cláudia Freitas
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Rio de Janeiro / Brasil
claudiafreitas@puc-rio.br
https://orcid.org/0000-0001-6807-8558

Elvis de Souza
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Rio de Janeiro / Brasil
elvis.desouza99@gmail.com
https://orcid.org/0000-0001-9373-7412

Resumo: Neste trabalho, apresentamos estudos descritivos e computacionais relacionados ao sujeito oculto. Em um primeiro momento, realizamos uma descrição de cunho quantitativo, tomando por base três *corpora* dos gêneros jornalístico, literário e enciclopédico. Especificamente, quantificamos o sujeito oculto em cada um dos *corpora*, e encontramos sujeitos omitidos em 24%, 41% e 46% das orações, respectivamente. Em um segundo momento, por meio de uma estratégia baseada em regras, reconstituímos esses sujeitos e os devolvemos aos *corpora*, com o objetivo de avaliar o quanto a omissão do sujeito é capaz de impactar o aprendizado automático de dependências sintáticas. Os resultados indicam que a reconstituição formal do sujeito pode melhorar a aprendizagem das dependências sintáticas em até 2% quando consideramos a métrica CLAS, evidenciando o papel relevante da modelagem linguística no aprendizado automático.

Palavras-chave: descrição linguística; sujeito oculto; omissão de sujeito; dependências sintáticas; linguística computacional; aprendizado de máquina; linguística de *corpus*.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1033-1058

Abstract: In this paper, we present descriptive and computational studies related to omitted subjects. Firstly, we develop a quantitative descriptive study based on three *corpora*, which consist of journalistic, literary and encyclopedic genres. Specifically, we quantify the omitted subjects in sentences for each of these *corpora;* omitted subjects were found in 24%, 41% and 46% of their sentences, respectively. Secondly, applying rule-based strategies, we reconstitute those subjects and place them back to the *corpora*, with the goal of evaluating how much the omission of subjects can impact the automatic learning of syntactic dependencies. The results indicate that the formal subject reconstitution can enhance the learning of syntactic dependencies in up to 2% according to the CLAS metric, highlighting the relevant role of linguistic modeling in the automatic learning process.

Keywords: linguistic description; omitted subject; syntactic dependencies; computational linguistics; machine learning; corpus linguistics.

Submetido em 07 de outubro de 2020

Aceito em 14 de dezembro de 2020

1 Introdução

A articulação entre o Processamento (automático) de Linguagem Natural (PLN) e os estudos linguísticos vem ganhando força nos últimos anos, alterando um pouco o quadro descrito em 2007 por Karen Sparck-Jones quando constata o distanciamento entre a linguística e a linguística computacional. Muito dessa reaproximação se deve ao trabalho de anotação de *corpora* que, como já apontado em Sampson (2001), é, também, um trabalho de descrição linguística.

Neste artigo, contribuímos com mais um elemento na aproximação entre os dois campos, e o fazemos não pelo viés da anotação, mas partindo de *corpora* já anotados para a descrição de um fenômeno linguístico de grande relevância para uma série de tarefas de PLN em português: o sujeito oculto. Após uma caracterização linguística do fenômeno, voltamo-nos para o PLN, a fim de medir o quanto a ausência de sujeitos em uma oração pode dificultar o processamento sintático automático.

Uma das áreas de atuação do PLN é a extração de informação (EI). Ainda que, tradicionalmente, a extração de informação consista na detecção automática de informações relativa a certos atores pré-definidos, como pessoas, lugares e organizações para indicar, simplificadamente, *quem faz o quê*, tomada em um sentido amplo, várias tarefas do PLN podem

ser agrupadas como variações de um jogo que consiste em vasculhar uma imensa quantidade de textos com o objetivo de encontrar algum tipo de conteúdo (ou informação), e incluímos aqui tanto opiniões quanto dados factuais sobre alvos pré-determinados. A partir daí, com a concatenação dos conteúdos extraídos, podem-se construir novos fatos, ou hipóteses, que serão explorados posteriormente. Em termos metodológicos, a busca pelo conteúdo se dá por meio da identificação de padrões que podem ser codificados como relações (proposições) com um número variável de argumentos e/ou modificadores, como indicado abaixo:

Maria$_{ARG1}$ *estava triste*$_{Pred}$
Maria$_{ARG1}$ *sorriu*$_{Pred}$
Maria$_{ARG1}$ *comprou*$_{Pred}$ *uma bicicleta*$_{ARG2}$
Maria$_{ARG1}$ *emprestou*$_{Pred}$ *dinheiro*$_{ARG2}$ *para o irmão*$_{ARG3}$ mês passado$_{MOD1}$

Ainda que consideremos os diferentes focos de interesse envolvidos na identificação e extração de conteúdos em textos, em boa parte deles a identificação do agente responsável pelas ações/atividades detectadas é crucial. Este *quem*, em geral, se manifesta como o sujeito da oração, e por isso a identificação do sujeito é de extrema relevância para uma boa parcela de tarefas, como identificação de papéis semânticos, identificação e análise de opiniões e sentimentos, extração de citações, além da extração de informação propriamente.

No entanto, na língua portuguesa, diferentemente do inglês (língua que concentra boa parte das pesquisas em PLN), a possibilidade de omissão do sujeito em contextos em que este é facilmente recuperável, o chamado "sujeito oculto", é um dado a mais a ser considerado. Trata-se de um fenômeno cuja resolução, por pessoas, costuma ser trivial, mas que, para as máquinas, impede a construção de relações (ou proposições) adequadas, porque deixa de levar em conta um dos argumentos da proposição. Marcus *et al*. (1993), no contexto do *Penn Treebank*, afirmam que a maneira mais simples de incluir informação sobre a estrutura de predicado-argumento é permitindo que a árvore sintática contenha, de maneira explícita, os elementos nulos.

Além disso, como apontam Hartmann *et al*. (2014) no contexto da língua portuguesa, a inserção de elementos nulos contribuiria para reduzir o problema da escassez de dados, sendo, portanto, um aspecto altamente relevante em abordagens de aprendizagem de máquina.

Ou seja, a presença de sujeitos ocultos traria um desafio adicional ao processamento sintático automático.

Por fim, no que se refere ao processamento humano e ao letramento, é possível que a omissão do sujeito – como, aliás, a omissão de qualquer termo na estrutura frasal, conforme sugerido em Finatto *et al.* (2011) – seja mais um elemento a trazer complexidade aos textos. Deste modo, estratégias capazes de identificá-lo com precisão contribuem para a verificação do grau de dificuldade de um texto. E, do mesmo modo, estratégias capazes de reconstituí-los promoveriam uma simplificação textual.

Do ponto de vista do PLN de língua portuguesa, o quanto de informação perdemos quando não explicitamos o sujeito? Precisamos de algum tratamento especial para resolver a questão, ou se trata de um fenômeno periférico? Tais perguntas poderiam ser facilmente respondidas pelos estudos descritivos, mas não sabemos de nenhum que se dedique ao tema de um ponto de vista quantitativo. É neste espaço que nos colocamos, trazendo, em uma primeira etapa, dados de grandes *corpora* com características textuais distintas: um *corpus* de textos jornalísticos, um *corpus* de textos literários e um *corpus* de textos enciclopédicos, que juntos somam mais de 17 milhões de unidades/[1] 685 mil frases. Os resultados mostram que a ocorrência de sujeitos omitidos não é desprezível, indo de 24% a 46%, conforme o tipo de texto.

Em uma segunda etapa, informados de que a quantidade de orações que não apresentam um sujeito sintático explícito é relevante, realizamos um experimento de PLN no qual reintroduzimos os sujeitos omitidos e treinamos um modelo de aprendizagem de máquina, com o objetivo de verificar o quanto a omissão de sujeito prejudica o processamento automático. Os resultados mostram uma melhora de até 2% no aprendizado quando reconstituímos os sujeitos, sugerindo que um *dataset* cuja construção leva em conta características linguisticamente motivadas é capaz de contribuir no processamento automático no nível sintático.

Ao longo do artigo, apresentamos estudos sobre o sujeito oculto na língua portuguesa a partir de duas perspectivas distintas,

[1] Ao longo do texto, usamos a palavra *unidade* como uma tradução do inglês *token,* isto é, uma unidade mínima de anotação, já tendo sido separadas as contrações de verbos e pronomes, preposições e artigos, etc.

mas complementares: de um ponto de vista linguístico-descritivo, quantificamos o fenômeno do sujeito oculto; de um ponto de vista linguístico-computacional, medimos o quanto este fenômeno dificulta o processamento automático. Em ambos os casos, utilizamos *corpora* e ferramentas públicos.

O presente estudo é conduzido sob o *framework* do projeto *Universal Dependencies* (NIVRE *et al*., 2016), uma abordagem multilíngue para o processamento computacional das línguas. Ao apresentar aqui os dados e procedimentos para o português, esperamos também contribuir para um quadro mais geral de descrição de diferentes línguas no que se refere à omissão do sujeito.

O restante do artigo se organiza da seguinte maneira: na seção 2, apresentamos o fenômeno do sujeito oculto em português, especificando o objeto que nos interessa tratar; na seção 3, indicamos alguns trabalhos sobre o tema na perspectiva dos estudos com *corpus* e do PLN, e na seção 4 detalhamos a metodologia do estudo descritivo, cujos resultados são analisados na seção 5. Por fim, na seção 6 relatamos um pequeno experimento cujo objetivo é medir o impacto da omissão do sujeito na aprendizagem sintática; e na seção 7 tecemos algumas considerações finais.

2 A omissão do sujeito em português

A língua portuguesa licencia a omissão de sujeito nas frases de diferentes maneiras: (i) o sujeito oculto propriamente, ou sujeito elíptico; (ii) o chamado sujeito indeterminado e, ainda, (iii) as orações sem sujeito (veja-se por exemplo LUFT, 2002). Especificamente, nos interessam os casos do primeiro tipo (frases 1, 5, 6 e 7, abaixo), e frases do segundo tipo (frase 2). Deste modo, igualamos o que gramáticas tradicionais distinguem, já que, em ambos os casos, existe um sujeito a ser explicitado, ainda que não haja sujeito formal. Pelo mesmo motivo, não levamos em conta as chamadas orações sem sujeito, como frases com verbos impessoais (frase 3) e com verbos indicativos de fenômenos da natureza (frase 4), já que, nesses casos, não há sujeito a ser explicitado.[2]

[2] Todos os exemplos foram retirados do corpus Bosque, a parte revista do projeto Floresta Sintá(c)tica (AFONSO *et al*., 2002; FREITAS *et al*., 2008).

1. "Eu tentei, o senhor Vance tentou, se for respeitado, urrah!", comentou.
2. Sempre que surge um problema, chamam-na.
3. Há, no ar, uma certa ideia de invasão.
4. Desde há já alguns anos que chove «a eito» na centenária Igreja de Ceide.
5. A empresa norte-americana informou à PF que não tem filiais ou representantes no Brasil.
6. Só depois é que levanto a cabeça para fazer um lançamento», reclama Neto.
7. Herbert Berger, diretor-superintendente da empresa, diz que o Charade «se aproxima do Honda Civic em tamanho e custa bem menos».

Os casos 1 a 4 são aqueles tradicionalmente mencionados quando se aborda o fenômeno da omissão de sujeito em português. Em (5) e (6), temos omissão de sujeito em orações subordinadas; e, em (7), a omissão em uma oração coordenada. Diferentemente das frases (1) e (2), nesses casos o sujeito se encontra nos limites da frase. No entanto, se, do ponto de vista discursivo, (1) e (2) são problemas mais interessantes, pois a explicitação do sujeito envolve a busca por referentes além da frase, do ponto de vista do processamento automático, que privilegia os limites da oração, as frases (5), (6) e (7) podem ser igualmente complexas. Em nosso estudo, fazemos duas contagens: uma para medir os sujeitos de orações principais, e outra para os sujeitos ocultos de orações subordinadas.

Por fim, sabemos também que, em português, é possível que o *se* materialize uma indeterminação do sujeito, como nas frases (8) e (9). No entanto, e diferentemente dos demais casos de indeterminação, com o *se* não é possível identificar quem é o sujeito; não é possível determiná-lo. Como nosso interesse está em apenas distinguir o sujeito oculto, também excluímos esses casos de nossa contagem.

8. Tem que se demonstrar através de contas e de raciocínios que o expurgo significará perda
9. Diga-se, de uma vez e claramente, o que se quer ou o que se quer mais.

3 Trabalhos relacionados

O trabalho de Hartmann *et al*. (2014) é o único que conhecemos que se debruça sobre o sujeito oculto em português no contexto do PLN. Tendo como foco principal a anotação humana de papéis semânticos, os autores exploram a inserção de elementos artificiais para representar sujeitos omitidos. O objetivo da inserção é preencher algumas lacunas na estrutura sintática das frases, a fim de facilitar a atribuição de papéis semânticos e, assim, melhorar o material de treinamento da tarefa. O *corpus* utilizado foi o PropBank-BR (DURAN; ALUÍSIO, 2012), que foi então analisado pelo anotador PALAVRAS (BICK, 2000). A partir da observação do texto anotado pelo PALAVRAS, os autores criaram regras para inserção automática dos elementos nulos.

Assim como no presente trabalho, o processo de criação de regras foi exploratório e incremental. Os elementos nulos foram preenchidos com pronomes pessoais retos levando em conta a forma flexional do verbo ("eu", "nós", e um genérico *SUBJ* (sujeito) para os demais casos). No entanto, ao que parece, o trabalho considerou uma sintaxe linear, sem informação da dependência, o que dificultou sensivelmente a identificação dos sujeitos (ou de sua ausência), haja vista a quantidade de itens intervenientes que podemos encontrar entre o sujeito e o verbo e também a posição do sujeito em português, que pode estar anteposto ao verbo (posição preferencial) ou posposto. A partir da análise de erros de uma amostra de 200 frases, os autores relatam que a estratégia funcionou em cerca de 80% dos casos e que, quando considerada a inserção por tipo do sujeito, os resultados são heterogêneos: a inserção do sujeito oculto é bem-sucedida em 88% quando o verbo corresponde às primeiras pessoas, mas corresponde a apenas 55.8% dos demais casos, o que se deve, sobretudo, a erros anteriores decorrentes da análise automática.

Do ponto de vista descritivo, a situação não é diferente, e isto certamente se deve à ausência de material com as características técnicas necessárias para o estudo: um *corpus* sintaticamente anotado e uma interface de busca em árvores que permita procurar pela ausência, já que não temos a tradição de anotar elementos nulos em *corpora* – o *Penn Treebank* o faz, e trataremos dele a seguir.

Apesar de já dispormos de bons *corpora* em língua portuguesa, nem sempre estão anotados sintaticamente. O vasto material do projeto AC/DC (SANTOS; BICK, 2000) é uma saudável exceção, mas a interface

de busca não permite buscar por dependências sintáticas ou por elementos nulos, que não estão anotados. Já a última versão dos *corpora* do projeto Floresta Sintá(c)tica (FREITAS *et al*., 2008) – que, assim como projeto AC/DC, é criado e mantido pela Linguateca (SANTOS, 2011) e tem seus *corpora* sintaticamente anotados, também, pelo *parser* PALAVRAS – contém, como um elemento "procurável", duas etiquetas atreladas ao verbo que indicam a ausência de sujeito: <*nofsubj*>, que indica a ausência de um sujeito formal (usada para os casos de oração sem sujeito e verbos indicativos de fenômeno da natureza) e <*nosubj*>, para os casos de sujeito oculto. O *corpus* Bosque é a parte revista de todo o material que compõe a Floresta e descobrimos, buscando pelos respectivos *procuráveis* em uma ferramenta de pesquisa em *treebanks* específica do projeto Floresta, o Milhafre,[3] que há 3622 orações sem sujeito explícito e 266 orações sem sujeito formal. Neste trabalho, queremos verificar a distribuição de sujeitos em diferentes gêneros textuais, e sobretudo naqueles em que a presença de um sujeito agente é crucial para tarefas posteriores: obras literárias, na qual a distribuição de falas entre personagens é um objeto de pesquisa (ELSON; MCKEOWN, 2010; RUANO SAN SEGUNDO, 2016) e textos enciclopédicos – especificamente, uma enciclopédia biográfica sobre a história do Brasil, na qual se pode extrair informações sobre personagens da história política brasileira (HIGUCHI *et al*., 2019).

No que se refere à anotação de *corpus*, ainda que o fenômeno do sujeito oculto não aconteça na língua inglesa, o *tagset* do *corpus Penn Treebank* utiliza a etiqueta PRO para os casos onde existe um sujeito não especificado ou não realizado. Tais casos referem-se especificamente a elementos nulos no imperativo (10); e construções de alçamento e controle, como em (11-13).

10. Go away!
11. $John_i$ seems to PRO_i like Mary
12. $John_i$ promised Mary PRO_i to *write* the book
13. John persuaded $Mary_i$ PRO_i to *write* the book

O *tagset* sintático do projeto *Universal Dependencies* possui uma etiqueta – *xcomp* – para os casos em que

[3] Milhafre. Disponível em: https://www.linguateca.pt/Floresta/milhafre. Acesso em: 8 out. 2020.

> um verbo ou adjetivo é um predicativo ou complemento oracional sem seu próprio sujeito – [isso] não significa que uma oração seja um *xcomp* apenas porque seu sujeito não está omitido. O sujeito deve necessariamente ser herdado de uma posição fixa na oração superior.[4]

Ou seja, a etiqueta *xcomp* inclui (dentre outros fenômenos) casos como 11 e 12 – e apenas eles receberão esta etiqueta.

4 Metodologia

O principal desafio deste trabalho está na metodologia – como encontrar algo sem materialidade, dado que não temos anotação de elementos nulos nos *corpora*. Neste trabalho, usamos a abordagem gramatical do projeto *Universal Dependencies* (UD). O projeto, cujo objetivo é facilitar o desenvolvimento de *parsers* multilíngues e a pesquisa linguística, propõe esquemas de anotação compartilháveis entre línguas para a anotação de classes de palavras, de informação morfológica e sintática. Atualmente, UD conta com mais de 150 florestas (*treebanks*) em 90 línguas diferentes. Como mencionamos na seção 3, o *tagset* de UD conta com uma etiqueta específica para certos casos de omissão de sujeito, e apenas eles. Nos demais casos já mencionados aqui – que correspondem aos exemplos (1-2) e (5-9) – não há uma etiqueta especial. Uma vez que nosso interesse está em medir os casos de omissão de sujeito, não abordaremos, neste momento, as diferenças entre os casos 10-13. Com o auxílio de uma ferramenta desenvolvida especialmente para lidar com *corpora* anotados seguindo o formalismo UD, fomos iterativamente desenvolvendo estratégias e filtros até identificar as frases que nos interessam.

4.1 Os *corpora*

A pesquisa foi realizada em três *corpora* com características distintas. O primeiro deles é o já referido *corpus* Bosque, mas dessa vez em sua versão UD, o Bosque-UD (versão 2.6). Trata-se de um *corpus*

[4] "(…) a verb or an adjective is a predicative or clausal complement without its own subject – [this] does not mean that a clause is an *xcomp* just because its subject is not overt. The subject must be necessarily inherited from a fixed position in the higher clause." *Universal Dependencies guidelines*. Disponível em: https://universaldependencies.org/u/dep/xcomp.html. Acesso em: 8 out. 2020.

composto por textos jornalísticos, com 9.366 frases divididas igualmente entre as variantes do Brasil e de Portugal. Dos três *corpora* utilizados, apenas o Bosque-UD teve sua anotação gramatical revista por linguistas.[5] O fato de a versão original do Bosque conter informações relativas à omissão do sujeito funcionou também como um gabarito para nosso método de procura.

O segundo *corpus* é o DHBB (acrônimo de Dicionário Histórico Biográfico Brasileiro) (HIGUCHI *et al*., 2019). O DHBB é uma enciclopédia sobre a história política brasileira a partir de 1930, criada pelo Centro de Pesquisa e Documentação de História Contemporânea do Brasil, da Fundação Getúlio Vargas (CPDOC/FGV). É um material que interessa especialmente pelo seu conteúdo, configurando-se como uma importante fonte de pesquisa. Desde 2018, o DHBB foi convertido em um *corpus* anotado, integrando o acervo do AC/DC. A versão 6.1 do *corpus* contém 7.700 entradas (ou verbetes), 314 mil frases, cerca de 14 milhões de palavras/16 milhões de unidades e está disponível para consulta e *download* na página da Linguateca.[6]

Por fim, o *corpus* OBras (SANTOS *et al*., 2018). Criado para ser a contraparte brasileira do *corpus* Vercial, o OBras contém obras literárias brasileiras que já estão em domínio público. É um corpus dinâmico e lança novas edições a cada dois meses. Para este trabalho, utilizamos a versão 9.0, que contém 263 obras literárias, 6.8 milhões de palavras/9.7 milhões de unidades. O OBras, assim como o DHBB e o Bosque, integra o AC/DC, o que significa que está ricamente anotado com informação sintática e semântica e disponível para buscas pela internet.

Embora todo o material aqui utilizado já exista em uma versão linguisticamente anotada e disponível para consultas linguísticas e para download,[7] a anotação não contém nenhuma etiqueta relativa à ausência do sujeito, e o AC/DC, embora permita buscas sintáticas, não permite buscas sobre árvores sintáticas. Por isso, reanotamos o material com a ferramenta UDPipe (STRAKA *et al*., 2016).

---

[5] A criação do Bosque-UD está detalhadamente descrita em Rademaker *et al*. (2017).

[6] Disponível em: https://www.linguateca.pt/acesso/desc_dhbb.html. Acesso em: 8 out. 2020.

[7] O OBRAS se encontra disponível em: https://www.linguateca.pt/OBRAS/OBRAS.html, e o DHBB, em https://www.linguateca.pt/acesso/desc_dhbb.html. Acesso em: 8 out. 2020.

4.2 A anotação dos *corpora*

A ferramenta de anotação utilizada foi o UDPipe (STRAKA *et al.,* 2016), uma ferramenta de código aberto que realiza sequencialmente as etapas de tokenização (segmentação do texto em unidades básicas, como palavras e sinais de pontuação), anotação gramatical, lematização e análise de dependências em qualquer *corpus* que esteja no formato CoNLL-U.[8] O UDPipe fornece modelos para quase todos os *treebanks* do projeto UD.[9] O modelo fornecido para o português (versão 2.5) tem índices de acerto (F1) de 96.4%, 95%, 87.2% e 83.1% para os níveis de classes gramaticais (POS), características morfológicas (*feats*), dependência sintática (*unlabeled attachment score* (UAS) e relação de dependência sintática (*labeled atachment score* (LAS)), respectivamente.[10] O *corpus* usado para a geração do modelo de língua portuguesa é o corpus Bosque-UD, já mencionado na seção anterior. Por ser um corpus linguisticamente revisto, o Bosque-UD (assim como o Bosque original) se presta também ao treino e à avaliação de modelos sintáticos, e é esta característica que possibilita a realização do experimento em que medimos a dificuldade de realizar análises sintáticas no que se refere à ausência de sujeitos (seção 5). Para tanto, usamos um modelo criado a partir da versão do Bosque-UD disponibilizada no lançamento 2.6[11] e treinado por nós, utilizando os parâmetros padrão do UDPipe.

O Quadro 1 apresenta os três *corpora* conforme a sentenciação (separação do texto em frases) e tokenização feitas pelo UDPipe (e

[8] O formato CoNLL-U é uma adaptação do formato CoNLL-X. As anotações são codificadas em arquivos de texto simples, com um *token* por linha, e colunas (no máximo 10) que codificam diferentes informações linguísticas, como *lema*, *pos* etc. Uma explicação detalhada do formato pode ser encontrada em https://universaldependencies.org/format.html. Acesso em: 8 out. 2020.

[9] Disponível em: http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_25_models_publications. Acesso em: 8 out. 2020.

[10] Especificamente, as medidas UAS e LAS (*unlabeled attachment score* e *labeled attachment score*, respectivamente) se referem aos acertos de encaixe das dependências sintáticas, sendo que, na segunda métrica, além do encaixe (isto é, além de saber qual o núcleo sintático de um determinado elemento), a relação de dependência sintática também deve estar correta.

[11] Disponível em: https://github.com/UniversalDependencies/UD_Portuguese-Bosque. Acesso em: 8 out. 2020.

que podem não corresponder exatamente àquelas feitas no contexto do projeto AC/DC):

QUADRO 1 – Apresentação quantitativa dos corpora, conforme processado pela ferramenta UDPipe

	DHBB	OBRAS	BOSQUE
Tamanho (Mb)	960	480	14
Tokens	16037286	7863261	227825
Or. Principal	480218	353662	9364
Or. Subordinada	341133	316297	6842
Total De Orações	821351	669959	16206

4.3 A busca pelo sujeito oculto

Para identificar o sujeito oculto, utilizamos o Interrogatório, uma ferramenta para busca e revisão de corpora anotados (de SOUZA; FREITAS, 2019). Nela, é possível realizar buscas sobre árvores sintáticas, desde que estejam em arquivos no formato CoNLL-U. Elencamos, a seguir, o procedimento para identificação dos casos de sujeito oculto:

QUADRO 2 – Passos para a identificação de sujeitos ocultos

1. Encontrar sentenças em que não exista um sujeito (simples, oracional ou sujeito de oração passiva) dependente sintaticamente do núcleo da oração principal (*root*)
2. Das frases encontradas, eliminar as seguintes:
   a. Construções em que *root* é o verbo haver impessoal (3ª pessoa do singular)
   b. Construções em que *root* não é verbo[12]
   c. Construções em que *root* é verbo que indica fenômeno da natureza
   d. Construções em que o *se* é um índice de indeterminação do sujeito

Nas etapas acima, o principal desafio está no item (2.d): distinguir, nos casos de ausência de sujeito, aqueles em que o *se* corresponde a um índice de indeterminação (*Diga-se que*...; *Trata-se de*...), dos casos em

[12] Comum em manchetes jornalísticas ou interjeições.

que o *se* é complemento (*Cortou-se*) ou parte de um verbo pronominal (*Formou-se*). Nos primeiros não há o que ser explicitado, nos últimos, há sujeito a ser explicitado. Dois outros fatores também tornam este o filtro mais difícil: a anotação UD não diferencia os casos (14-15) de (16), ambos recebem a etiqueta *expl*.; e a distinção entre esses casos e o caso (17), no qual o *se* recebe a etiqueta de *obj*, ainda é pouco confiável de um ponto de vista automático.

14 Recorde-se aliás que, em Dezembro de 1992, quando a China realizou a maior explosão não nuclear de sempre.

15 Diga-se, de uma vez e claramente, o que se quer ou o que se quer mais.

16 Formou-se em engenharia em 1931.

17 Penteava-se segundo a moda do tempo, mas sem afetação.

A partir da análise manual, fizemos 3 filtros (d1; d2; d3) para identificar as ocorrências que nos interessam, e assim separamos os casos do tipo (14-15), que não contam como omissão do sujeito, dos demais casos:

d1. Casos em que o *se* recebe a anotação morfológica de gênero não especificado
Veja-se que «absoluta prioridade» não é simplesmente uma expressão, mas um princípio constitucional (...)

d2. Casos em que o *se* se associa a um verbo no infinitivo
Tem que se demonstrar através de contas e de raciocínios que o expurgo significará perda.

d3. Casos em que o *se* se associa a um verbo transitivo indireto ou intransitivo[13]
Pense-se em Kingsley Amis, Malcolm Bradbury e Albert Finney.

[13] Notamos que a forma de buscar as construções no *corpus* (isto é, o filtro) não corresponde, necessariamente, a uma análise correta. Neste exemplo, o *se* é exatamente do mesmo tipo do filtro d1, mas, como mencionamos, nem sempre podemos contar com uma análise sintática perfeita no caso do *se*. A forma de buscar indica apenas que, nesse caso, as ocorrências que gostaríamos de encontrar estão anotadas, na grande maioria das vezes, dessa maneira.

De forma complementar, fizemos também uma busca por sujeitos ocultos em orações subordinadas, utilizando as mesmas estratégias listadas no Quadro 2.[14]

5 Resultados e análise

Anterior à apresentação dos resultados, precisamos garantir que aquilo que recuperamos com as buscas e os filtros é o que desejamos. Esta validação é crucial no contexto do processamento automático, sobretudo porque em dois dos *corpora* analisados estamos lidando com o resultado de uma análise sintática que não foi revista. Procedemos a uma verificação manual de uma amostra, a fim de medir o grau de confiança que podemos ter nos resultados, já que apenas o Bosque-UD foi revisto. Foram analisadas até 20 frases por filtro (alguns filtros devolveram menos de 20 ocorrências), considerando cada *corpus*, totalizando 572 frases. A Tabela 1 traz os resultados da análise e a Tabela 2, complementar, indica a quantidade total de casos recuperados por filtro, bem como o quanto esses casos representam considerando o total de orações principais e subordinadas em cada *corpus*. Chamamos de *busca ingênua* a busca por qualquer frase que não tenha um sujeito. A coluna *Aval* (avaliados) da Tabela 1 indica o total de ocorrências de cada filtro; a coluna *Corr* indica a quantidade de ocorrências corretas, isto é, que atendem às especificações da busca/filtro.

A partir da Tabela 1, vemos que os resultados dos filtros variam por *corpus*, e o primeiro dado que chama a atenção é a importância de um material revisto, já que os números do Bosque superam os dos demais *corpora* em todos os cenários, e no que se refere às orações principais, isto é ainda mais evidente. Nos demais *corpora*, os resultados indicam que o que capturamos, quando tentamos encontrar o sujeito omitido, está correto em pouco mais da metade das vezes. Vemos, também, que é mais difícil acertar a procura nas orações subordinadas que nas principais, e isso se deve igualmente a limitações do processamento automático. Quando nos detemos nos resultados de cada um dos filtros, temos uma imagem mais nítida do que recuperamos.

[14] O único filtro não replicado nas orações subordinadas foi o 2b, relacionado às frases sem verbo, uma vez que há uma série de construções que atendem a essa especificação, como adjuntos adverbiais, que nada têm a ver com a omissão do sujeito.

TABELA 1 – Resultados da análise manual, por filtro

	DHBB				**OBras**				**Bosque-UD**			
	Principal		**Subordinada**		**Principal**		**Subordinada**		**Principal**		**Subordinada**	
	Aval.	**Corr**	**Aval.**	**Corr**	**Aval**	**Corr**	**Aval**	**Corr**	**Aval**	**Corr**	**Aval**	**Corr**
Busca ingênua	20	20	20	20	20	20	20	20	20	20	20	20
V. haver	20	20	20	20	20	17	20	20	20	20	20	20
Nominais	20	15	--	--	20	14	--	--	20	20	--	--
V. Natureza	0	0	4	4	20	20	20	20	1	1	2	2
D 1 (SE)	20	7	20	1	20	9	20	2	20	18	5	4
D 2 (SE)	20	5	20	1	20	4	20	1	2	2	20	3
D 3 (SE)	20	1	20	3	20	11	20	6	20	15	20	11
TOTAL	120	68	104	49	140	95	120	69	103	96	87	60
Total de acertos	**56%**		**47%**		**67%**		**57%**		**93%**		**69%**	

TABELA 2 – Distribuição das ocorrências por filtro, por corpus.

	DHBB				**OBras**				**Bosque-UD**			
	Principal		**Subordinada**		**Principal**		**Subordinada**		**Principal**		**Subordinada**	
Busca ingênua	226122	47%	190381	56%	172124	48%	156040	49%	2777	29%	2617	38%
V. Haver	1083	0,2%	861	0,2%	5181	1,5%	4395	1,4%	124	1%	148	2%
Nominais	31599	6,5%	735	0,2%	43326	12%	2766	0,8%	1145	12%	50	0,7%
V. Natureza	0	0%	5	0%	113	0%	147	0%	1	0%	2	0%
D 1 (SE)	59	0%	757	0,2%	215	0%	154	0%	31	0%	5	0%
D 2 (SE)	79	0%	2692	0,8%	229	0%	1609	0%	2	0%	24	0%
D 3 (SE)	26425	5,5%	13189	3,8%	4915	1,4%	6447	2%	57	0,6%	75	1%

Os filtros relativos à busca ingênua, aos verbos que indicam fenômenos da natureza e ao verbo *haver* impessoal trazem os resultados esperados, obtendo 100% de acertos, exceto pelo OBras, que contém 3 erros que se devem a construções do tipo *há de v-inf*, como *há de comer* e *há de saber*. Já o filtro relativo às frases sem verbo é mais dependente de uma boa análise sintática, e por isso mesmo o filtro é preciso quando aplicado ao Bosque. O grande responsável pelos erros é o filtro do *se*:

tanto no DHBB quanto no OBras, as frases recuperadas estão, na imensa maioria, erradas – e, portanto, onde esperaríamos orações sem sujeito, temos um sujeito omitido. Além disso, no caso do DHBB, que tem uma estrutura de texto previsível, vimos que muitos erros são, na verdade, fruto de uma mesma estrutura que se repete em vários verbetes. Uma construção como *transferindo-se*, por exemplo,[15] respondeu por 10 dos 19 erros e por 8 dos 13 erros encontrados em orações subordinadas e principais, respectivamente. No Bosque-UD, se temos bons resultados para o *se* quando este se encontra em orações principais, a qualidade da análise cai sensivelmente quando estamos diante de orações subordinadas, e o filtro d2 é responsável pela maior parte dos erros. Como indicamos na seção 4.3, devido à inconsistência (e dificuldade) de análise, mesmo no material revisto, fizemos uma busca não exatamente pelo que queríamos, mas por como nos parecia que o *se* estava anotado. Como podemos observar, essa estratégia foi pouco eficaz: ainda que seja precisa quando aplicada às orações principais, dá conta de poucos casos; no caso das orações subordinadas, recupera muitas construções, mas quase todas são casos de sujeitos omitidos. Por outro lado, quando levamos em conta os dados da Tabela 2, vemos que os casos de *se* respondem por uma porção muito pequena do total de casos filtrados no *corpus*.

Por outro lado ainda, como na omissão dos sujeitos só há duas possibilidades de resposta (estar diante de um sujeito omitido ou não), se simplesmente descartamos os filtros do *se* nos *corpora* não revistos todos os erros passam a acertos (e vice-versa), e desse ponto de vista os resultados melhoram sensivelmente. Ou seja, considerando os números do DHBB, e apenas com relação aos filtros do *se* tomados como um todo, passamos de 21% e 8% de acertos para orações principais e subordinadas, respectivamente, para 78% e 91%. No OBras, também apenas no que se refere ao *se*, quando eliminamos os filtros passamos de 40% e 15% de acertos para 60% e 85%. Diante dos resultados, optamos por prosseguir da seguinte maneira: eliminação de todos os filtros do *se* no DHBB e no OBras, e eliminação, no Bosque-UD, dos filtros d2 e d3 apenas nos casos das orações subordinadas. Com a alteração, a análise dos filtros traz resultados bastante positivos quanto aos resultados das buscas pelo sujeito oculto (TABELA 3): 89% de acertos no DHBB, 88%

[15] Como em *Transferindo-se para o Partido Social Cristão (PSC), em novembro de 1986 concorreu a deputado federal constituinte*.

no Bosque e 83% no OBras. Lembramos ainda que o *se*, responsável absoluto pelos erros, interfere pouco no resultado final, visto sua baixa frequência quando consideramos cada *corpus* na íntegra. Adicionalmente, a análise da amostra revelou que, no DHBB, os casos de *se* que não permitem omissão de sujeito (e que, portanto, são considerados erros na nova contagem) referem-se em sua imensa maioria a construções com *tratar-se de*, de modo que uma eliminação simples dessas construções torna os resultados ainda mais precisos. Especificamente, encontramos 347 casos de *tratar-se* no DHBB e 258 no OBras. Com isso, na prática, os resultados das buscas pelos sujeitos ocultos são ainda mais precisos do que indica a Tabela 3.

Diante dos resultados positivos, prosseguimos com a contagem. A Tabela 4 apresenta, finalmente, o resultado da quantificação do sujeito oculto nos três *corpora*. Começamos a contagem com a *busca ingênua* e, das ocorrências obtidas, fomos excluindo, com os filtros, os casos em que embora não haja sujeito, não há uma omissão.

TABELA 3 – Resultado final da análise dos filtros

	DHBB	**OBras**	**Bosque-UD**
Total de acertos	89%	83%	88%

TABELA 4 – Distribuição das ocorrências de sujeito oculto por *corpus*

	DHBB		OBras		Bosque-UD	
	Principal	**Subord.**	**Principal**	**Subord.**	**Principal**	**Subord.**
Busca ingênua	226122	190381	172124	156040	2777	2617
v. haver impessoal	1083	861	5181	4395	124	148
Or. Sem verbo	31599	0	43326	0	1145	0
Fenômenos da natureza	0	4	100	127	1	2
Filtros se	Não se aplica	Não se aplica	Não se aplica	Não se aplica	90	5
Filtro "tratar-se de"	181	163	150	123	Não se aplica	Não se aplica
Total	193259(40%)	189353(55%)	123367(34%)	151395(47%)	1417(15%)	2462(36%)
Total de frases com sujeito oculto	**382612 (46.5%)**		**274762 (41%)**		**3879 (24%)**	

Como suspeitávamos, o *corpus* DHBB é o que apresenta a maior quantidade de omissões de sujeito, o que não espanta dada a natureza de seu conteúdo: verbetes biográficos ou temáticos, nos quais o tema/foco da frase dificilmente se altera, e, por isso, a omissão é o recurso estilístico utilizado para deixar o texto não repetitivo. Do ponto de vista da identificação e da extração automática de informação, os resultados alertam para o fato de que em quase metade das orações (46.5%) a extração de relações entre argumentos de verbo fica prejudicada devido à ausência de um sujeito sintático. O *corpus* de obras literárias também tem um número expressivo de orações em que o sujeito foi omitido (41%), o que se explica, igualmente, em temos de estilística – lembremos das omissões de sujeito nos discursos relatados que introduzem falas de personagens, por exemplo. No Bosque-UD, composto por notícias de jornal, ainda que a frequência de sujeitos ocultos seja bem menor (24%) que nos demais materiais, não é insignificante, e traz impactos em tarefas do PLN como a extração de citação. Também vemos, na Tabela 4, que a presença de sujeito oculto é maior em subordinadas, mas que a diferença entre omissões quanto ao tipo de oração varia entre 15% e 20%. Levando em conta que frequentemente a omissão do sujeito na oração principal não se desfaz nas subordinadas, e que, nesses casos, o sujeito está fora do âmbito da sentença, esta característica é um desafio a mais para o processamento automático do texto e da informação.

Em resumo, não são poucos os casos de omissão do sujeito em português, independente de gênero, ainda que os números variem em função do tipo de texto. Do lado do PLN, o alto índice indica o desafio na construção de proposições informativas – e mesmo que em alguns casos, sobretudo o de orações subordinadas, a recuperação do elemento sujeito possa ser feita no âmbito da frase, isso envolve algum trabalho de pós-processamento. Outro desdobramento da alta frequência dos sujeitos omitidos é um aumento da dificuldade para o processamento automático no nível sintático, já que a ausência de um elemento para preencher a posição do sujeito pode desencadear outros erros. Na seção a seguir, realizamos um experimento simples para avaliar o problema.

6 Experimento: omissão do sujeito e aprendizagem automática

Dos três *corpora* utilizados em nosso estudo, apenas o Bosque é um *corpus* revisto. Por isso, este foi o material utilizado para verificar

em que medida a explicitação dos sujeitos poderia facilitar o aprendizado automático. Para tanto, reconstituímos os sujeitos e os devolvemos às frases, seguindo o procedimento do Quadro 3, no qual os elementos novos são as etapas 3 e 4, esta última inspirada na solução de Hartman *et al*. (2014).

QUADRO 3 – Etapas do experimento de busca e reconstituição de sujeitos ocultos

1. Encontrar sentenças em que não exista um sujeito (simples, oracional ou sujeito de oração passiva) dependente do núcleo da oração principal (*root*)
2. Das frases encontradas em (1) eliminar as seguintes:
   a. Construções em que *root* é o verbo haver impessoal (3ª pessoa do singular)
   b. Construções em que *root* é verbo que indica fenômeno da natureza
   c. Construções em que *root* não é verbo[16]
   d. Construções em que o *se* é um índice de indeterminação do sujeito
3. Das frases encontradas em (1), identificar aquelas em que *root* é verbo de oração adverbial, contém pessoa e número compatíveis com o sujeito de *root* e antecede *root*.
   a. Devolver este sujeito para à esquerda do verbo ao qual se relaciona.
4. Das frases que não foram eliminadas, incluir um elemento sujeito imediatamente anterior ao verbo a que se associa. O sujeito é um pronome pessoal reto e corresponde aos elementos flexionais indicados pelo verbo, ou seja, um verbo na 1ª pessoa do singular recebe o pronome *eu*. Nos casos de formas participiais, a informação de gênero também é levada em conta. No caso de infinitivo, inserimos um elemento genérico *SUBJ*.

Na etapa 1, identificamos todos os casos em que uma oração não tem um sujeito associado, e na etapa 2 eliminamos, desses casos, aqueles em que não há sujeito a ser encontrado. As etapas 3 e 4 têm o objetivo de reconstituir os sujeitos nas ocorrências restantes, sendo que a etapa 3 devolve o sintagma sujeito completo. A seguir apresentamos alguns exemplos de frases com seus sujeitos reconstituídos (sublinhados):

> Original: Quando o povo suíço recusou, em 92, a adesão ao Espaço Económico Europeu, como já fizera com a ONU, cometeu um grave engano.
>
> Reconstituído: quando o povo suíço recusou, em 92, a adesão ao Espaço Económico Europeu, como já fizera com a ONU, o povo suíço cometeu um grave engano.

[16] Comum em manchetes jornalísticas, como em "PT no poder".

> Original: Os dirigentes da Fenprof avisam no entanto desde já que se Couto dos Santos insistir nalgumas das directrizes dos seus antecessores arranjará lenha para se queimar.
>
> Reconstituído: Os dirigentes de a Fenprof avisam em o entanto desde já que se Couto de os Santos insistir em algumas de as directrizes de os seus antecessores Couto Santos arranjará lenha para se queimar.
>
> Original: Quando a gente se diz engajado, corre o risco de evocar modelos anteriores e o engajamento hoje deve encontrar formas novas.
>
> Reconstituído: Quando a gente se diz engajado, a gente corre o risco de evocar modelos anteriores e o engajamento hoje deve encontrar formas novas.

A etapa 4 devolve sujeitos "genéricos", isto é, formas pronominais informadas por traços morfossintáticos do verbo ou de predicadores, e um genérico *SUBJ* para o caso em que tais informações não estão disponíveis (verbos no infinitivo, por exemplo). Ainda que o desenvolvimento de estratégias para a reconstituição correta seja um exercício desafiador e relevante, nosso foco aqui está apenas em verificar se a presença de um sujeito é um elemento capaz de facilitar a aprendizagem automática. E, para isso, interessa devolver os sujeitos às orações que de fato precisam de um (em oposição a todos os casos excluídos pelo filtro 2). O índice de acerto acima de 80% de cada estratégia (cf. TABELA 2) sugere que os resultados da reconstituição, de um ponto de vista da estrutura da oração, são confiáveis.

Separamos o Bosque-UD reconstituído nas partições *treino* (incluímos a partição *dev* no treino), e *teste*, segundo a distribuição do Bosque-UD.[17] Utilizando a ferramenta UDPipe, criamos um novo modelo (com os sujeitos reconstituídos) e o avaliamos. Criamos, ainda, um cenário alternativo de avaliação, que corresponde ao Bosque-UD clássico. Isto é, no treino, utilizamos sujeitos reconstituídos, mas no

[17] A divisão de um corpus (ou de um *dataset*) em partições de treino (*train*), desenvolvimento (*dev*) e teste (*test*) são próprias para o aprendizado de máquina, e indicam respectivamente o conjunto de dados que será usado para treinar (ou aprender), para realizar ajustes e para avaliar o modelo criado.

teste consideramos o texto original, com sujeitos ocultos. Quisemos, com isso, reproduzir um cenário real, no qual os textos não terão seus sujeitos reconstituídos, porque para reconstituí-los é necessária uma análise sintática prévia, e é justamente isso que estamos medindo.

A Tabela 4 apresenta os resultados da aprendizagem com e sem a reconstituição do sujeito em termos de métricas específicas para a avaliação de dependências sintáticas: UAS, LAS e CLAS (*content-word labeled attachment score*). Como podemos observar, os resultados são melhores no *corpus* com o sujeito reconstituído, sugerindo facilidade na aprendizagem. A maior alteração está na métrica CLAS, com aumento de 2%, e este é um dado relevante, uma vez que CLAS representa melhor os resultados da análise sintática por tratar de relações entre palavras de conteúdo/classes abertas. Especificamente, a métrica CLAS não leva em conta as relações entre um sinal de pontuação e qualquer elemento (visto que serão sempre uma relação do tipo *punct)* e as relações entre determinantes, preposições, auxiliares e conjunções e qualquer outra palavra, visto que serão sempre relações de um mesmo tipo: *det, case*; *aux* e *mark*, respectivamente (NIVRE; FANG, 2017). Por isso, CLAS é uma medida mais sensível ao aprendizado das relações sintáticas. Por outro lado, quando observamos os resultados relativos à reconstituição apenas no treino, a melhora, ainda que permaneça, é menor. Se este último dado traz algum desapontamento no que se refere à aprendizagem de aspectos sintáticos, sinaliza também que, para tarefas cujo foco não é a análise sintática, mas que necessitem de análise sintática prévia, como a anotação de papéis semânticos, a reconstituição de sujeitos é uma estratégia que merece ser considerada.

TABELA 5 – Comparação da aprendizagem sintática em diferentes cenários relativos à omissão dos sujeitos

	Bosque-UD v.2.6 clássico	Bosque-UD v.2.6 com reconstituição no treino e na validação	Bosque-UD v.2.6 com reconstituição só no treino
UAS (F1)	84.81	85.66	85.34
LAS (F1)	80.63	81.85	81.01
CLAS (F1)	72.67	74.81	72.83

7 Considerações finais e desafios futuros

Dentre as várias utilidades que um *corpus* oferece, fazemos uso de duas: *corpus* para medir um fenômeno, e *corpus* para treinar um sistema e gerar um modelo de língua. Nossa conclusão é a de que o tratamento do sujeito oculto em português é relevante no PLN por dois motivos: (i) porque é um fenômeno altamente frequente, e não levá-lo em consideração tem como consequência limitar as possibilidades da extração automática de conteúdos; e (ii) porque a reconstituição do sujeito, quando possível, é capaz de facilitar a aprendizagem automática de dependências sintáticas, o que é positivo para todas as tarefas que, em algum momento, se utilizam deste tipo de análise linguística. Com relação a esse segundo aspecto, chamamos a atenção para a relevância da modelagem linguística na aprendizagem automática. Sem qualquer interferência na forma de aprendizagem, apenas lidando com informação linguística, conseguimos um impacto positivo que pode ser de até 2% no aprendizado automático. Um terceiro cenário que pretendemos testar consiste em aplicar as regras de reconstituição de sujeito nos resultados da anotação automática e, então, reanotar sintaticamente o material com os sujeitos reconstruídos. Isso nos dará outra medida relacionada aos benefícios da reconstituição do sujeito na análise sintática.

Diante dos resultados obtidos, outra tarefa se impõe: a reconstituição precisa do sujeito em termos discursivos (e não apenas sintáticos ou formais), o que esbarra também na resolução de correferência.

Como um resultado adicional, mas não inesperado, ratificamos a necessidade e relevância de *corpora* com anotação linguística de qualidade. Quando se trata de *treebanks* revistos (e públicos), a língua portuguesa dispõe apenas do Bosque, em suas variadas versões, desde 2002. E, ainda assim, temos fenômenos linguísticos que carecem de um tratamento sistemático, como é o caso da classificação do pronome *se*.

Apesar de não ser o foco deste trabalho, a identificação de sujeitos ocultos, por um lado, e sua reconstituição, por outro, são de grande valia também para a área de simplificação textual e de leitura. De um ponto de vista didático, um *corpus* com sujeitos artificialmente (bem) reconstituídos pode funcionar como "gabarito" para exercícios de leitura, por exemplo. Porém, para que os resultados sejam confiáveis, é crucial garantir a qualidade da anotação sintática subjacente. Embora

um índice de acertos que se aproxima dos 90% seja suficientemente bom para uma série de tarefas de PLN, reconhecemos que ainda há espaço para melhorias (veja-se, por exemplo a página *NLP Progress*,[18] um repositório para o monitoramento da evolução de tarefas de PLN que elenca o estado da arte para as tarefas mais comuns. Infelizmente, grande parte das tarefas têm como alvo a língua inglesa).

Por fim, lembramos que o ganho que tivemos pode ser ampliado por meio de técnicas mais elaboradas de aprendizado de máquina. Para isso, disponibilizamos o *corpus* Bosque-UD reconstituído, além das regras que permitem a identificação do sujeito, bem como sua reconstituição.[19] Nesse contexto, chamamos a atenção para a relevância da modelagem linguística no aprendizado automático, tendo em vista a resolução de tarefas de PLN.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Iniciação Científica concedida a Elvis de Souza no âmbito do projeto "Construção de datasets para o PLN de Língua Portuguesa". Número do processo da bolsa: 128693/2019-3.

Contribuição dos autores

Cláudia Freitas foi responsável pela concepção do trabalho, e Elvis de Souza, pela preparação dos dados. A análise dos dados e a redação foram realizadas por ambos.

Referências

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: A Treebank for Portuguese. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2002), 3rd, 2002, Las Palmas de Gran Canaria. *Proceedings* […]. Las Palmas de Gran Canaria: ELRA, 2002. p. 1698-1703.

[18] Disponível em: http://nlpprogress.com/. Acesso em: 8 out. 2020.

[19] Disponível em: https://github.com/alvelvis/desocultando-sujeitos. Acesso em: 30 nov. 2020.

BICK, E. *The parsing system palavras*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Aahus, Dinamarca: Aarhus Universitetsforlag, 2000.

DURAN, M. S.; ALUÍSIO, S. M. Propbank-Br: a Brazilian Treebank Annotated with Semantic Role Labels. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 12), 8th., 2012, Istambul, *Proceedings* […]. Istambul: ELRA, 2012. p. 1862-1867.

ELSON, D.; MCKEOWN K. Automatic Attribution of Quoted Speech in Literary Narrative. *In:* CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI 10), 24th., 2010, Atlanta, *Proceedings* […]. Atlanta: The AAAI Press, 2010. p. 1013-1019.

FINATTO, M. J.; SCARTON, C.; ROCHA, A.; ALUÍSIO, S. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. *In:* 8TH BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL 2011), 8th., 2011, Cuiabá, *Proceedings* […]. Cuiabá: SBC, 2011. p. 49-58.

FREITAS. C.; ROCHA, P.; BICK, E. Um mundo novo na Floresta Sintá(c)tica – o treebank do Português. *Calidoscópio*, São Leopoldo, RS, v. 6, n. 3, p. 142-148, 2008. DOI: https://doi.org/10.4013/cld.20083.03

HARTMANN, N. S.; DURAN, M. S.; ALUÍSIO, S. M. Filling the Gap: Inserting an Artificial Constituent Where a Subject Is Omitted in Portuguese. *In:* WORKSHOP ON TOOLS AND RESOURCES FOR AUTOMATICALLY PROCESSING PORTUGUESE AND SPANISH (TORPOR), I., São Carlos, *Proceedings* […]. São Carlos: SBC, 2014. Disponível em: http://www.nilc.icmc.usp.br/semanticnlp/includes/projects/brazilis/artigos/ToRPorEsp,%202014.pdf. Acesso em: 8 out. 2020.

HIGUCHI, S.; SANTOS, D.; FREITAS, C.; RADEMAKER, A. Distant Reading Brazilian Politics. *In:* CONFERENCE OF THE ASSOCIATION DIGITAL HUMANITIES IN THE NORDIC COUNTRIES (DHN 2019), 4th., 2019, Copenhagen. *Proceedings* […]. Copenhagen: University of Copenhagen, 2019. p. 190-200.

JONES, K. S. Computational Linguistics: What about the Linguistics? *Computational Linguistics*, Cambridge, MA, v. 33, n. 3, p. 437-441, 2007. DOI: https://doi.org/10.1162/coli.2007.33.3.437

LUFT, C. P. *Moderna gramática brasileira*. Rio de Janeiro: Globo Livros, 2002.

MARCUS, M.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, Cambridge, MA, v. 19, n. 2, p. 313-330, 1993. DOI: https://doi.org/10.21236/ADA273556

NIVRE, J.; de MARNEFFE, M.C.; GINTER, F.; GOLDBERG, Y.; HAJIČ, J.; MANNING, C.D.; McDONALD, R.; PETROV, S.; PYYSALO, S.; SILVEIRA, N.; TSARFATY, R.; ZEMAN, D. Universal Dependencies v1: A Multilingual Treebank Collection. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'16), 10th., Portorož, *Proceedings* [...]. Portorož: ELRA, 2016. p. 1659-1666.

NIVRE, J.; FANG, C. Universal Dependency Evaluation. *In:* UNIVERSAL DEPENDENCIES WORKSHOP (UDW 2017), 2017, Gothenburg, *Proceedings* [...]. Gothenburg: Association for Computational Linguistics, 2017. p. 86-95.

RADEMAKER, A. CHALUB, F.; REAL, L.; FREITAS, C.; BICK, C.; de PAIVA, V. Universal Dependencies for Portuguese. *In:* INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DEPLING 2017), 4th., Pisa, *Proceedings* [...]. Pisa: Linköping University Electronic Press, 2017. p. 197-206.

RUANO SAN SEGUNDO, P. A Corpus-Stylistic Approach to Dickens' Use of Speech Verbs: Beyond Mere Reporting. *Language and Literature,* [*S.l.*], v. 25, n. 2, p. 113-129, 2016. DOI: https://doi.org/10.1177/0963947016631859

SAMPSON, G. *Empirical Linguistics*. London: Continuum, 2001.

SANTOS, D.; BICK, E. Providing Internet Access to Portuguese Corpora: the AC/DC project. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2000), 2nd., Atenas, *Proceedings* [...]. Atenas: ELRA, 2000. p. 205-210.

SANTOS, D. Linguateca's Infrastructure for Portuguese and How It Allows the Detailed Study of Language Varieties. *OSLa*: *Oslo Studies in Language,* Oslo, v. 3, n. 2, p. 113-128, 2011. DOI: https://doi.org/10.5617/osla.100

SANTOS, D.; FREITAS, C.; BICK, E. OBras: A Fully Annotated and Partially Human-Revised Corpus of Brazilian Literary Works in the Public Domain. 2018. Disponível em: https://opencor.gitlab.io/corpora/santos18obras. Acesso em: 8 de out. 2020.

de SOUZA, E.; FREITAS, C. ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente. *In:* WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TILic 2019), VI., 2019. Salvador. *Proceedings* [...]. Salvador: SBC, 2019. p. 15-18.

STRAKA, M.; HAJIC, J.; STRAKOVÁ, J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *In*: TENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'16), 10th., Portorož, *Proceedings* [...]. Portorož: ELRA, 2016. p. 4290-4297.